J. Ehlers G. Schäfer (Eds.)

Relativistic Gravity Research

With Emphasis on Experiments and Observations

Proceedings of the 81 WE-Heraeus-Seminar
Held at the Physikzentrum, Bad Honnef, Germany
2-6 September 1991

Springer-Verlag
Berlin Heidelberg New York
London Paris Tokyo
Hong Kong Barcelona
Budapest

Editors

Prof. Dr. Jürgen Ehlers
Max-Planck-Institut für Astrophysik
Karl-Schwarzschild-Str. 1, W-8046 Garching, FRG

Priv.-Doz. Dr. Gerhard Schäfer
Arbeitsgruppe Gravitationstheorie der
Max-Planck-Gesellschaft an der Universität Jena
Max-Wien-Platz 1, O-6900 Jena, FRG

ISBN 3-540-56180-3 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-56180-3 Springer-Verlag New York Berlin Heidelberg

Typesetting: Camera ready by author/editor using the TEX macro package from Springer-Verlag
58/3140-543210 - Printed on acid-free paper

Editors' Preface

This volume of Lecture Notes in Physics contains the proceedings of the "81. WE-Heraeus-Seminar" of the "Dr. Wilhelm Heinrich Heraeus und Else Heraeus-Stiftung", entitled "Aktuelle Entwicklungen in der Erforschung der relativistischen Gravitation" ("Recent Developments in Relativistic Gravity Research"). The seminar, held at the Physikzentrum Bad Honnef, Germany, on 2–6 September 1991, was intended to bring together scientists from Germany with research interest in experimentally or observationally oriented relativistic gravity, and particularly to inform younger scientists about recent developments in this field. The selected topics were:

Gravitational Lensing;
Relativistic Celestial Mechanics, Astrometry, Geodesy;
Metrology;
Gravitational Waves;
Compact Objects, Black Holes;
Tests of Newton's Law of Gravity;
Matter Wave Interferometry.

Constraints on strong-field relativistic gravity obtained recently were not included in the talks delivered at the seminar. For more details the reader is referred to the following two articles:
T. Damour and G. Schäfer, New Tests of the Strong Equivalence Principle Using Binary-Pulsar Data, *Phys. Rev. Lett.*, **66** (1991) 2549.
J.H. Taylor, A. Wolszczan, T. Damour and J.M. Weisberg, Experimental Constraints on Strong-Field Relativistic Gravity, *Nature*, **355** (1992) 132.

The seminar was attended by 60 people; 26 participants presented 31 talks of which 20 lasted 50–60 minutes. The proceedings give a balanced record of the seminar. However, not all talks have been included in these proceedings. The following list gives papers which were presented at the seminar but which have been published elsewhere:
F.V. Kusmartsev, E.W. Mielke and F.E. Schunck, Gravitational Stability of Boson Stars, *Phys. Rev. D*, **43** (1991) 3895; Stability of Neutron and Boson Stars: A New Approach Based on Catastrophe Theory, *Phys. Lett. A*, **157** (1991) 465.
H.-P. Nollert, U. Kraus, A. Rebetzky, H. Herold, T. Maile and H. Ruder, Relativistic Light Bending Near Neutron Stars, *Proc. 23rd ESLAB Symp. on Two Topics in X-Ray Astronomy*, Bologna, 13–20 Sept. 1989 (ESA SP-296, Nov. 1989), p. 551.
H.-P. Nollert, U. Kraus, H. Ruder and H. Herold, Relativistic Light Deflection and Light Curves of X-Ray Pulsars, *Proc. of The Sixth Marcel Grossmann Meeting on General Relativity*, ed. by H. Sato (World Scientific, Singapore 1992).
N. Salié, Influence of Time Dependent Gravitational Fields on Superconducting Oscillatory Circuits, *Astron. Nachr.*, **307** (1986) 335.

Garching — *J. Ehlers*
Jena — *G. Schäfer*
July 1992

Acknowledgements

On behalf of the participants the Editors would like to thank the "Dr. Wilhelm Heinrich Heraeus und Else Heraeus-Stiftung" for having included the meeting in the WE-Heraeus Seminar Series. We gratefully acknowledge the generous financial support.

Contents

Gravitational lensing

Jürgen Ehlers and Peter Schneider

Max-Planck-Institut für Astrophysik, Karl–Schwarzschild–Str. 1, W–8046
Garching bei München, FRG

1 Introduction

1.1 Historical Remarks

In his *Opticks*, published in 1704, Sir Isaak Newton raised the question: "Do not Bodies act upon Light at a distance, and by their action bend its Rays, and is not this action strongest at the least distance?" In view of the dominance of the corpuscular conception of light in the 18th century it is strange that the possible bending of light by gravity was computed only as late as 1784. Stimulated by his friend John Michell, Henry Cavendish then calculated the bending of light by a spherical body of mass M. Assuming light corpuscels to move like material particles, he found the deflection angle to be

$$\hat{\alpha} = \frac{2GM}{c^2 r} = \frac{R_s}{r} \quad , \tag{1}$$

provided the photon's speed is c at infinity and its closest distance r from the center of the body is much larger than R_s, the length now called the Schwarzschild radius of the body. Michell had discovered that a body of radius $R \leq R_s$ would not be visible to a distant observer, and this and related remarks motivated Cavendish to calculate the bending angle. These findings appear to be the earliest gravito–optical effects ever contemplated.

In 1801 the Bavarian astronomer Johann von Soldner, unaware of Cavendish's work, after studying Laplace's rediscovery of "black holes" in 1795 again found (1) and concluded that the effect, if it existed at all, was practically negligible on account of the accuracy with which angles could be measured at his time. (Recall that the first stellar parallaxes, which are comparable in size, were measured in 1838 only.)

The question of gravitational light deflection was taken up again by Albert Einstein. In 1907 and again in 1911 he also found the law (1), guided solely by his principle of equivalence. Finally in 1915, in possession of his field equation he noted, almost in passing, that space curvature doubles the bending so that (1) has to be replaced by

$$\hat{\alpha} = \frac{2R_s}{r} \quad , \tag{2}$$

the so–called *Einstein angle*. The law (2) has been verified by VLBI measurements to within an accuracy of 0.003 (1σ).

Several physicists, among them Lodge 1919, Eddington 1920, Chwolson 1924, Mandl 1936, Einstein 1936, soon realized that light deflection may lead to multiple images and changes of apparent brightness. The observability of such effects was considered very unlikely, though, due to the small probability for sufficient alignment of sources and deflectors, taken to be stars in our galaxy. However, in 1937 in two remarkably prescient papers Fritz Zwicky considered the possible astronomical importance of gravitational light bending by external galaxies and concluded that "the probability that nebulae which act as gravitational lenses will be found becomes practically a certainty". Though correct, the verification of this prediction had to wait until 1979 when D. Walsh, R.F. Carswell and R.J. Weymann tentatively interpreted a "double quasar" as a pair of images of one quasar and within one year A. Stockton and P. Young et al. identified the lensing galaxy. Since then more than 10 cases of multiply imaged quasars and several gravitationally lensed images (arcs, rings) of extended sources have been found. Even before — and, of course, with increased activity after — these discoveries the theory of gravitational lenses has been and is elaborated; one of the pioneers is Sjur Refsdal, now at the Hamburg Observatory, who pointed out, in particular, cosmological applications of the gravitational lens effect.

1.2 Remarks on the astrophysical significance of gravitational lensing

The study of gravitational lensing is of interest for several reasons, among them:
successful models of gravitational lens configurations based on (2) provide *evidence for the* operation of the *law of gravity on galactic* (and possibly supergalactic) *scales*;
galaxies as gravitational lenses act as *natural telescopes* permitting to see otherwise invisible, very distant objects;
light deflection can explain unusual observed shapes of cosmic objects such as *arcs* and *rings*; it can be used to *determine masses* of deflecting, visible and dark matter;
microlensing, i.e. time–dependent differential light deflection by compact objects within the observed light bundle may help to elucidate the size and structure of the energy source of QSOs, and it may be used to determine masses of the deflecting compact objects.
Finally, lensing may possibly be used to determine *cosmological parameters*, in particular the value of the Hubble constant, and statistical lens theory provides, inter alia, inhomogeneity — *corrections to* the idealized *observational relations* ($m-z$, diameter — z etc.) which would hold in a homogeneous universe.

The following sections are intended as an introduction to and overview of the, by now quite extensive, field of gravitational lensing. For more details, we refer to a forthcoming book (Schneider, Ehlers & Falco 1992, hereafter SEF) the conference reports Moran, Hewitt & Lo (1989), Mellier et al. (1990), Kayser & Schramm (1992), and the excellent review article by Blandford & Narayan (1992, hereafter BN), and papers quoted in those sources. The notation used here follows that of SEF.

2 Foundations. The deflection mapping

2.1 Heuristic derivation of the lens equation

Consider a situation as sketched in the spacetime diagram Fig.1. A source emits, continually or in burst events S_1, S_2, ...light which propagates in ray bundles I_1, I_2, I_1', ...to an observer where it arrives at events O_1, O_2, ...The various rays are deflected by intervening matter at D; therefore the observer will see several images corresponding to these bundles, which have fluxes S_1, S_2, ...and are separated by angular distances θ_{12}, ...A change of luminosity of the source at S_1 will be seen in the images with a time delay Δt_{12}.

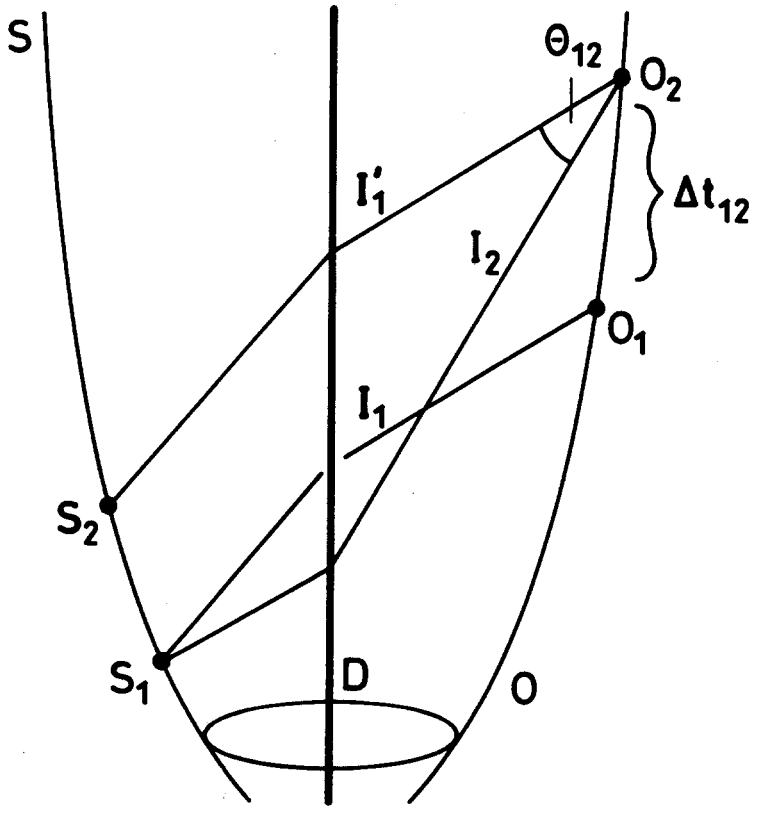

Fig. 1. A source S emits light at events S_1, S_2 which is received by the observer at events O_1, O_2. At O_2, the observer sees two images of the source, separated by an angle θ_{12}. The signal emitted at S_1 arrives at O_2 with a delat Δt_{12} relative to its arrival at O_1

The source and deflector redshifts, z_s and z_d, and the angular separations θ_{1d}, θ_{2d} of the images from the deflector may also be observable, as also the shapes, light distributions, spectra, and polarizations of the images. The primary aim of lens theory

is to determine relations between the observables θ_{12}, θ_{id}, Δt_{12}, S_1, S_2, z_s, z_d, the image shapes etc., and the (hypothetical) properties of the source, the deflector and the assumed cosmic background parameters H_0, Ω_0, q_0, $\tilde{\alpha}$ (clumpiness, see below).

It should be clear that such relations can be obtained only under simplifying assumptions. Before stating them and entering a detailed discussion of the bits and pieces of theory needed to obtain the desired relations, it is useful to infer — or rather to guess — the *form of the lens equation* simply from Fig.2. It shows the locations $\hat{S}$, $\hat{D}$ and $\hat{O}$ of a source S, a deflector D and an observer O.

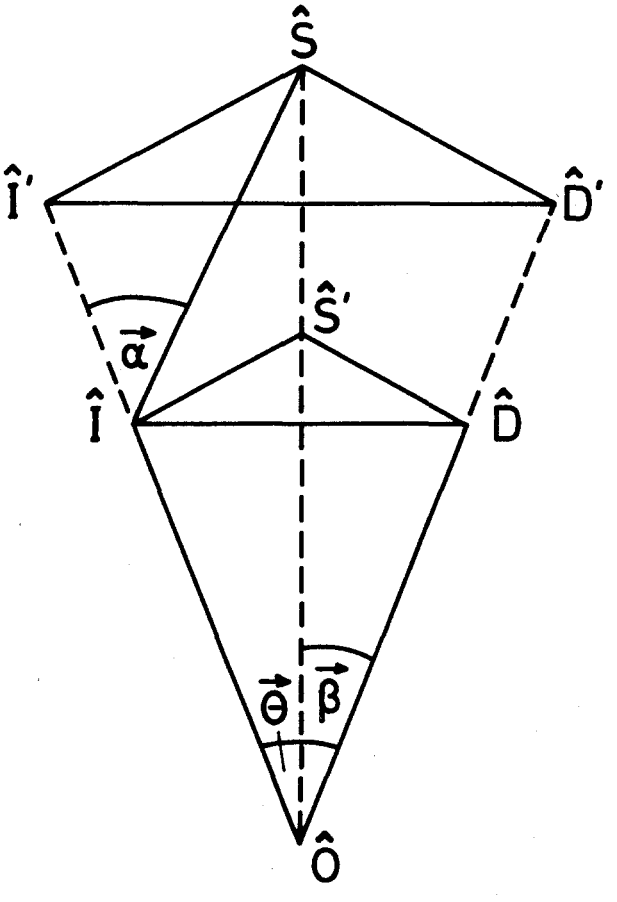

Fig. 2. Spatial projection of a lensing situation as described in the text

Without the gravitational field of D, light would proceed from $\hat{S}$ to $\hat{O}$ along the dashed line. In presence of the field, each light path is curved, but in small–angle approximation it may be approximated by the broken line $\hat{S}\hat{I}\hat{O}$. Light from $\hat{D}$ to $\hat{O}$ is (nearly) unaffected by D's field. The deflection angle at $\hat{I}$ is represented by the difference between the unit tangent vectors of the incoming and outgoing raypaths,

$\hat{\boldsymbol{\alpha}} := \mathbf{e}_{\rm in} - \mathbf{e}_{\rm out}$. The similar triangles $\hat{I}$, $\hat{D}$, $\hat{S}'$ and $\hat{I}'$, $\hat{D}'$, $\hat{S}$ are contained in the *lens plane* and the *source plane*, respectively, both orthonal to the *optical axis* $\hat{O}\hat{D}\hat{D}'$. The figure shows that the angular separation $\boldsymbol{\theta}$ of the source's image, $\hat{I}$, from the deflector $\hat{D}$ is related to the (unobservable) unperturbed separation $\boldsymbol{\beta}$ of $\hat{S}$ from $\hat{D}$ by the *lens equation*

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{D_{ds}}{D_s}\,\hat{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \quad . \tag{3}$$

D_s and D_{ds} denote the distances of $\hat{S}$ from $\hat{O}$ and $\hat{D}$, respectively. To obtain eq.(3), use has been made of

$$\boldsymbol{SI'} = D_{ds}\hat{\boldsymbol{\alpha}}\ , \quad \boldsymbol{D'S} = D_s\boldsymbol{\beta}\ , \quad \boldsymbol{D'I'} = D_s\boldsymbol{\theta} \quad . \tag{4}$$

Since the figure is supposed to represent a process in an expanding universe containing an inhomogeneous matter distribution, it is not obvious which 3-dimensional space is displayed in Fig.2, and which metric is used to define distances, angles and straight lines. This geometry must be such that the paths of light rays are geodesics, that angles equal the physical angles, and distances are related to angles by (4). Moreover, the dependence of the deflection angle $\hat{\boldsymbol{\alpha}}$ on the mass distribution of the deflector has to be found, the "timeless" description has to be augmented by finding the time–delays referred to above, and the unobservable distances have to be related to observable redshifts. All this will be done below.

2.2 Geometrical optics and Fermat's principle

2.2.1 The WKB approximation

Before carrying out the programme just outlined, we recall some facts about *short–wave asymptotic solutions of Maxwell's equations* in an arbitrarily curved spacetime, to put the following considerations on a firm basis.

The (real) phase S of a locally approximately plane and monochromatic electromagnetic wave

$$F_{\alpha\beta} = 2\mathcal{R}\mathrm{e}\left\{\mathrm{e}^{\mathrm{i}S} k_{[\alpha} A_{\beta]}\right\} \tag{5}$$

with frequency 4–vector $k_\alpha = -S_{,\alpha}$ in a matter–free region obeys, in leading WKB–approximation, the *eikonal equation*

$$g^{\alpha\beta} S_{,\alpha} S_{,\beta} = 0 \quad , \tag{6}$$

and its complex amplitude A_α is transported along the *rays*, defined by

$$\frac{dx^\alpha}{dv} = k^\alpha = -g^{\alpha\beta} S_{,\beta} \quad , \tag{7}$$

according to the *transport equation*

$$\frac{DA_\alpha}{dv} = -\frac{1}{2}\, k^\beta{}_{;\beta} A_\alpha \quad . \tag{8}$$

These statements form the basis of geometrical optics. Eq.(7) implies the rays to be *null geodesics*, and (8) specifies how energy–momentum and polarization are transported.

Since, for a rapidly changing phase S, the frequency measured by an observer with 4–velocity U^α is $\omega = -\frac{dS}{d\tau} = -S_{,\alpha}\frac{dx^\alpha}{d\tau} = -k_\alpha U^\alpha$, the frequency ratio between a source and an observer is given by

$$1 + z = \frac{\omega_s}{\omega_o} = \frac{(k \cdot U)_s}{(k \cdot U)_o} \tag{9}$$

where, according to (7), k^α is parallel on a ray.

2.2.2 Fermat's principle

For the purposes of lens theory the most useful way to characterize light rays, i.e. null geodesics, is *Fermat's principle.* Its role for gravitational lensing was recognized by one of us (PS) in 1985. The general form of this principle as given below was stated by I. Kovner in 1990; its proof is due to V. Perlick in 1990. The principle says: a smooth null curve γ connecting a source event S to an observer worldline l is a light ray if, and only if, its arrival time τ on l is stationary under all first–order variations of γ within the set of smooth null curves from S to l,

$$\delta\tau = 0 \quad . \tag{10}$$

To prove this, one considers a family $x^\alpha(v,\varepsilon)$ of null curves connecting S to l, $0 \leq v \leq 1$, $|\varepsilon| < \varepsilon_0$. Then, with a dot denoting covariant differentiation along the curves, one obtains by the ususal integration by parts,

$$0 = \delta\,\frac{1}{2}\int_0^1 \dot{x}_\alpha \dot{x}^\alpha dv = (\dot{x}_\alpha \delta x^\alpha)_{v=1} - \int_0^1 \ddot{x}_\alpha \delta x^\alpha dv \quad ,$$

i.e.,

$$\dot{x}_\alpha \delta x^\alpha = \int_0^1 \ddot{x}_\alpha \delta x^\alpha dv \quad .$$

Here, the left–hand side refers to the end point of the unvaried curve $x^\alpha(v,0)$, δx^α denotes the variation of that end point, and the integrand is to be evaluated at $(v,0)$. If u^α denotes the unit tangent of l at the end point, $\delta x^\alpha = \delta\tau\, u^\alpha$, thus

$$\delta\tau(u_\alpha \dot{x}^\alpha) = \int_0^1 \ddot{x}_\alpha\, \delta x^\alpha\, dv \quad . \tag{11}$$

Since u^α is timelike and $\dot{x}^\alpha$ is lightlike, $u_\alpha\, \dot{x}^\alpha \neq 0$. The proof, therefore, reduces to the assertion that the right–hand side of (11) vanishes for all admissible variations if, and only if, $x^\alpha(v,0)$ is geodesic. The necessity of that condition is obvious, for if $x^\alpha(v,0)$ is geodesic, we may choose v to be an affine parameter so that $\ddot{x}_\alpha = 0$. The proof of sufficiency is trickier. One has to show that there are sufficiently many admissible variations δx^α such that, with a suitable choice of v, $\ddot{x}^\alpha = 0$. This proof will not be reproduced here, see Perlick (1990) or SEF.

For a *static spacetime*, i.e. if the metric has the special form

$$ds^2 = \mathrm{e}^{2U/c^2} c^2\, dt^2 - \mathrm{e}^{-2U/c^2} dl^2 \tag{12}$$

with U and the conformally rescaled spatial metric $dl^2 = \gamma_{ab} dx^a\, dx^b$ independent of the time coordinate t, Fermat's principle reduces to

$$c\,\delta t = \int_{\tilde{\gamma}} \mathrm{e}^{-2U/c^2} dl \quad , \tag{13}$$

where $\tilde{\gamma}$ denotes the spatial projection of a light ray and δt denotes the variation of the arrival time; now the end points have to be kept fixed under variations of the spatial curve $\tilde{\gamma}$. Eq.(13) has the form of the classical Fermat principle for a medium with index of refraction

$$n = \mathrm{e}^{-2U/c^2} \quad . \tag{14}$$

The variational principle (13) may be rephrased as follows: The spatial paths of light rays in a static spacetime are the geodesics of 3–space with respect to the (Riemannian) optical metric $\mathrm{e}^{-4U/c^2}\gamma_{ab}$ corresponding to the Lagrangian

$$L(x^a, e^a) = \mathrm{e}^{-2U/c^2}\sqrt{\gamma_{ab} e^a e^b} \quad ; \tag{15}$$

now $e^a = \frac{dx^a}{dl}$, and $a, b = 1, 2, 3$. The resulting Euler–Lagrange equation

$$\frac{De^a}{dl} = \frac{-2}{c^2}\left(\gamma^{ab} - e^a e^b\right) D_b U \quad ,$$

or

$$\frac{D\mathbf{e}}{dl} = \mathbf{D}_\perp \log n = -\frac{2}{c^2}\mathbf{D}_\perp U \quad , \tag{16}$$

states that the curvature of the ray–path is equal to the component of $\mathbf{D}\log n$ orthogonal to the ray direction. Apart from the fact that now the metric is Riemannian and D is its covariant derivative, this is exactly the law of classical geometrical optics (see, e.g., Sommerfeld 1959, §48A).

2.3 The deflection angle

We are now in a position to derive the light deflection by a nearly static, isolated mass distribution which produces a weak gravitational field. Then, U in (12) may be taken to be the Newtonian gravitational potential, and in linear approximation (with respect to U/c^2) to Einstein's field equation, dl^2 is Euclidean, so that

$$ds^2 = \left(1 + \frac{2U}{c^2}\right) c^2\, dt^2 - \left(1 - \frac{2U}{c^2}\right) d\mathbf{x}^2 \quad . \tag{17}$$

Defining the deflection "angle" $\hat{\boldsymbol{\alpha}}$ as the difference between the asymptotic "in" and "out" values of the ray tangent $\mathbf{e}$, we obtain from (16)

$$\hat{\boldsymbol{\alpha}} = \frac{2}{c^2}\int \nabla_\perp U\, dl \quad . \tag{18}$$

Under realistic conditions the deflection is very small, $|\hat{\boldsymbol{\alpha}}| < 10^{-4}$, say. For a point mass, $U(\mathbf{x}) = -GM/|\mathbf{x}|$, and integration over the unperturbed ray $\mathbf{x}(l) = \boldsymbol{\xi} + l\mathbf{e}$ with impact vector $\boldsymbol{\xi} \perp \mathbf{e}_{\mathrm{in}}$ leads to the Einstein angle

$$\hat{\boldsymbol{\alpha}} = \frac{4GM}{c^2}\,\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|^2} \quad .$$

Let us now assume not only that the total deflection angle is very small, but that the extent of the deflecting mass in the direction $\mathbf{e}_{\mathrm{in}}$ is so small that the value of $\nabla_\perp U$ on the actual ray differs but little from that on the unperturbed, straight ray. Then, one may integrate over the unperturbed ray in (18); thus for such a *geometrically thin lens* the deflection angle is equal to the sum of the Einstein angles of its mass elements. Accordingly, all mass elements in an infinitesimal cylindrical tube parallel to $\mathbf{e}$ have the same impact vector. We may therefore project all mass elements of the lens onto a plane orthogonal to $\mathbf{e}$, passing through some (rather arbitrary) "centre" of the lens, and characterize the deflector by the resulting surface mass density $\Sigma(\boldsymbol{\xi})$. Thus,

$$\hat{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \frac{4G}{c^2}\int_{\mathbf{R}^2} \frac{(\boldsymbol{\xi}-\boldsymbol{\xi}')\Sigma(\boldsymbol{\xi}')}{|\boldsymbol{\xi}-\boldsymbol{\xi}'|^2}\, d^2\xi' \quad ,$$

where the integral is over the *lens plane* just introduced, and $\boldsymbol{\xi}$ is a 2–dimensional vector in that plane. This equation, which may be rewritten in terms of the *deflection potential*

$$\hat{\Psi}(\boldsymbol{\xi}) = \frac{4G}{c^2}\int \Sigma(\boldsymbol{\xi}')\ln\frac{\boldsymbol{\xi}-\boldsymbol{\xi}'}{D_d}\, d^2\xi'$$

as

$$\hat{\boldsymbol{\alpha}} = \nabla\hat{\Psi} \quad , \tag{20}$$

is a basic relation of lens theory. (Here the distance D_d of the deflector from the observer has been introduced for dimensional reasons only.) Note that the two $\frac{U}{c^2}$–contributions in (17), the time dilation term and the space curvature term, contribute equally to $\hat{\boldsymbol{\alpha}}$. This causes Einstein's angle to be twice "Newton's".

The foregoing consideration shows that, for the case where source, deflector and observer are part of an isolated, nearly static system with a weak gravitational field, the lens equation (3) with (19), (20) holds. In this case Fig.2 should be interpreted as referring to the 3–dimensional background x–space with Euclidean metric $d\mathbf{x}^2$ which occurs in (17).

2.4 The time delay

2.4.1 'Local' treatment

Next we wish to calculate the *arrival time delay* for a ray from the source S to the observer O caused by the gravitational field of the deflector D. According to (17) and $ds^2 = 0$, the delay is

$$\Delta t = \frac{1}{c}\,\Delta l - \frac{2}{c^3}\int U dl \quad . \tag{21}$$

It consists of a geometrical and a potential term. Let us for the moment concentrate on the latter, the Shapiro delay. A straightforward calculation shows that, under the assumptions made already — small deflection angles, weak, static field, thin lens — we have approximately

$$\Delta t_{\mathrm{pot}} = -\frac{1}{c}\,\hat{\Psi} + \mathrm{const.} \tag{22}$$

The first term agrees with the deflection potential whose argument is given by $\boldsymbol{\xi} = D_d\boldsymbol{\theta}$, and the second term depends on the locations of S, D and O, but not on the ray considered which is specified by $\boldsymbol{\xi}$, the position of the image in the lens plane (Fig.2). Eq.(22) holds not only for "physical" rays deflected by the "true" angle (20), but for all "kinematically possible" rays proceeding along a broken line SIO. This fact can be used to rederive the lens equation (3) by means of Fermat's principle: Calculating — here for the case of a Euclidean background space — Δl (which we leave to the reader) and extremising the Δt of eq.(21) with respect to variations of the position of the deflection point I using (22) and (20), indeed reproduces (3).

2.4.2 The time delay in an expanding universe

In real lensing situations, deflectors and sources are at large cosmological distances; thus the foregoing "local" treatment needs to be generalised. For this purpose one assumes that the spacetime metric can on average be represented by a *Robertson–Walker metric*

$$ds^2 = c^2dt^2 - R^2(t)\,d\sigma_k^2 \quad , \tag{23a}$$

$$d\sigma_k^2 = du^2 + S_k^2(u)(d\vartheta^2 + \sin^2\vartheta\,d\phi^2) \quad , \tag{23b}$$

$$S_k(u) = \begin{cases} \sin u & 1 \\ u & \text{if } k = 0 \\ \sinh u & -1 \end{cases} \quad , \tag{23c}$$

and that the actual metric deviates substantially from (23a) only in "small" regions around lumps of matter such as galaxies. One may then again use Fig.2 to obtain the lens equation, provided one now interprets (i) $\hat{S}$, $\hat{D}$, $\hat{O}$, $\hat{I}$ and the lines connecting them as the projections of source, deflector, light rays into the "comoving 3–space" with metric (23b), and (ii) distances to be *angular diameter distances*. Indeed, with these interpretations eqs.(4) remain valid except for the common factor R_s^{-1} on the right–hand sides, whence (3) holds again.

To obtain the potential time–delay in the cosmological case, we use the fact that the significant, $\boldsymbol{\xi}$–dependent part of $\Delta t_{\rm pot}$ in (22) arises near the deflector and therefore amounts, at the observer, to the red–shifted value

$$\Delta t_{\rm pot} = \frac{-1}{c}\,(1+z_d)\,\hat{\Psi} + \text{ const.} \tag{24}$$

It remains to compute the geometric time delay. It is due to the path–difference between the perturbed and unperturbed rays whose 3–space projections may be taken, for the present purpose, to be geodesics. Rewriting the RW metric (23) in the conformally static form

$$ds^2 = R^2[\mu]\,(d\eta^2 - d\sigma_k^2)$$

we see that $c\Delta t_{\rm geom} = R_0\Delta\eta_{\rm geom} = R_0(\sigma_{ds} + \sigma_d - \sigma_s)$ (see Fig.3). Consider first the case $k = 1$. Then, from spherical trigonometry,

$$\cos\sigma_s = \cos\sigma_{ds}\cos\sigma_d - \sin\sigma_{ds}\sin\sigma_d\cos\hat{\alpha} \quad .$$

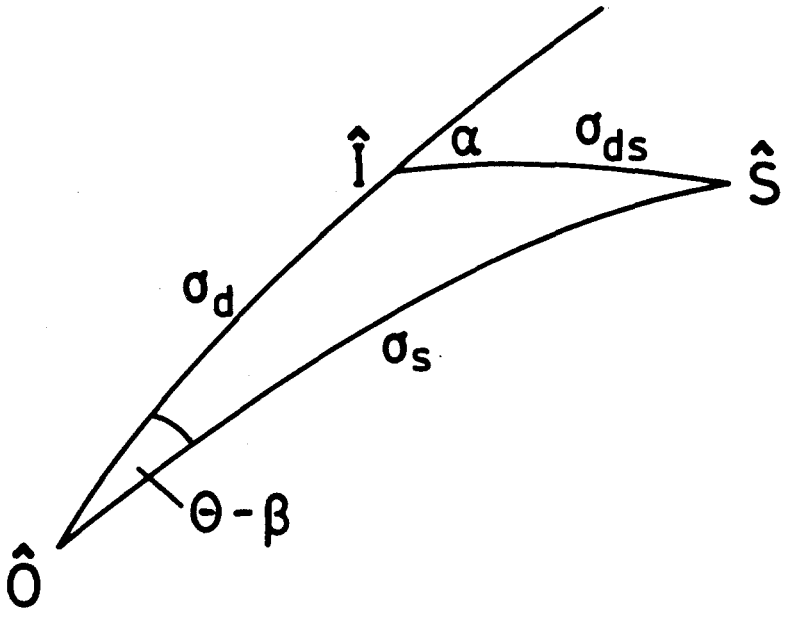

Fig. 3. Spatial projection of the world lines of source, observer and deflection event near world line of the lens, into the comoving 3-space of an Robertson–Walker universe model. Shown are also the spatial paths of unlensed ($\hat{S}\hat{O}$) and deflected ($\hat{S}\hat{I}\hat{O}$) ray paths

After a little calculation this gives, for $\hat{\alpha} \ll 1$,

$$\Delta t_{\rm geom} = \frac{R_0}{c} \frac{\sin\sigma_{ds} \sin\sigma_d}{2\sin\sigma_s} \hat{\alpha}^2 \quad .$$

If the σ–distances are expressed in terms of the angular diameter distances,

$$\begin{aligned} D_{ds} &= R_s \sin\sigma_{ds} \ , \quad D_d = R_d \sin\sigma_d \quad , \\ D_s &= R_s \sin\sigma_s \quad , \\ (\boldsymbol{\theta} - \boldsymbol{\beta}) \sin\sigma_s &= \hat{\boldsymbol{\alpha}} \sin\sigma_{ds} \quad , \end{aligned}$$

there results for the total time delay

$$c\Delta t = (1 + z_d) \left\{ \frac{D_d D_s}{2 D_{ds}} (\boldsymbol{\theta} - \boldsymbol{\beta})^2 - \hat{\Psi}(\boldsymbol{\xi}) \right\} + \text{ const.}$$

The same expression holds for $k = -1$ and $k = 0$.

It is convenient to introduce a dimensionless deflection potential $\tilde{\Psi}$ via

$$\hat{\Psi}(\boldsymbol{\xi}) = 2R_s\tilde{\Psi}(\boldsymbol{\theta}) \quad ,$$
$$\tilde{\Psi}(\boldsymbol{\theta}) = \int \frac{dm'}{M} \ln\left|\boldsymbol{\theta} - \boldsymbol{\theta}'\right| \quad , \tag{25}$$

where $dm'/M = D_d^2\,\Sigma(\boldsymbol{\theta}')\,d^2\theta'/M$ represents the fraction of the source's mass "visible" by the observer in the solid angle $d^2\theta'$.

The result can be written as

$$c\Delta t = \hat{\phi}(\boldsymbol{\theta},\boldsymbol{\beta}) + \text{ const.} \quad , \tag{26a}$$

$$\hat{\phi} = (1+z_d)\left\{\frac{D_d D_{ds}}{2D_{ds}}\,(\boldsymbol{\theta}-\boldsymbol{\beta})^2 - 2R_s\tilde{\Psi}(\boldsymbol{\theta})\right\} \quad . \tag{26b}$$

As in the "local" case, these formulae for Δt hold for all "kinematically possible" light rays along SIO. Fermat's principle, specialized to $\frac{\partial\phi}{\partial\boldsymbol{\theta}} = 0$, therefore again gives the lens equation (3),

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{2D_{ds}}{D_d D_s}\,R_s\,\frac{\partial\tilde{\Psi}}{\partial\boldsymbol{\theta}} \quad . \tag{26c}$$

The factor $\frac{2D_{ds}}{D_d D_s}$ is due to the RW–geometry while $R_s\,\frac{\partial\tilde{\Psi}}{\partial\boldsymbol{\theta}}$ accounts for the local deflection.

The unobservable distances in (26) can be related to red shifts of source and deflector, if a cosmological model is used. Since we are concerned with processes occurring well after hydrogen recombination, we may use an "on average Friedmann" dust model and take, for simplicity, $\Lambda = 0$.

Under plausible assumptions discussed in detail in SEF, one can derive the following formula:

$$(1+z_d)\,\frac{D_d D_{ds}}{D_{ds}} = \frac{c}{H_0}\cdot\frac{1}{\chi_d - \chi_s} \quad , \tag{27}$$

where H_0 denotes the Hubble constant and $\chi(z;\Omega_0;\tilde{\alpha})$ is a function depending on the redshift z, the density parameter Ω_0 of the universe and the smoothness parameter $\tilde{\alpha}$ which measures which fraction of the mass is smoothly distributed, i.e. not bound in galaxies or clusters. The function χ can be computed for any statistically homogeneous Friedmann model. (The angular diameter distances appearing in (27) and used in computing χ are those introduced by Dyer& Roeder(1973).)

2.4.3 Summary

The *result* of these considerations can be summarized as follows:

The time delay for rays SIO is given by

$$c\Delta t = \hat{\phi} + \text{ const.} \quad , \tag{28}$$

where

$$\hat{\phi}(\boldsymbol{\theta},\boldsymbol{\beta}) = \frac{c}{2H_0}\,\frac{(\boldsymbol{\theta}-\boldsymbol{\beta})^2}{\chi_d - \chi_s} - 2R_s(1+z_d)\,\tilde{\Psi}(\boldsymbol{\theta}) \tag{29}$$

is called the *Fermat potential.* The images of a source at $\boldsymbol{\beta}$ are located at the stationary points of the Fermat surface $\boldsymbol{\theta} \to \hat{\phi}(\boldsymbol{\theta},\boldsymbol{\beta})$; they are obtained by inverting the lens equation

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \frac{2R_s H_0}{c}(1+z_d)(\chi_d - \chi_s)\,\frac{\partial\tilde{\Psi}}{\partial\boldsymbol{\theta}} \quad . \tag{30}$$

Hence, the angular separation $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\theta}_j$ between any two images is

$$\boldsymbol{\theta}_{ij} = \frac{2R_s H_0}{c}(1+z_d)(\chi_d - \chi_s)\left\{\frac{\partial\tilde{\Psi}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_i) - \frac{\partial\tilde{\Psi}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_j)\right\} \quad , \tag{31}$$

and the time delay is

$$\begin{aligned}\Delta t_{ij} = \frac{2R_s}{c}(1+z_d)\left\{\frac{1}{2}\,\boldsymbol{\theta}_{ij}\cdot\left(\frac{\partial\tilde{\Psi}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_i) + \frac{\partial\tilde{\Psi}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_j)\right)\right.\\ \left. - \left(\tilde{\Psi}(\boldsymbol{\theta}_i) - \tilde{\Psi}(\boldsymbol{\theta}_j)\right)\right\} \quad .\end{aligned} \tag{32}$$

(The dot between 3-vectors indicates Euclidean inner products.) The *distortion* of an image near $\boldsymbol{\theta}$ is given by the inverse μ_{ij} of the 2×2 Jacobian matrix

$$\frac{D\boldsymbol{\beta}}{D\boldsymbol{\theta}} = (\beta_{ij}) \tag{33}$$

of the map (30). While this distortion is, of course, not observable, the relative distortion of two images at $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, given by $\mu_{ij}(\boldsymbol{\theta}_1)\,\beta_{jk}(\boldsymbol{\theta}_2)$, may be observable. In the above equations, the dimensionless deflection potential $\tilde{\Psi}$ is given, in terms of the mass distribution of the deflector, by (25), and the χ–values refer to z_d and z_s, respectively.

3. Magnification, odd–number theorem and magnification theorem

The results derived above concern geometrical and kinematical variables. What can be said about *fluxes* of deflected light bundles? The form (5) of the electromagnetic field tensor implies that the corresponding energy–momentum tensor, averaged over a period or wavelength, has the same form as that of a stream of photons with number 4–current density $N^\alpha \propto k^\alpha$, each photon carrying 4–momentum $P^\alpha = \hbar k^\alpha$:

$$T^{\alpha\beta}_{\text{eff}} = N^\alpha P^\beta \quad .$$

Moreover, as long as there are no interactions of the photons with matter, $T^{\alpha\beta}_{\text{eff}\ ;\beta} = 0$, whence photons are conserved,

$$N^\alpha{}_{;\alpha} = 0 \quad .$$

From this one concludes (see, e.g., SEF) that the ratio I_ω/ω^3 of the specific intensity by the cube of the circular frequency ω of a light bundle is observer independent (locally Lorentz invariant) and constant along the central ray of the bundle. The frequency shift of a distant source is (practically) not affected by lensing, since each photon looses as much energy in climbing out of the potential well of the deflector as it gained when falling into it. Hence, I_ω is not affected by lensing, so that the (monochromatic or integrated) observed flux $S_\omega = I_\omega \Delta\omega$ *does* change just like the

solid angle $\Delta\omega$ which a small area at the source subtends at the observer, due to differential deflection. Thus, lensing changes the flux by the *magnification* (factor)

$$\mu = |\det(\mu_{ij})| = |\det(\beta_{ij})|^{-1} \tag{34}$$

whose ratio for two images gives the observable flux ratio or relative magnification.

The results derived so far imply two general theorems about lensing:
The number of images of a point source produced by a transparent mass distribution acting as a lens, is always odd (Burke 1981), and
at least one of the images is amplified ($\mu \geq 1$) relative to the case where there is no lens between source and observer, other things being equal (Schneider 1984). Proofs of these assertions are given, e.g. in SEF.

Both of these theorem also apply for more complicated situations than the ones considered here, where a single optically thin deflector acts on the light rays. The odd-number theorem was treated by McKenzie (1985) for a fairly general spacetime. The magnification theorem also applies in a general spacetime, owing to the focusing equation. For the case that a light ray is deflected by several geometrically thin matter distributions between source and observer, both theorems have been proved in the framework of gravitational lensing in Seitz and Schneider (1992).[1]

4. An example: The point–mass lens

Let us specialize the general relations considered so far to the case of a spherically symmetric lens of negligible size, formally idealized as a point–mass. Then it suffices to consider the lens map in the plane containing source, deflector and observer. Eq.(30) then reduces to

$$\beta = \theta - \frac{2R_s D_{ds}}{D_d D_s}\,\theta^{-1} \quad , \tag{35}$$

see Fig.4. The image positions of a source at β are given by

$$\theta_{+,-} = \frac{1}{2}\left(\beta \pm \sqrt{4\alpha_0^2 + \beta^2}\right) \tag{36}$$

where

$$\alpha_0 = \sqrt{\frac{2R_s D_{ds}}{D_d D_s}} \tag{37}$$

is the (angular) *Einstein radius*, the value of $\theta_+ = -\theta_-$ for $\beta = 0$. The angular separation of the images,

$$\Delta\theta = \theta_+ - \theta_- = \sqrt{4\alpha_0^2 + \beta^2} \quad , \tag{38}$$

[1] The basic idea for the proof of the odd-number theorem is to consider the Poincare index of the vector field $\boldsymbol{\theta} - \boldsymbol{\beta}$, and applying the index theorem. The magnification theorem considers light rays where the light travel time has an absolute minimum; such light rays cannot have passed through a caustic; otherwise, an even shorter light ray could be constructed. Minimal light travel time and the non-negativity of the surface mass density of the deflectors then yields the desired result.

is at least $2\alpha_0$, and the unlensed position β is

$$\beta = \theta_+ + \theta_- \quad . \tag{39}$$

Moreover, by (36) and (27),

$$\alpha_0^2 = |\theta_+\theta_-| = \frac{2R_sH_0}{c}\,(1+z_d)(\chi_d - \chi_s) \quad , \tag{40}$$

and the relative magnification turns out to be

$$\nu := \frac{\mu_+}{\mu_-} = \left(\frac{\sqrt{4+\tilde{\beta}^2}+\tilde{\beta}}{\sqrt{4+\tilde{\beta}^2}-\tilde{\beta}}\right)^2 \quad . \quad \left(\tilde{\beta} := \frac{\beta}{\alpha_0}\right) \tag{41}$$

Use of (32) and some algebra leads to the following expressions for the arrival time delay $\Delta t = t_- - t_+$:

$$\begin{aligned} \Delta t &= \frac{2R_s}{c}\,(1+z_d)\left[\frac{\theta_+^2-\theta_-^2}{2\,|\theta_+\theta_-|} + \ln\left|\frac{\theta_+}{\theta_-}\right|\right] \\ &= \frac{R_s}{c}\,(1+z_d)\,(\nu^{1/2} - \nu^{-1/2} + \ln\nu) \quad . \end{aligned} \tag{42}$$

Finally we note that the Einstein radius can be inferred from

$$\alpha_0 = \frac{\Delta\theta}{\nu^{1/4}+\nu^{-1/4}} \quad . \tag{43}$$

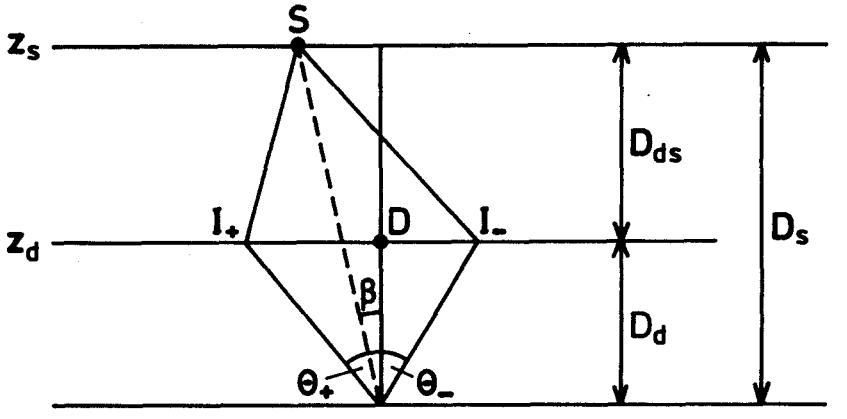

Fig. 4. Lensing by a point mass at S with images I_+, I_- and observer at O

Therefore, in this simple model situation, we see:

(i) the mass $M = \frac{c^2}{2G}R_s$ of the lens can be found from the observables Δt, z_d, θ_+, θ_- or alternatively from Δt, z_d, ν;

(ii) the product $R_s H_0$ — and therefore, with (i), the Hubble constant — can be obtained from θ_+, θ_-, z_d, z_s, provided one knows the cosmological density parameter Ω_0 and clumpiness parameter $\tilde{\alpha}$ which enter χ;

(iii) three point–mass lensing observations would give the three lens masses as well as the cosmological parameters H_0, Ω_0, $\tilde{\alpha}$;

(iv) the combination of (40) and (43) provides an equation between the three observables θ_+, θ_-, ν, and thus a test of the model.

The case considered here is admittedly very idealized, but it nevertheless indicates the possibility of using gravitational lens observations to determine masses and cosmological parameters, and the equations provide orders of magnitude and typical scales. Of cource, the analysis becomes much more involved for extended lenses whose mass distribution has to be determined too. Clearly, resolved images of extended sources contain much more information than images of point sources.

It is perhaps useful to insert here a remark about so-called Einstein rings, first predicted by Chwolson in 1924, as mentioned above, the case where source, lens and observer are colinear. Rotational symmetry about the line of sight then implies that, according to geometrical optics, a point source appears as a ring to the observer. However, in this ideal case infinitely many light rays with exactly equal path length intersect at the observer, and therefore diffraction and interference invalidate geometrical optics. In contrast, any real source is extended and not strictly monochromatric. Then, (i) those parts of a wave train emitted from a source point located not on, but near to the axis, which reach the observer from directions corresponding to the two geometrical images of the source point, arrive at the observer separated in time, if the time delay exceeds the coherence time; and (ii) waves from different source points are incoherent. Therefore, unresolved, extended sources, nearly aligned with deflector and observer, produce nearly the same ring images which geometrical optics predicts for the idealized case.

5 Singularities of lens mappings

5.1 The lens equation as Lagrangean mapping

The lens mapping, eq.(30), was obtained assuming that the "angles" β, θ are very small; therefore we may take β and θ to vary in small neighbourhoods of the origins of two copies of $\mathbb{R}^2$. To study these mappings mathematically it is convenient to consider them as maps from the whole of $\mathbb{R}^2$ to $\mathbb{R}^2$, remembering that they are "realistic" near the origins only. Absorbing factors into $\tilde{\Psi}$ and $\hat{\phi}$ and changing notation, we write the lens map

$$f : \mathbb{R}^2 \to \mathbb{R}^2$$

now as

$$\mathbf{x} \to \mathbf{y} = \mathbf{x} - \nabla\Psi(\mathbf{x};\mathbf{p}) = \nabla\left(\frac{1}{2}\mathbf{x}^2 - \Psi(\mathbf{x},\mathbf{p})\right) \quad . \tag{44}$$

Here $\mathbf{p}$ denotes a set of N parameters characterizing a family of lensing situations, $\mathbf{p} \in \mathbb{R}^N$. Examples of such parameters are the core radii of galaxies, their ellipticities, the separation between lens components, redshifts, and the distance ratio D_{ds}/D_s.

Alternatively, the source position $\mathbf{y}$ corresponding to an image $\mathbf{x}$ can be characterized by

$$\nabla\phi = 0\,, \quad \phi = \frac{1}{2}(\mathbf{x}-\mathbf{y})^2 - \Psi(\mathbf{x};\mathbf{p}) \quad . \tag{45}$$

(As before, the gradient operator refers to $\mathbf{x}$.)

To avoid confusion, we shall use the terminology appropriate to the *physical meaning* of the symbols; thus $\mathbf{y}$ is called the source position, $\mathbf{x}$ the image position (although according to mathematical terminology $\mathbf{x}$ is the pre-image, $\mathbf{y}$ the image variable for the map f defined by (44)).

One basic question of lens theory is: How many images exist for a given source, i.e. what is the set $f^{-1}(\mathbf{y})$, and how does it depend on $\mathbf{y}$? For localized sources the deflection potential Ψ increases like $\ln r$, the deflection angle decreases like r^{-1}. Therefore, as one would expect, the map f is bijective for large $|\mathbf{x}|$. Also, f is surjective, i.e. for any source position $\mathbf{y}$, there is at least one image $\mathbf{x}$ such that $\mathbf{y} = f(\mathbf{x})$. Thirdly, if the Jacobian matrix $\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right)$ is invertible at $\mathbf{x}$, the map f is locally invertible at $\mathbf{x}$. Therefore, if $\mathbf{y}$ moves inwards from infinity along a curve, there will for a while be a unique image $\mathbf{x}$ until possibly a point is reached where $\det\left(\frac{D\mathbf{y}}{D\mathbf{x}}\right) = 0$. The set of all points $\mathbf{x}$ where this equation holds is called the *critical set*, its image the *caustic set* of f. In the theory of singularities of maps one studies the behaviour of f and f^{-1} near critical and caustic points, respectively, in general. Lens theory is concerned with the special case of *gradient maps*, see eq.(44).

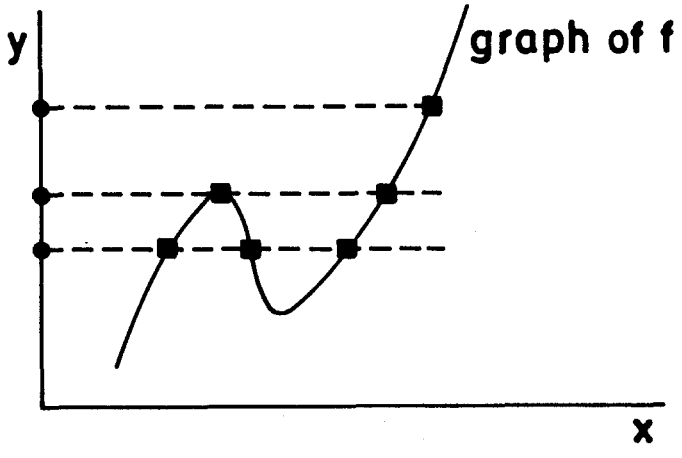

Fig. 5. The graph of the lens mapping illustrates the dependence of the number and position of images on the source position

One may study such maps from several points of view, two of which we mention.

Firstly, one may consider f as defining a 2–surface in $\mathbb{R}^4 = \{(\mathbf{x},\mathbf{y})\}$. On this 2–surface, the symplectic form $\Omega = dy_i \wedge dx_i$ vanishes; thus the 2–surface is a Lagrangean submanifold of $(\mathbb{R}^4, \Omega)$. The lens map may then be viewed as a *projection* of this 2–

surface onto the $\mathbf{y}$–plane, see Fig.5. This is useful since it gives a "quasi–intuitive" insight into the way how $f^{-1}(\mathbf{y})$ changes with $\mathbf{y}$, and since such *Lagrangean maps* have been studied extensively by mathematicians.

In the second view, based on (45), one thinks of $\mathbf{x} \to \phi(\mathbf{x},\mathbf{y};\mathbf{p})$, for each fixed value of the *control parameters* $(\mathbf{y}, \mathbf{p})$, as a surface in $(\mathbf{x}, \phi)$–space. For large $\mathbf{x}$, this arrival time surface approaches the paraboloid $\phi = \frac{1}{2}(\mathbf{x} - \mathbf{y})^2$. The images, in this context also called *states*, then correspond to those points where the tangent plane to the surface is parallel to the plane $\nabla\phi = 0$. (This second view corresponds to the so–called "static model" of catastrophe theory, which is popular also in discussions of phase transitions.)

By eqs.(44) and (45), the Jacobian matrix of the lens mapping has components

$$\frac{\partial y_i}{\partial x_k} = \phi_{ik} = \delta_{ik} - \Psi_{ik} \quad . \tag{46}$$

Since for $x \to \infty$, $\Psi_{ik} \to 0$, the critical set is bounded and, of course, closed, hence compact. The caustic set is compact and in addition of measure zero (Sard's theorem).

At a regular, i.e. non–critical image $\mathbf{x}^{(0)}$, the magnification μ (eq.(34)) is finite, and near $\mathbf{x}^{(0)}$, f is locally diffeomorphic. At a critical image $\mathbf{x}^{(0)}$, the magnification is formally infinite; however, if instead of a point source an extended source is considered or if wave optics is taken into account, the "physical" magnification becomes finite, but in general large.

It follows from the above that if $\mathbf{y}$ moves, the number of its images can change only if $\mathbf{y}$ enters or crosses the caustic set. Thus knowing the critical and caustic sets provides a qualitative overview of a lens mapping.

5.2 Generic singularities of lens mappings

Let us write $D := \det(\phi_{ik})$. At a critical image $\mathbf{x}^{(0)}$, the rank of $A := (\phi_{ik})$ can be 1 or 0. In the first case one can have, at $\mathbf{x}^{(0)}$, either $\nabla D \neq 0$ or $\nabla D = 0$. The only *stable* critical points, i.e., those which remain critical under arbitrary, small changes of the deflection potential, are those with rank $A = 1$ and $\nabla D \neq 0$. These are the only critical points which occur generically, i.e. in "almost all" lens mappings.

Generically, the critical set of a lens mapping consists of finitely many closed, smooth, compact curves without end points. Let $\mathbf{T}^{(0)}$ denote the tangent of a critical curve at $\mathbf{x}^{(0)}$. Then, in general, $A\mathbf{T}^{(0)} \neq 0$, but at isolated points it may happen that $A\mathbf{T}^{(0)} = 0$. Thus, a critical curve in general consists of open arcs where $A\mathbf{T} \neq 0$, separated by points at which $A\mathbf{T} = 0$. The images of critical curves, the caustics, have *cusps* (spikes) at those points which correspond to critical points with $A\mathbf{T} = 0$.

Whenever a source point crosses a caustic where the latter is smooth, two new images appear on opposite sides of the critical curve; the corresponding segments of critical curves which consist of "double images", are called *folds*. Near and inside a cusp, a source point has three images which merge when the source reaches the cusp, and only one image survives if the source has passed through the cusp. The point on the critical curve where this happens, i.e. where the tangent of the critical curve is in the kernel of A, is also called a cusp point of the lens mapping. Both folds and cusps are stable, i.e., they are preserved under all small deformations of the deflection potential, and smooth maps $\mathbb{R}^2 \to \mathbb{R}^2$ have no other kinds of stable singularities

than folds and cusps (Whitney 1955). Near a fold, the amplification diverges like the inverse square root of the distance from the caustic; near a cusp, it diverges even stronger. These facts can be established by approximating the Fermat–potential at a critical point by a polynomial of suitably high order and then studying the resulting representative mapping.

If one considers not a single lens mapping, but a family of those depending on some parameters $\mathbf{p}$, the critical curves and caustics depend on the value of $\boldsymbol{p}$. Qualitative changes of the pattern of critical curves and caustics, called *metamorphoses*, can occur for particular values $\mathbf{p}^{(0)}$ and at special points $(\mathbf{x}^{(0)}, \mathbf{y}^{(0)})$. These higher–order singularities are also useful to survey lensing models; they are needed too to study caustics and self–intersections of null cones of spacetimes, but we shall not pursue this topic here, but refer the reader to Chap. 6 of SEF (see also Blandford & Narayan 1986).

6 Observed gravitational lens phenomena

Having outlined the theoretical foundations and concepts, we now turn to the astrophysical aspects of gravitational lensing, starting by describing some of the cosmic sources for which the lensing phenomenon plays an essential role: multiple images of QSOs have been observed, ring-shaped radio sources are generally believed to be created by the lens action of a foreground galaxy, and long, narrow arcs in clusters of galaxies are thought to be highly distorted images of background galaxies. We describe these classes of objects in turn, ending this section with a brief account on microlensing and its observations.

6.1 Multiple QSOs

6.1.1 0957+561

The discovery of the first gravitational lens system came about by identifying the optical couterpart of a source in a radio catalog (for the history of this discovery, see Walsh 1989). The system 0957+561, the famous 'double quasar', consists of two QSO images, separated by 6.1 arcseconds. The spectra of these images are essentially indistinguishable, except for a slight reddening of the B image relative to the A image; in particular, they share the redshift of $z_s = 1.41$ to within the errors of its determination. Thus, the optical spectroscopy yielded strong indications for the source to be gravitationally lensed. For this to be the case, a massive galaxy (with an estimated mass of $10^{12} M_\odot$) must be responsible for the image splitting. In fact, this galaxy was found independently by two groups (Stockton 1980, Young et al. 1980); it has a redshift of about $z_d = 0.36$ and is the brightest member of a compact cluster of galaxies. With such a massive galaxy in the foreground of a QSO, gravitational lensing of the background source is unavoidable, making this a secure example of the lensing phenomenon (at least if we stick to the cosmological interpretation of the redshift as a distance indicator).

Nevertheless, initially doubts were raised against the lensing interpretation, for at least two reasons: if this was a lens system, why are only two images observed, whereas theory predicts an odd number of images (Sect. 4)? Second, the radio structures of

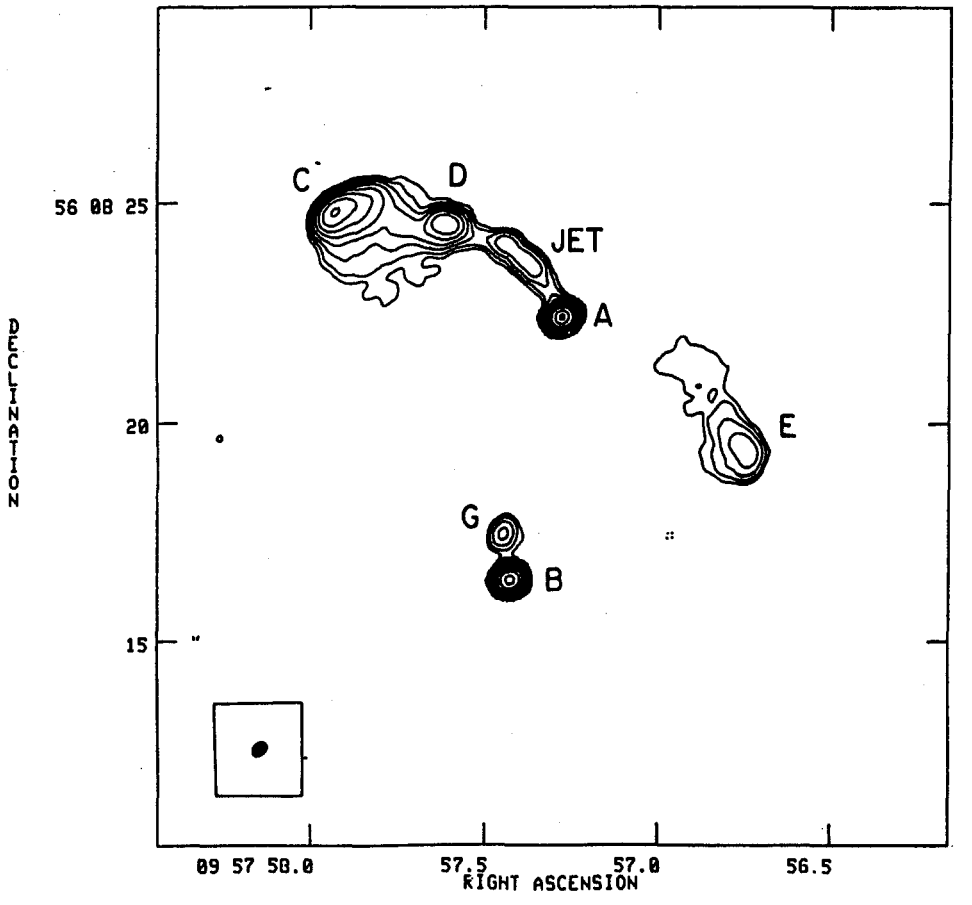

Fig. 6. A VLA map of the gravitational lens system 0957+561, taken at 6 cm wavelength. The compact components A and B coincide with the optical QSOs, and component G is situated at the position of the main lens galaxy. In contrast to component B, component A has a jet and extended radio lobes (C, D, and E). The separation between A and B is 6.1 arcseconds (from Roberts et al. 1985)

the two images appear very different (see Fig. 6). If A and B are indeed images of the same source, should they not share a similar radio morphology? Both concerns were justified at the time of discovery of this system, since lens theory up to this time mainly dealt with fairly special mass distributions, which cannot be applied to this particular lens system. The usefulness of the odd number theorem is limited by the fact that nothing is said about the flux ratios of the images. The third image can indeed be present but be too weak for being observable. In fact, component G in Fig. 6 cannot be the third image of the QSO, since it is resolved (whereas A and B are unresolved) and its radio spectrum is steeper than that of the two QSO images;

hence, G is probably the radio image of the lens galaxy. If a third image, which would be expected to lie very close to the center of the galaxy, is present as part of component G, its flux must be less than 2% of that of image A for being compatible with the VLBI observations. The fact that A has extended radio structure such as a jet and outer lobes, whereas B has not, is explained by placing the source close to a caustic formed by the lens: the compact component is situated inside the caustic curve and thus multiply imaged, whereas the extended source components are outside the caustic and have a single image only. Thus, image A is a – distorted – image of the whole source. This interpretation is supported by a slight extension of B in the direction of G, so that a small part of the jet is multiply imaged.

VLBI observations of 0957+561 (Gorenstein et al. 1988) have erased all doubts about the lensing nature of this system. Both compact components (A and B) share the same compact radio morphology, consisting of a compact core (unresolved at milliarcsecond resolution) and two jet components. The VLBI maps of the two images can in fact be related to each other by a linear transformation, as expected from lens theory: the scale of the VLBI source is much smaller than the scale of the deflector mass distribution, so that the mapping should be locally linear (the linear transformation between the two images is the product of the distortion matrix at one image and the inverse of the distortion matrix at the other image, see the end of Sect. 2; the four components of this relative distortion matrix have been determined by the observations and provide strong constraints on any model for this system). Most impressive, the parities [= sign of the determinant of the Jacobian matrix (33)] of the two VLBI images are different, i.e., one is a distorted mirror image of the other, as predicted by all lens models for 0957+561.

An intrinsic flux variation of the source should show up in both images, but occur with a time delay equal to the difference in the light travel time along the two corresponding light rays [see (26)], thus providing the possibility of measuring the time delay. Fortunately, 0957+561 is a variable source, both optically and in the radio, though not very strongly variable. Flux measurements of both components have been taken for more than a decade, so it appears to be fairly easy to obtain the time delay from these data. Unfortunately, this is not the case: the cross correlation of these two light curves turns out to be fairly involved, since the optical data are interrupted by monthly (observations avoid full moon) and yearly gaps (the source is observable for about eight months per year), and random gaps due to bad weather. Therefore, some kind of interpolation of the light curves of both images is needed before cross correlating them. The result of the correlation depends fairly strongly on the method used; it is therefore not surprising that several different values for Δt have been claimed in the literature. Recently, a more sophisticated statistical method was used to obtain Δt from both, the optical and the radio data (Press, Rybicki & Hewitt 1992a,b), yielding a result of 540 ± 12 days.

The determination of the Hubble constant from the measurement of the time delay, as discussed above, requires knowledge on the mass distribution of the lens. However, in this respect the system 0957+561 is a fairly complicated one, since the cluster in which the main lensing galaxy is embedded makes a significant contribution to the deflection. Unfortunately, basically nothing is known about the mass distribution of the cluster. The construction of lens models proceeds by assuming that the deflection caused by the cluster varies slowly over the region of the size of the image separation;

the contribution by the cluster is then described by the lowest order terms of a Taylor expansion of the deflection angle around the center of the main galaxy (the validity of this approach can be questioned, see Kochanek 1991). A 'plausible' ansatz for the shape of the mass distribution in the lens galaxy is chosen, compatible with what is known about the matter distribution in elliptical galaxies, and the parameters of the model are varied to obtain a match of all observables with the model. Needless to say that such an approach yields models which are far from being unique. Even worse, there are invariance transformation of lens model parameters which have no impact on the observables (Gorenstein, Falco & Shapiro 1988; e.g., adding a uniform mass sheet to the lens acts like a Gaussian thin lens and is equivalent, in terms of the lens model, to decreasing the separation between lens and source). This degeneracy can be broken by obtaining additional observables; in the case of 0957+561, the velocity dispersion of the lens galaxy (a measure for the lens mass) has been observed, thus allowing the degeneracy to be broken. It thus seems that a point estimate of the Hubble constant is possible from this lens system. Taking the lens model from Falco, Gorenstein & Shapiro (1991), assuming that the velocity dispersion measures the total mass of the lens galaxy, and taking a cosmological model with $\Omega_0 = 1$, one obtains a value for H_0 which is less than $50\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (the 'best' value being closer to 40). The uncertainties, however, are still considerable. First, as already mentioned, the lens model is not unique, and one can construct equally good models which would yield different values for H_0. Second, it is not clear whether the velocity dispersion of the stars probes the total matter of the lens galaxy; if the stellar mass is more centrally concentrated than the dark matter in the galaxy, the effective dispersion can be larger than the measured value by up to a factor $\sqrt{1.5}$. Third, additional inhomogeneities around the line-of-sight to the QSO can perturb the propagation of light and thus affect the lens mapping; in this respect it is interesting to know that spectroscopy of the galaxies around 0957+561 indicate that they do not all belong to a cluster at $z_d = 0.36$, but that there seems to be an additional concentration at a higher redshift of about 0.54.

With all the difficulties and uncertainties mentioned, it might appear that the determination of the Hubble constant from lensing will not yield more accurate values than the classical method of 'climbing up the distance ladder'. However, it should be stressed that the latter method measures the Hubble constant from nearby objects, whereas lensing permits to obtain measurements of H_0 on truly cosmic scales. It is by no means evident that both methods should yield the same results, in fact: since the probability that a galaxy (on which an observer is situated to measure H_0) is placed in an overdense region of the universe, the local Hubble expansion is likely to be slower than the mean Hubble expansion, yielding a systematically higher value of H_0 from local measurements. This was demonstrated quantitatively by Turner, Cen & Ostriker (1991). Hence, measuring H_0 from lensing might provide the only way to obtain that value of H_0 which enters the Friedmann equations.

6.1.2 2237+0305

As a second example for a multiply imaged QSO, we display in Fig. 7 an isophotal plot of the system 2237+0305, which was discovered in the course of the CfA redshift survey of galaxies (Huchra et al. 1985). A redshift $z_d = 0.04$ spiral galaxy shows broad emission lines in the core, indicative of a QSO at redshift $z_s = 1.69$. More detailed imaging showed that the QSO is split into four images, situated nearly (though not completely) symmetrically around the center of the galaxy. Spectroscopy of the four individual images revealed that they indeed belong to the same source; their spectra differ only in the continuum colour, which is easily understood by differential reddening caused by the interstellar medium of the lensing galaxy. Owing to the small redshift of the lens in this system, the light distribution of the deflector can be mapped in great detail. Assuming a constant mass-to-light ratio across the galaxy, a lens model can be constructed with just one single parameter (this mass-to-light ratio), which has been shown be consistent with the observed positions of the images; hence, a constant mass-to-light ratio is at least compatible with the observations (however, this lens model is again far from unique, see Kent & Falco 1988, Kochanek 1991). The nearness of the lens in this system makes this also an ideal target for investigating microlensing (see below).

6.1.3 Other cases; discussion

Presently, we know about seven firm cases of multiply imaged QSOs, and an additional handful of good candidates for lensed QSOs. Verification of a candidate system to be truly a gravitational lens system is fairly difficult in practice. 'Similarity' of QSO spectra is not a quantitative measure, and the 'lensing community' has learned its lesson from systems like 1145+071, where two QSOs with very similar optical spectra pretend to be an ideal lensing candidate, but radio observations have shown that only one QSO is a radio source, with very strict upper limits on the radio flux ratio: it is almost certainly not a lens. In order for a system to be called a 'lens system' instead of 'candidate system', one or more of the following criteria should be satisfied: (1) more than two images with similar spectra, (2) agreement of optical and radio flux ratio, (3) a candidate deflector between or very near to the images. The spectra need not be identical in all details, since they can be modified by differential absorption in the lens, intrinsic variability of the source (together with the time delay), and microlensing (see below). Of the seven known multiple QSOs, four are quadruples, two are doubles, and there is one triple QSO (2016+112). The large fraction of quadruples has not been predicted by fairly simple models, since the probability for a lens to cause four images is smaller by a factor of about 10 than for forming double images; it can only be understood by accounting for selection effects (amplification bias, see below). The two doubles both have a visible lensing galaxy. The triple system 2016+112 is not very well understood yet; it appears that a generic successful model must allow for multiple deflection, i.e., two deflectors at different redshifts (for which there is indication from the objects in its field). Of the four quadruples, two have the images arranged nearly symmetrically (as in 2237+0305), and the fluxes of the four images are comparable; in the other two systems, two of the four images are very close together, and much brighter than the other two, indicating that these images are close to a critical curve.

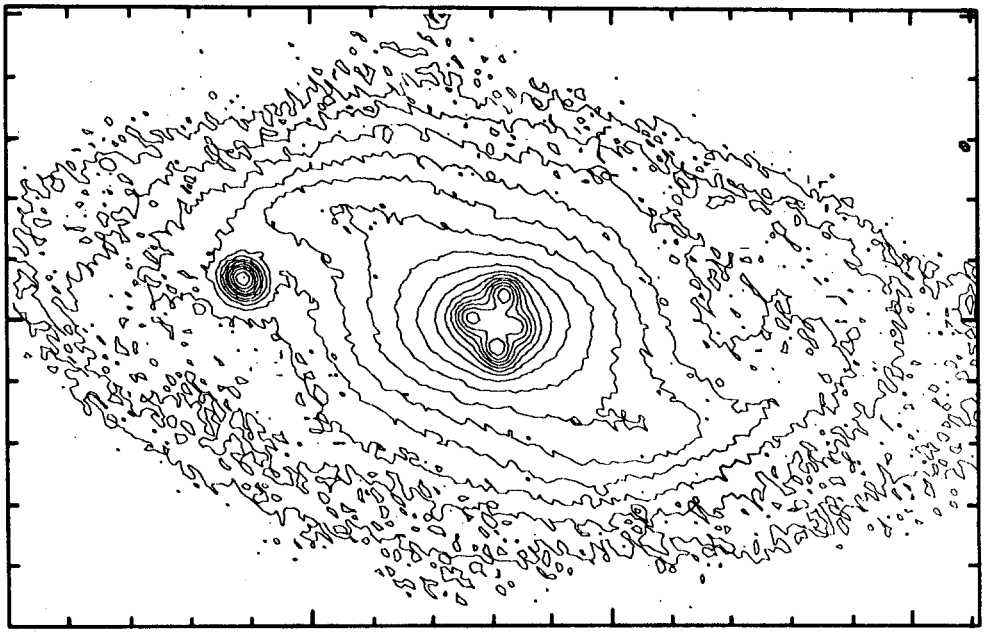

Fig. 7. Optical isophotes of the gravitational lens system 2237+0305, where a QSO with redshift $z_s = 1.69$ is split into four images by a foreground spiral galaxy at redshift $z_d = 0.04$. The maximum image separation in this system is 1.8 arcseconds (from Yee 1988)

Except for the ill-understood system 2016+112, all multiply imaged QSO systems have an even number of images, in contrast to what theory would predict. The canonical explanation for the absence of the odd image is its large demagnification: a light bundle traversing the inner core of the lensing galaxy is so much overfocused that the resulting flux of the corresponding image is very small and undetectable. The characteristic value for the surface mass density of a lens (or critical density) is

$$\Sigma_{\rm cr} = \frac{c^2}{4G}\frac{D_s}{D_d D_{ds}} \quad ; \tag{47}$$

a lens with surface mass density larger than $\Sigma_{\rm cr}$ in at least one point can produce multiple images of a source which is sufficiently well aligned. Conversely, for a centrally-condensed axially-symmetric mass distribution to be able to form multiple images, the central surface mass density Σ_0 must exceed $\Sigma_{\rm cr}$. If $\Sigma_0 = \kappa_0 \Sigma_{\rm cr}$, the central image will have a magnification of about $\mu = (\kappa_0 - 1)^{-1}$, which can be very small if the core radius of the deflector is small. None of the observed multiple QSOs requires a model with finite core size.

6.2 Rings

As we have seen in Sect.4, if a source is exactly aligned with an axisymmetric deflector of sufficient strength, a ring-shaped image of the source occurs. If the source is extended, the exact alignment is no longer necessary for a ring to appear; rather, the size of the inner – or tangential – caustic must be smaller than or comparable with the source size. Ring-shaped images were predicted by Chwolson in 1924. Five of them were discovered during the last few years, and we will here describe the two best-studied ones. All observed ring-shaped images appear in the radio and have been discovered in radio surveys.

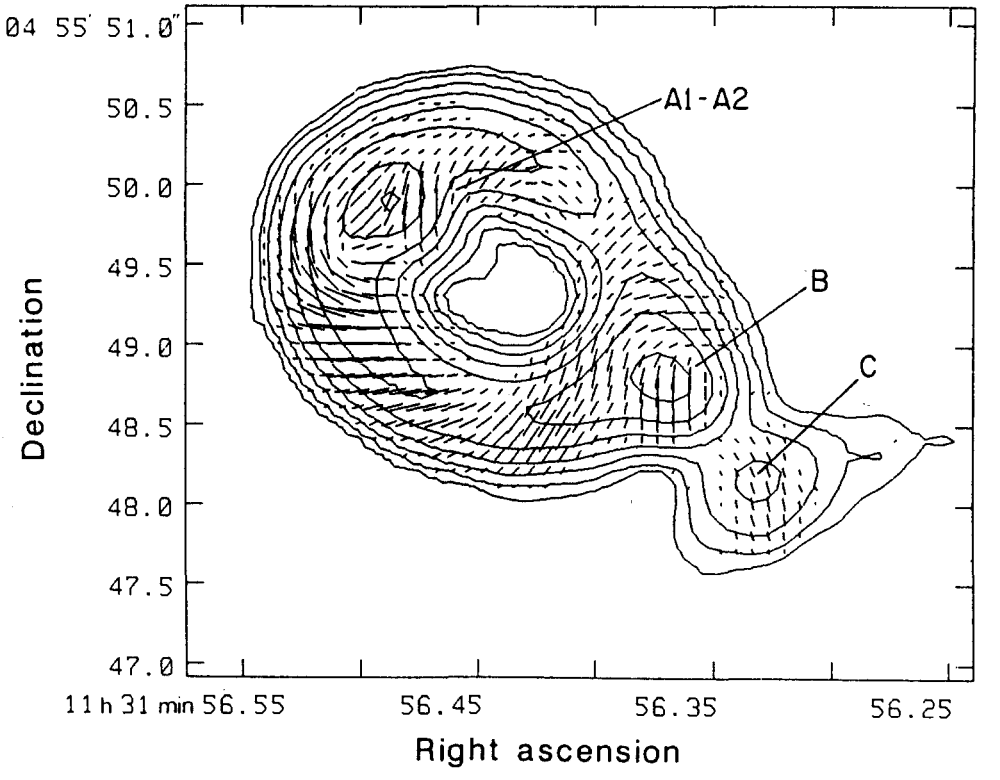

Fig. 8. A 6 cm radio map of the ring MG1131+0456 (from Hewitt et al. 1988)

6.2.1 MG1131+0456

The first ring discovered is MG1131+0456 (Hewitt et al. 1988), of which a 6 cm map is displayed in Fig.8. Besides the ring, the map shows several compact components (A, B and C in the figure). Optical information about this source is sparse. Its optical counterpart is a faint extended object, but it is unclear whether it corresponds to the radio source or to a possible lens galaxy. Spectroscopy reveals two spectral breaks

which were preliminarily assigned to redshifts of 0.85 and 1.13 and might indicate the redshifts of source and lens. The lack of any strong emission line shows that the source does not belong to a class of objects known in the Galaxy, and its radio spectral index as well as its optical color are compatible with the appearance of a radio galaxy.

The strongest argument for 1131+0456 being a gravitational lens system is derived from a theoretical study by Kochanek et al. (1988), who successfully tried to reconstruct the image from a gravitational lens model, which was chosen to be an elliptical galaxy. The applied reconstruction technique simultaneously yields the model parameters of the lens and the intrinsic intensity distribution of the source. It might appear on first sight that such a method could reproduce nearly any image morphology, but this is it not the case: since observations are available for two radio frequencies as well as for the flux polarization, the *same* lens model must apply to these three independent maps, and does indeed! (For instance, if the compact components A and B would not lie on opposite sides of the ring, but would subtend an angle of, say, $\pi/2$ as seen from the center of the ring, it would be impossible to reconstruct all three maps from a single gravitational lens model.) Moreover, the reconstructed source brightness distributions are typical for radio sources, strengthening the lens interpretation. Recently, Hammer & Le Fevre (1991) have applied deep optical imaging, the result indicating a ring-like structure with an opening in the ring coinciding in position with the opening seen in the 2 cm radio map, thus providing direct observational evidence for the lensing nature of this source.

6.2.2 MG1654+1346

An even better example for a radio ring is provided by the source MG1654+1346, which is displayed in Fig.9. Here, a radio galaxy is seen with its radio core – which conincides with the optical image of a QSO (Q) at a redshift of $z_s = 1.7$ – and its two radio lobes (B and C), one of which is deformed into a ring by the galaxy G, showing a redshift of $z_d = 0.25$. This gravitational lens system is ideal for determining the mass of the galaxy G inside the ring-shaped image.

Ring images in general constrain the lens model better than multiple point images, since the maps of extended images contain more information than just image position and magnification ratios of point images. The reconstruction technique mentioned above has been recently extended to account for finite resolution of the observations (Kochanek and Narayan 1992), providing a powerful tool for fitting lens models and reconstructing intrisic brightness profiles of sources. Still, one should remember that lens models obtained from ring systems are also not unique; the most robust parameters of such models are the lens mass inside the ring image, the ellipticity of the matter distribution and its major axis.

6.3 Luminous arcs

Although observed earlier, the announcement of R. Lynds and V. Petrosian in 1986 to have observed "a hitherto unknown type of spatially coherent extragalactic stucture", located in clusters of galaxies, "with narrow arc-like shape [and] enormous length", and independently by G. Soucail and coworkers in Toulouse, provided us with a new

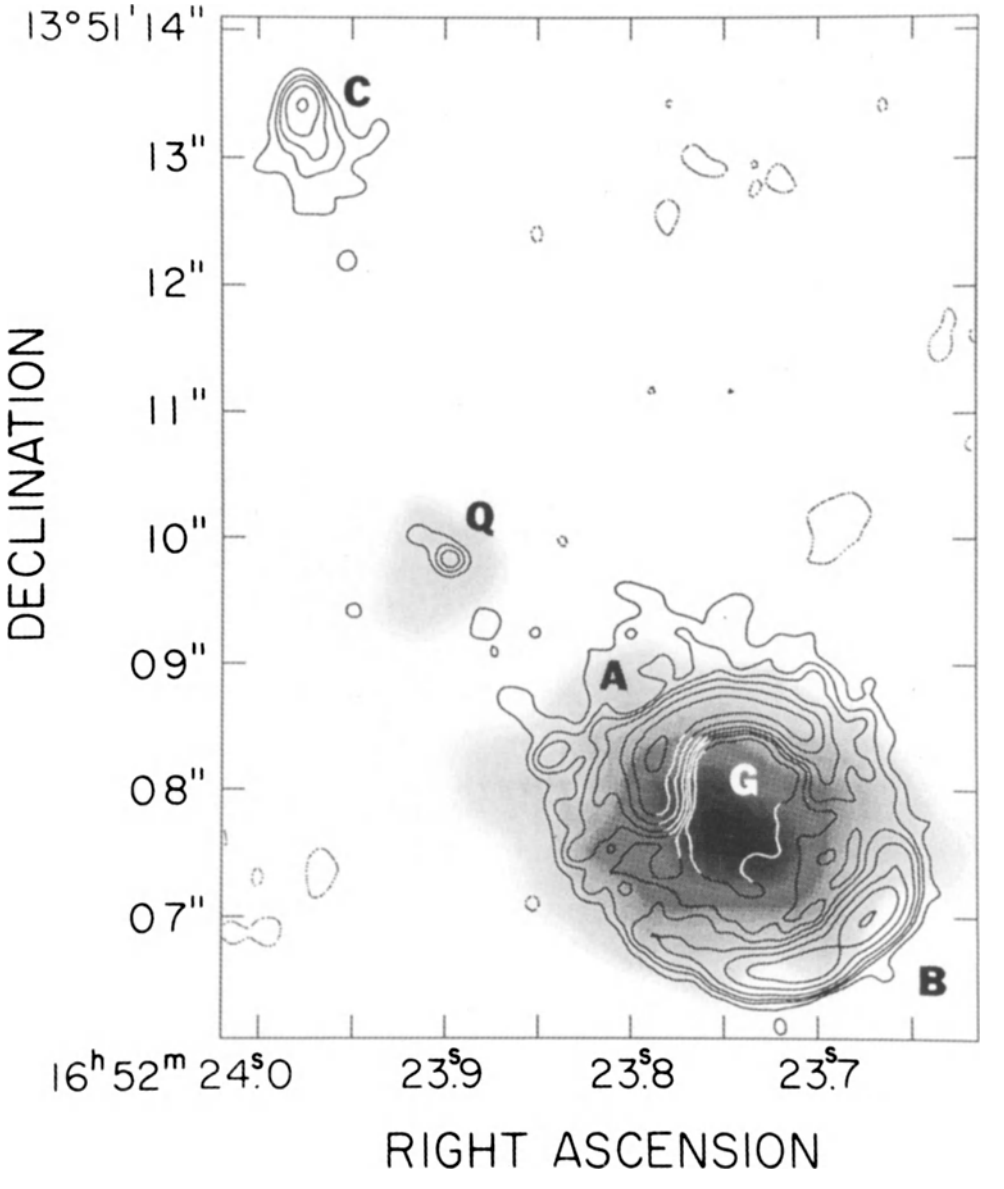

Fig. 9. The contours show a 8 GHz radio map of MG1654+1346, superimposed on an R-band optical image of the QSO Q and the galaxy G (from Langston et al. 1990)

type of gravitational lens systems. Whereas several different explanation for the occurrence of such arcs have been given initially, models which put the source of the arc in the cluster of galaxies in which they are observed were falsified once the redshift of the arc in A370 (Fig. 10) was measured to be $z_s = 0.724$, whereas the cluster has a redshift $z_d = 0.37$. Today we know about 10 such giant luminous arcs, A370 being the best studied of them.

Presently, the nature of the arcs is interpreted as galaxies at high redshift being gravitationally lensed by foreground clusters. The highly elongated shape of the images is due to the distortion by the lens mapping, with the source being situated close to a cusp singularity of the lens. (There are also "straight arcs", which are most conveniently interpreted as sources lying close to so-called beak-to-beak singularities, one of the types of metamorphoses mentioned in Sect. 5.) Note that the length of the arc in A370 is about 21 arcseconds, and its width is roughly 2 arcseconds, so that the deformation by lensing is indeed huge; correspondingly, the total flux of the image is much larger than that of the unlensed source. It is only this high magnification which allows spectroscopy of such intrinsically faint extended sources – clusters of galaxies forming arcs provide us with (cheap) 'natural telescopes'.

Detailed modeling of arcs in several clusters of galaxies has provided us with some crucial information about the mass distribution in clusters: First, the amount of dark matter, inferred previously from studies of the dynamics of cluster galaxies, agrees with that obtained from the lens models (as was the case for galaxy-mass lenses, lens models for arcs are far from unique, with the mass inside a circle of radius given by the distance of the arc from the center of the cluster being the most robust parameter). Second, the dark matter is not tied to individual galaxies, but spread more evenly throughout the clusters. Third, the mass inside clusters of galaxies must be distributed more compactly than previously thought; otherwise, clusters would not be able to form caustics and thus multiple images. Once a complete sample of clusters (selected, say, by their X-ray flux; it seems that the X-ray luminosity of a cluster is a good indicator of its lensing power) is scrutinized for the occurrence of arcs, statistical information about the mass spectrum of clusters, their core radii etc. will be available (see Bergmann 1992, Wu and Hammer 1992).

If a cluster is sufficiently compact to form such spectacular long arcs, it will also deform other background sources lying close to the direction to the center of the cluster. Examples of this effect can be seen in Fig. 10, where various elongated blue images are seen (A1 through A6). These images share the property that they are elongated in the tangential direction with respect to the cluster center. Such tangential elongation is expected from light deflection. Indeed, the redshift of component A5 has been measured to be $z_s = 1.305$ which thus stems from a background source. Since the density of faint background galaxies is very large (Tyson 1988), these sources can in principle be used to 'map' the mass distribution of compact clusters. This is not a trivial task, both from observation and theory. Very deep images have to be obtained in order for the source density to be sufficiently high. Sources (galaxies) are intrinsically not round, and therefore intrinsic ellipticity has to be disentangled from lens-induced deformation. This clearly is an ambitious statistical problem, treated in considerable depth in the literature (e.g., Kochanek 1990, Miralda-Escude 1991), but has been demonstrated to work in practice (Tyson, Valdes & Wenk 1990). In principle,

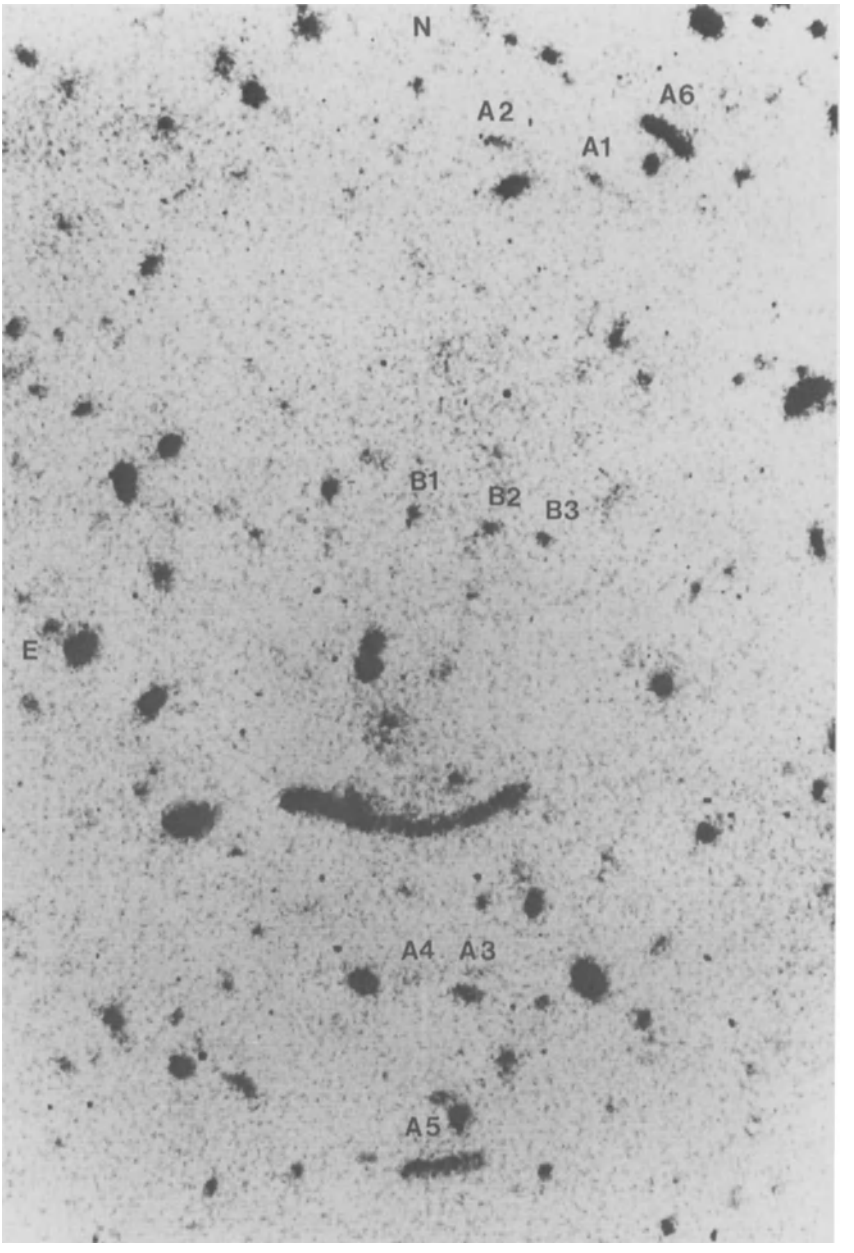

Fig. 10. The cluster of galaxies A370, with the luminous arc situated to the south of the center of this image, and various blue, elongated images, labeled A1 through A6 and B1 through B3. The redshift of the cluster is $z_d = 0.37$, the redshift of the arc is $z_s = 0.724$, and the arclet A5 has redshift $z_s = 1.305$ (courtesy of G. Soucail and B. Fort)

the redshift distribution of this faint background galaxy population can be obtained from such studies of arclets around a sample of clusters of galaxies.

6.4 Microlensing

If a source is observed through a galaxy, its light bundle does not traverse a smooth distribution of matter, but clumpy material (e.g., stars). If the light bundle is sufficiently small, it will 'feel' the graininess of the mass distribution. Whereas the deflection angle caused by an individual star (at its Einstein radius) is of the order of microarcseconds (this motivates the term 'microlensing') and thus unobservable, the distortion of the cross-sectional area of the light bundle by clumps leads to an observable magnification. Since the relative alignment of source and lensing galaxy changes in time, so does the magnification: microlensing leads to a lens-induced variability of sources seen through a galaxy. This effect, in general, is extremely difficult to distinguish from intrinsic variability of sources, but for multiply imaged QSOs, such microlensing could be observed: an intrinsic variation of the source will be seen in all images, with their respective time delay, whereas variability due to microlensing is uncorrelated between the individual images.

The best lens system for observing microlensing is again 2237+0305, owing to the small redshift of the lens (so that velocities in the lens plane correspond to large effective velocities in the source plane) and to the fact that the time delay of the images is of the order of one day, so that intrinsic flux variations of the source will be seen nearly simultaneously in all four images. In fact, uncorrelated flux variations in at least three images have been observed (Corrigan et al. 1991), ranging up to half a magnitude. This is a clear signature of microlensing.

It is relatively difficult to obtain quantitative conclusions from such microlensing observations. One can compare observed lightcurves with numerical simulations, but due to the statistical nature of the effect definite conclusions will only be possible if a sample of microlensing events is observed, which means that a sufficiently long data track of the source must be available. Nevertheless, the observations of microlensing in 2237+0305 have led to the following conclusions: the observed flux variations are compatible with the picture in which microlensing is produced by a normal stellar mass spectrum (Wambsganss, Paczynski & Schneider 1991). The size of the emission region of the optical continuum source of 2237+0305 is just compatible with a simple accretion disc model (Rauch & Blandford 1991, Jaroszynski, Wambsganss & Paczynski 1992); the observed variations lead to an upper limit of the source size.

In addition, microlensing can be used to obtain information about the brightness structure of sources. For example, the lightcurve of an extended source crossing a caustic is a convolution of its one-dimensional brightness profile and an universal function, the point-source magnification function near folds. Hence, from a well-observed high-magnification event in an observed microlensing light curve, the one-dimensional brightness profile of a source could in principle be reconstructed (Grieger 1990). The broad-line region in QSOs, probably too extended to be magnified as a whole by microlensing, has intrinsic structure. Therefore, differential magnification across the broad line region can affect the line profiles of the broad emission lines (Schneider & Wambsganss 1990). In fact, line profile differences in 2237+0305 have been observed by Fillipenko (1989). For further application of microlensing, see Sect. 12.4 of SEF.

7 Theoretical expectations for lens systems

It is reassuring that (nearly) all observed gravitational lens systems can be understood in terms of a single simple model for the deflector, namely an elliptical gravitational lens. Here, 'elliptical' means a fairly rough characterization of the property of the deflecting mass: for the reconstruction of observed lens systems, it basically does not matter whether mass distributions with elliptical isodensity contours, elliptical isopotential curves, or quasi-elliptical lenses are used (such as axially-symmetric lenses with a superposed quadrupole term). A we shall explain below, all these matter models share some generic properties which are essential for reconstructing observed lens systems.

Generic properties of lenses are best discussed by considering the caustic structure of the lens. For intuition, it is simpler to understand the caustic structure of axially-symmetric lenses; therefore, in Sect. 7.1 we shall discuss the qualitative behaviour of an axi-symmetric lens model. It must be kept in mind that axially-symmetric lenses have non-generic properties, so that they can hardly be used for applications in real lens systems. The generic properties of 'elliptical lenses' will be described in Sect. 7.2.

7.1 Disks as lenses

Consider a disk of radius ϱ, surface mass density $\Sigma(\xi)$ and mass $M = 2\pi \int_0^\varrho \Sigma(\xi)\xi d\xi$. We choose as the length scale in the lens plane the Einstein radius

$$\xi_0 = \sqrt{2R_\mathrm{S} \frac{D_\mathrm{d} D_\mathrm{ds}}{D_\mathrm{s}}}$$

of the disk's total mass; then its dimensionless radius is $x_0 = \varrho/\xi_0$, and we define $x := \xi/\xi_0$. In the following, we investigate in turn the cases of a homogeneous and an inhomogeneous disk; the latter provides a generic description for centrally condensed axially-symmetric lenses.

7.1.1 Homogeneous disk

First, we treat the homogeneous case, $\Sigma = \Sigma_0$. Then, $M = \pi\varrho^2\Sigma_0$, and the resulting lens mapping is

$$y = \begin{cases} x - (x/x_0^2) & \text{for } |x| \le x_0 \\ x - 1/x & \text{for } |x| \ge x_0 \end{cases} \quad . \tag{48}$$

It can easily be inverted. For $x_0 < 1$ (equivalent to $\Sigma_0 > \Sigma_\mathrm{cr}$, a source at y has three images if $0 \le y \le (1-x_0^2)/x_0$, one inside the disk at $x = x_0^2\, y/(x_0^2-1)$ and two outside, given by the corresponding solutions of the point-mass lens, $x = \frac{1}{2}\left(y \pm \sqrt{y^2+4}\right)$; two images if $y = (1-x_0^2)/x_0$, one at the rim of the disk, the other one outside, at x_0^{-1}; one image if $y > (1-x_0^2)/x_0$, at $x = \frac{1}{2}\left(y + \sqrt{y^2+4}\right)$. For $x_0 > 1$, there is one image only. It is inside the disk at $x = x_0^2\, y/(x_0^2-1)$ if $0 \le y \le (x_0^2-1)/x_0$, outside at $x = \frac{1}{2}\left(y + \sqrt{y^2+4}\right)$ otherwise.

Note that in each case the magnification of the image inside the disk is $(1-\kappa_0)^{-2}$, with $\kappa_0 = \Sigma_0/\Sigma_{\rm cr} = 1/x_0^2$. Its parity is always positive, and it corresponds to a maximum (minimum) of the Fermat potential if $x_0 < 1$ ($x_0 > 1$).

In the special case $x_0 = 1$, i.e., $\kappa_0 = 1$, all points of the disk are mapped into the point $y = 0$ on the optical axis. Thus, this "lens" indeed acts as a perfect thin lens, the focus of which is at the observer. For a fixed physical mass density Σ_0, the value of κ_0 depends on the distances. The value $\kappa_0 = 1$ corresponds to $\varrho = \xi_0$, i.e., to $2R_{\rm S}D_{\rm d}D_{\rm ds} = \varrho^2 D_{\rm s}$. This equation is (naturally) symmetrical in $D_{\rm d}$, $D_{\rm ds}$ and reduces, in the case of Euclidean distances when $D_{\rm s} = D_{\rm d} + D_{\rm ds}$, to the elementary formula $(1/D_{\rm d})+(1/D_{\rm ds}) = (1/f) =: (2R_{\rm S}/\varrho^2)$ for a thin lens of focal length f. If the observer is between the lens and the focus, there can be a single image of the source only, but if the focus is between the lens and the observer, three images are possible. The magnification of the image within the disk is large if the observer is near the focus.

7.1.2 Inhomogeneous disk

In the "interesting" case $x_0 < 1$, the circle of radius $y = (1 - x_0^2)/x_0$ separates the three-image region from the single-image region. The corresponding circle with radius x_0 in the lens plane – the rim of the disk – is the curve where images fuse or are created. However, this circle is not a critical curve in the strict sense of catastrophe theory since the mapping given by (48) is not differentiable there. But suppose we smooth the edge of the disk so that $\kappa(x) = \Sigma(x\xi_0)/\Sigma_{\rm cr}$ becomes differentiable. Then there is a point $x_{\rm c}$ close to x_0 where (dy/dx) vanishes, and the corresponding circle in the source plane is indeed a fold-type caustic. Let the smoothing be done in such a way that, (i) a central region $|x| \le x_1$ of the disk remains homogeneous, and (ii) for positive x, the scaled deflection angle $\alpha(x)$ has only one maximum, as in the homogeneous case. Inhomogeneous disks of this kind form an open set within the collection of axially symmetric lenses. Their qualitative properties will now be described. We put

$$s = \sqrt{\frac{2R_{\rm S}D_{\rm d}}{\varrho^2}} \quad , \qquad p = \sqrt{\frac{D_{\rm ds}}{D_{\rm s}}} \quad , \tag{49}$$

so that $0 < p < 1$ and $x_0^{-1} = ps < s$. Note that $(2R_{\rm S}/\varrho)$ is a measure of the compactness of the lens, s increases with the distance of the lens from the observer, and for fixed s, p increases with the distance of the source from the lens. The following cases are possible:

1) $s \le 1$. Then $x_0 > 1$ and, as in the homogeneous case, there are no critical points. The observer is so close to the lens that, even for distant sources, the light deflection does not suffice to produce more than one image; the past light cone of the observer is smooth, i.e., free of caustics; see Fig. 11a,b.

2) $s > 1$. This case is illustrated in Fig. 11c. The large caustic C is indicated as a thick line. In three-space, it consists of a conical surface M and an axis A, a half-line meeting M at the vertex V. Sources so close to the lens that $ps < 1$, have one image only; the corresponding source planes do not intersect C. If $ps = 1$, $x_0 = 1$; the source plane intersects C in the vertex V only. The corresponding critical set is the whole homogeneous "core" $|x| \le x_1$ of the disk; it consists of degenerate critical circles all

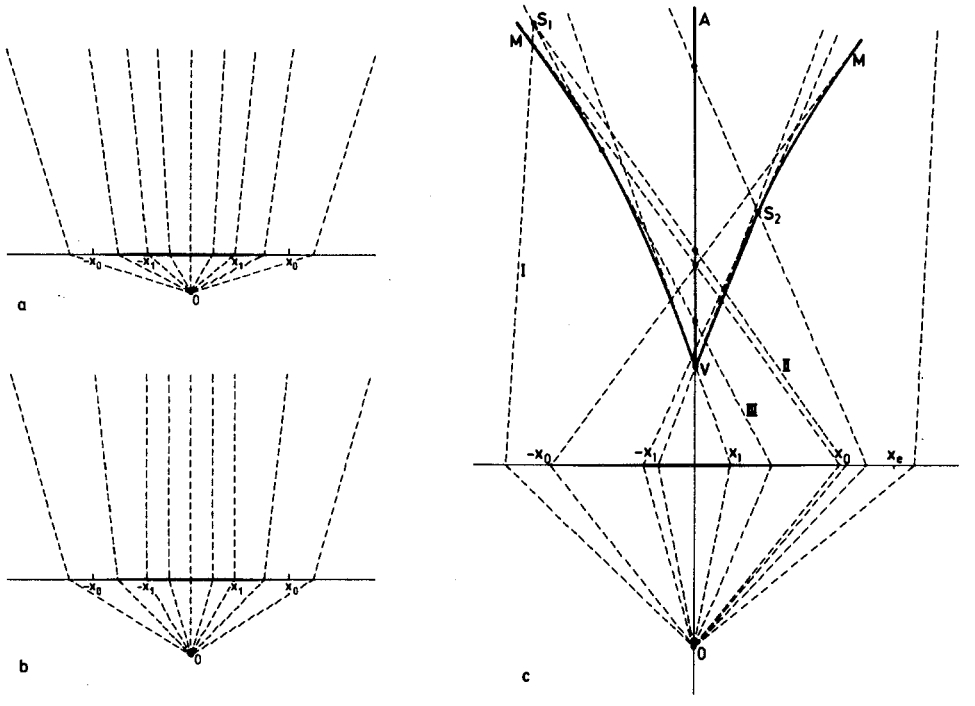

Fig. 11. Light bending by an inhomogeneous disk. (a) For observers close to the disk, there is no caustic. Followed backwards from O, all rays diverge. (b) Rays from a source "at infinity" are focused at the observer by the homogeneous inner part of the disk. (c) For observers at larger distances than in case (b), the large caustic has the shape of an inverted tent with a pole. Details are explained in the text

mapped into V. If $ps > 1$, $x_0 < 1$, and C intersects the source plane in a point (the image of tangential critical circles in the lens plane) and a circle (the image of a radial critical circle).[2] If we consider a sequence of point sources on the optical axis starting closely behind the lens, the observer will see a point until the source has reached V, when suddenly the source is seen as a disk. Sources at larger distances produce a central point-image and an Einstein ring, first inside, then outside the disk.

It is also instructive to locate the points conjugate to O on all light rays reaching O. Let $\mathbf{x}'$ denote the point at which a ray crosses the lens plane. Then, if

$$|\mathbf{x}'| \geq x_e := \left[\frac{2D_s}{(D_s - D_d)} \int_0^{x_0} x\kappa(x)dx\right]^{1/2} ,$$

[2] The determinant of the Jacobian matrix (46) reduces for axially-symmetric lenses to $\left(\frac{y}{x}\right)\left(\frac{dy}{dx}\right)$. Critical curves for which the first factor vanishes are called tangential critical circles, those where the second vanishes radial critical circles. For the geometrical interpretation of these terms, see Sect. 8.1.1 of SEF.

the ray does not contain a point conjugate to O; if $x_0 \leq |\mathbf{x}'| < x_e$, it contains exactly one (simple) conjugate point, which is located on the axis A; if $x_1 < |\mathbf{x}'| < x_0$, the ray contains two (simple) conjugate points, one on A and one where the ray touches M; and if $|\mathbf{x}'| = x_1$, the ray contains only one (degenerate) conjugate point, the vertex V of the large caustic. Clearly, the latter is the set of points conjugate to O on some ray through O. (Note that the preceding description holds for the rays and the caustic in spacetime although the figure and the discussion refer to its projection into the three-space.)

The discussion just given remains valid if the homogeneous inner disk shrinks and is removed, $x_1 \to 0$. Then M has a cusp, not only a vertex, at V, and only the central ray passes V. An example is provided by the disk representing a homogeneous, spherical mass.

7.2 'Elliptical' lenses

As we mentioned in the introduction to this section, for modeling observed gravitational lens system, a generic set of lens models sufficies for (nearly) all observed systems (with the exception of 2016+112, where multiple light deflection must probably be taken into account). The types of 'elliptical' lenses described above work equally well in reconstructing images, because the number of observational constraints is fairly small: for multiply imaged QSOs, one has to match the image positions, their flux ratios and, if the source has a compact resolved radio structure (such as for 0957+561), the relative magnification matrices. Adjusting the free parameters of a lens model from each of the classes mentioned allows a match of observed and theoretical properties of the lens system. On the other hand, the observed ring images provide more constraints on a lens model; however, ring images typically probe the lens mapping only close to the critical curve of the lens and is thus sensitive only to a small number of properties of the mass distribution, which can be matched by any of the models listed above. These remarks imply good and bad news: the good news is that our ideas about the rough mass distribution in galaxies are in accord with gravitational lens observations; the bad news is that, without additional information, multiply imaged QSOs and radio rings do not allow to investigate the mass distribution inside lensing galaxies in any detail.

7.2.1 Evolution of the caustic structure

Consider a matter distribution with fixed surface mass density $\Sigma(\boldsymbol{\xi})$, and consider the separation between lens and source as a variable quantity; e.g., let $p = D_{ds}/D_s$. We assume that the mass distribution is of one of the types mentioned above, and that Σ is finite everywhere and decreases outwards.

For sufficiently small p, the lens is too weak to produce critical curves, and thus multiple images. For $p = p_1$, the Jacobian matrix at the center of the lens will have a zero eigenvalue, whereas the second eigenvalue is positive. This characterizes a lips catastrophe, one of the metamorphoses mentioned in Sect. 5. For increasing p, the size of the lips caustic grows, and the Jacobian matrix at the center has one eigenvalue of either sign. At $p = p_2$, the larger of the two eigenvalues becomes zero, leading again to a lips catastrophe. For $p \gtrsim p_2$, there are now two lips-shaped caustics, one inside

the other, and oriented mutually perpendicularly. Increasing p further, the size of the inner caustic grows, i.e., its two cusps approach the outer caustic. At $p = p_3$, the cusps 'touch' the outer caustic, and at these two points, hyperbolic umbilics occur (another type of metamorphoses). For $p > p_3$, there are two closed caustic curves, one with four and one with no cusps; we will call the first one the "tangential" caustic (since this caustic occurs if the tangential critical circle of a symmetric lens is slightly perturbed) and the latter one "radial" caustic. For $p > p_3$, no more metamorphoses occur in general ; nevertheless, the qualitative morphology of the caustics changes for increasing p: for $p \gtrsim p_3$, the two cusps which were formed during the hyperbolic umbilic transition lie inside the radial caustic, whereas the other two cusps of the tangential caustic lie outside the radial one. At $p = p_4$, these two latter cusps cross the radial caustic, so that for $p > p_4$ the tangential caustic lies completely inside the radial one. This is the case which is most relevant for modeling the observed multiple image QSOs and ring systems.

If a cusp lies outside the radial caustic, a source close to it will have one or three highly-magnified images, without any additional images; we term such a cusp "naked". Naked cusps probably are relevant for explaining some of the luminous arcs: the generic model for the occurrence of arcs is that of an extended source just inside a cusp. In order to avoid additional images of the corresponding source, the cusp must be naked. Straight arcs can occur if an extended source lies within a lips caustic.

7.2.2 Imaging properties

It is usually assumed that the central part of galaxies is sufficiently compact, so that for typical lens and source redshifts, $\kappa \gg 1$ at the center of the lens. In fact, the observational evidence for the occurrence of even number of images (in contrast to the expected odd number) supports this view, as it is usually assumed that the 'missing' image is situated close to the center of the lens, where the surface mass density is so large as to cause a strong demagnification of the corresponding image. Assuming this to be the case, the typical caustic of an "elliptical lens" is that corresponding to the case $p > p_4$, i.e., where the tangential caustic with its four cusps lies entirely inside the radial caustic. We consider this case now in somewhat more detail; see also Fig. 12.

If a source is situated close to the center of the source plane (like the cross in Fig. 12a), it will have five images, one of which will be close to the center of the lens and thus strongly demagnified, whereas the other four will lie more or less symmetrically around the lens center. The more symmetric the image arrangement is, the less are the differences in their magnifications. The two GL systems 2237+0305 and 1413+117 are probably typical examples for such a lensing configuration. Moving the source closer to the tangential caustic, but not close to one of its cusps (like the triangle in Fig. 12c), two of the four images will come closer together and become brighter. The GL system 1115+080, and probably also 0414+053 can be accounted for with such a lensing geometry.

If the source lies close to, but on the inside of a cusp (the triangle in Fig. 12a), three bright images will be very close together; depending on the resolution of the observations, such a triple image is hard to resolve, so that it appears to an observer as a single, very bright image. It is unclear at the moment whether such a system has

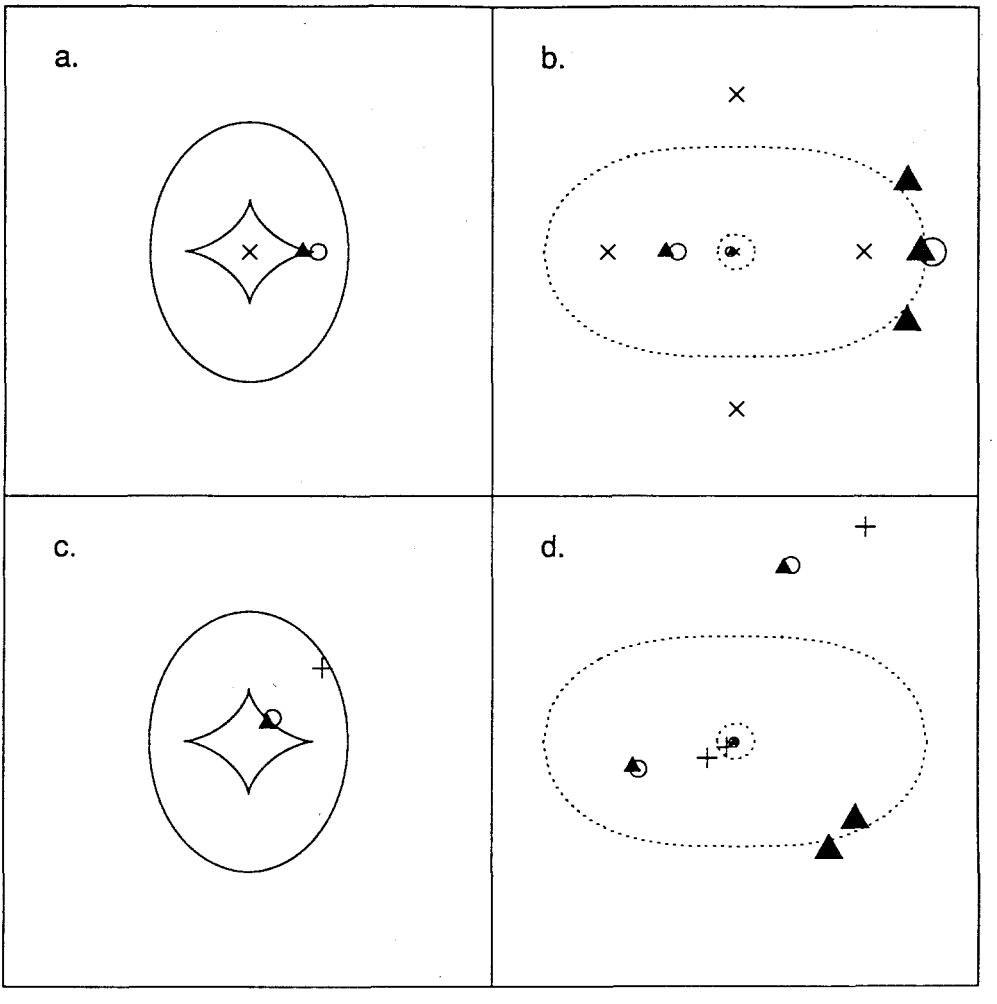

Fig. 12. The caustic curves (a) and (c) in the source plane and critical curves (b) and (d) in the lens plane of a generic 'elliptical' lens, with $p > p_4$ (see text). The symbols in the source plane indicate three different source positions in each of the frames (a) and (c), with the corresponding image positions plotted by the respective symbols in (b) and (d). The relative size of the symbols in the lens plane should indicate the magnification of the images (from Blandford and Narayan 1992)

been found already. Due to the large overall magnification, such GL systems should be included in flux-limited samples preferentially (see the discussion on amplification bias in the following section), and one might therefore suspect that some of the GL candidate systems with large flux ratios of the images (most noticibly, UM425=1120+019) can be accounted for by such a lens arrangement. In all the three cases just described,

the fifth image is usually much fainter than the other images, if the lens is sufficiently compact, and can therefore easily escape detection.

Consider next the case that a source lies outside the tangential caustic, but inside the radial caustic. If it is not too close to the latter (like the circle in Fig. 12c), two of its three images will have comparable magnification, whereas the third close to the center of the lens will still be strongly demagnified. This is the lensing geometry which is expected to occur most frequently, due to its relatively large probability. The existing sample of GL systems, however, does not contain a large fraction of such systems (one example is 0957+561); we interpret this discrepancy as being due to strong selection biases, to be explained in more detail in Sect. 8. Briefly, optical QSO surveys bias against GL systems with two images of comparable brightness, since one of the selection criteria is the stellar appearance of the source (Kochanek 1991). Convolving a pair of images of comparable brightness with the seeing disc, it appears elliptically, therefore extended, and will be excluded from further examination (by spectroscopic means). Furthermore, the total magnification of such systems is much smaller than for the cases described above, and thus, the amplification bias is less effective for such GL systems than for the 4-image cases.

Moving the source closer to the radial caustic (like the cross in Fig. 12c), the image with negative parity moves towards the radial critical curve and becomes weaker. Depending on the separation from the radial caustic, the third image can become of comparable brightness to the negative parity image, but for this to occur, these two images must be fairly close together, so that they are difficult to resolve observationally [note that a pair of images near a radial critical curve need not necessarily be highly magnified; whereas the scaling $\mu \propto (\Delta\theta)^{-1}$ is of course also valid for radial critical curves, the constant of proportionality can be fairly small for realistic lens models]. If the double nature of the B image in 2345+007 can be confirmed, it can be accounted for by such an arrangement. Also, the two GL candidates 1635+267 and 0023+171 can be of that type.

The foregoing discussion has shown that a single "elliptical lens" can account for nearly all types of observed gravitationally lensed QSOs (and it can also account for the observed radio rings). From that we conclude that the existing GL systems pose no problem in understanding their qualitative features (again, except for 2016+112) with a simple lens model, although the lens parameters required from observations, in particular the lens mass, appear in some cases to be fairly unusual. On the other hand, since most multiple QSO gravitational lens systems are so easily reproduced by simple models, this modeling procedure can not infer much information about the matter distribution of the lenses; the only solid model parameters are the mass inside a circle traced by the observed images and the 'ellipticity' of the lens.

8 Further applications

In this final section, we want to present some additional applications of gravitational lensing, of vastly different scale: from the detection of dark matter in our Galaxy to the large-scale structure of the universe.

8.1 The size of Lyα clouds

Each QSO at sufficiently high redshift, such that the Lyα emission line can be observed from Earth, shows a large number of narrow absorption lines shortward of the Lyα emission line. This so-called 'forest' is interpreted as being due to intervening material in the line-of-sight to the QSO, and the absorption lines being due to the Lyα transition. It is less clear what the nature of the absorber is; one can measure the redshift, equivalent width (yielding the column density of neutral hydrogen) and line width. Gravitational lens systems provide us with an invaluable tool to determine the size of this absorbing material.

Consider a lens system in which a high-redshift QSO has two observable images. The transverse separation between the corresponding light bundles varies as a function of redshift: it is zero at the source, and largest at the redshift of the lens. Suppose an absorption line is seen in the spectra of both images, at the same redshift; then the size of the absorbing material must be at least as large as the separation between the two light bundles at the respective redshift. Hence, redshift coincidences of absorption lines leads to a lower bound on the size of the absorbing material. If, in addition, the equivalent widths of the corresponding lines in both spectra are strongly correlated, it could be concluded that the two light bundles actually have crossed the same cloud (as opposed to crossing two different clouds at the same redshift).

Two lens systems have turned out to be useful for such studies. The lensing nature of one of them, 2345+007, however, has not yet been totally clarified, but it is an excellent candidate system. The large angular separation of 7.3 arcseconds makes this a very useful target for these spectroscopic investigations, although the faintness of the QSO images renders such an investigation difficult. A second system, UM673 with an image separation of 2.2 arcseconds, is brighter and thus higher-quality spectroscopy can be obtained. Results of such a study are reported in Smette et al. (1992).

8.2 Dark matter in our Galaxy

Observations of the rotation curve of our Galaxy and that of other spiral galaxies indicate the presence of a halo composed of dark matter. It is unclear what the nature of this dark matter is; candidates are: weakly interacting elementary particles, brown dwarfs and 'Jupiters', or black holes. If the dark matter such of such compact objects, these can lead to lensing effects of background sources: for stellar mass lenses, the corresponding angular separation of split images would be much too small to be detectable, but the magnification could be observed. The problem with this idea is that the probability of finding a single lens sufficiently close (within its Einstein radius) to the line of sight of an extragalactic object is about 10^{-6}. Hence, for observing this effect a huge number of source must be photometrically monitored.

Paczynski (1986) suggested monitoring of the stars of a nearby galaxy, the Large Magellanic Cloud or M31. In fact, two groups are currently carrying out such an observational program (see the corresponding contributions in Kayser & Schramm 1992). The difficulties for such a program are enormous; we want to mention only a few of them. First, many stars are intrinsically variable, and it will be a major task to separate intrinsic variability from lens-induced one. A magnification event will be a 'once in a lifetime' event, that is, events will not recur. The program will work by

comparing the brightnesses of sources from consecutive observations. The required large number of stars implies that the amount of data produced in the course of the program will be huge.

On the other hand, such an experiment is probably the only way to clarify the nature of the dark matter in the halo of our Galaxy. It will be sensitive to compact objects of masses between $10^{-6} M_\odot$ and about $1 M_\odot$. A calibration experiment, also suggested by Paczynski (1991), can be used to test the sensitivity of the observational program, including the software for data analysis. Whatever the outcome of the program will be, it certainly will produce the most useful database for stellar variability and is thus worth the effort.

8.3 Amplification bias

8.3.1 General ideas

In a completely smooth universe, there is a unique relation between the luminosity L of a source and the flux S we observe from this source, $S = L/(4\pi D_L^2)$, where $D_L(z)$ is the luminosity distance to a redshift z, which can be directly computed from the parameters Ω and H_0 of the cosmological model. However, our universe is not homogeneous, at least not on small scales. In a locally inhomogeneous universe (or clumpy universe), the flux observed from a source depends, besides its luminosity and redshift, also on the propagation of its light bundle through the universe, which is affected by the local inhomogeneities. Since gravitational light deflection can affect the flux from a source, the relation between flux and luminosity has to be modified, $S = \mu\, L/(4\pi D_L^2)$, where μ is the magnification caused by light deflection. Since, in general, the magnification of any given source is unknown, the best one can hope for is to obtain a probability distribution $p(\mu, z)$ for the magnification. Suppose we knew such a probability distribution. Now, let $\Phi(L, z)$ be the number density of sources at redshift z with luminosity greater than L. In a completely smooth universe, the number counts of these sources, i.e., the number of sources $n_0(S, z)$ per unit redshift interval at z and with flux greater than S would be

$$n_0(S, z) = C(z)\, \Phi(4\pi D_L^2\, S, z) \quad , \tag{50a}$$

where $C(z)$ is a constant volume factor. In a clumpy universe, with given $p(\mu, z)$, the source counts $n(S, z)$ would be

$$n(S, z) = C\, \frac{1}{\langle\mu\rangle} \int d\mu\; p(\mu, z)\; \Phi(4\pi D_L^2\, S/\mu, z) \quad , \tag{50b}$$

and C is the same as in (50a); $\langle\mu\rangle\,(z)$ is the mean magnification, which is determined from overall flux conservation (concerning this point, see Ehlers & Schneider 1986). Eq.(50b) implies that even from well-measured source counts, the intrinsic luminosity function cannot be directly inferred. In particular, it implies that in a flux-limited sample of sources, there may be some sources which are magnified above the flux threshold of the sample, and which, without this magnification, would be too faint to enter the sample. This effect is called "amplification bias", and it can severely affect the source counts of compact extragalactic sources (presumably mostly relevant to QSOs and BL Lac objects). Whether or not the amplification bias is important

depends on $p(\mu, z)$, which in turn depends on the population of lenses in the universe, the redshift and the size of the sources, and on their intrinsic luminosity function: the steeper $\Phi(L, z)$, the more biasing can occur.

Equation (50b) can be applied to the whole sky, and thus the overall source counts can be affected. It can also be applied to certain regions in the sky; an example of this will be given below. Whether or not the amplification bias is relevant for our observations of QSOs cannot be decided by theory only; however, there are at least three observational results which strongly suggest that it is indeed the case.

8.3.2 Evidence for the amplification bias

The probability for any given high-redshift source to be multiply imaged is less than 1% (as already estimated by F. Zwicky in 1937!). This does not mean, however, that the fraction of multiply imaged sources in flux-limited samples is as small as this — due to the amplification bias, this fraction can be considerably larger, as magnification and image splitting are not mutually independent (e.g., if the amplification bias is due to microlenses in galaxies, the galaxy hosting the magnifying stars can cause multiple images).

As argued above, the effect of the amplification bias increases with increasing steepness of the luminosity function. Since the luminosity function for faint QSOs is flat, whereas it is steep for luminous QSOs, one expects the amplification bias to be most effective for apparently luminous QSOs. Taking these two arguments together, one would therefore expect, if the amplification bias applies to our source counts, to find more multiply imaged QSOs among the apparently more luminous sources than for fainter ones. In fact, this has been recently verified observationally (for a review and references, see Surdej 1990): in two samples of apparently luminous QSOs, the fraction of "interesting" objects (those which show either multiple optical components, or elongated structure, or fuzz around the QSO image) is greater than 20%, and the fraction of "good lensing candidates" is about 5%. This has to be compared with the theoretical fraction given above, and the low fraction of multiply imaged sources found in fainter optical QSO samples and in samples of radio sources (Burke 1990). In addition, since the QSOs in the samples are at high redshift, the fuzz observed around many of them cannot be due to the host galaxy, unless it would be unusually luminous; rather, it is supposed (although not verified) that the fuzz is due to a foreground galaxy, lying just on top of the QSO and thus being a potential lens and magnifier. It appears that these observational results, if they stand the test of time, can only be interpreted if one assumes that the source counts of QSOs at the bright end are substantially affected by the amplification bias!

If a substantial fraction of the apparently bright QSOs are magnified, one should expect to see the lens (or matter associated with it) along the line-of-sight to some of these QSOs. One possible way to detect this is to count the number of galaxies around bright, high-redshift QSOs, and see whether they are overdense. In fact, several groups have reported on an overdensity of galaxies around high-redshift QSOs (for a summarizing discussion, see Narayan 1992). The observational situation is not very clear yet, owing to the extreme difficulties encountered by these observations.

As a third evidence for the influence of the amplification bias on source counts, we want to mention the results of M. Stickel and his collaborators (Stickel, Fried & Kühr

1988a,b, 1989): out of only a few tens of BL Lac objects with redshift measured from their (weak) broad emission lines, three have a foreground galaxy just on top of them (i.e., with misalignment less than 1 arcsecond). Although the statistical significance of this result cannot be easily estimated, these close alignments seem to clearly point towards magnification of the underlying source by the foreground galaxy. Whether or not this provides sufficient evidence for the suggestion by Ostriker & Vietri (1985), that at least part of the BL Lac population is due to microlensing of the optical continuum of an optically violently variable QSO (OVV) (so that the continuum is magnified to outshine the broad emission line flux, and thus turn the OVV into a BL Lac object), is unclear. Certainly, lensing does not avoid the necessity to have relativistic beaming in these sources, e.g., because of the observed superluminal motion. Some of the close alignments of BL Lac objects and foreground galaxies can be used to constrain the core size of the galaxies (Narayan & Schneider 1990).

Finally, direct evidence for amplification bias comes from detailed optical imaging of high-redshift ($z > 1$) 3CR radio galaxies (Hammer & Le Fevre 1990, and references therein). Out of the 31 such objects, 27 have high-quality optical images made; of those, $\sim 75\%$ have multiple optical components. Four of the sources have a foreground galaxy within 5 arcsec, where the hypothesis of a chance projection would predict 0.5 such associations. Another 5 of these sources have a foreground Abell and/or Zwicky cluster in their line-of-sight. Two of these sources are likely multiple-image lens candidates. From a case by case analysis, those four sources with a foreground galaxy can be shown to be significantly ($\gtrsim 0.5$ mag) magnified, and the same is true for at least one source with a foreground cluster. In addition, for at least three of these 3CR radio galaxies, the magnification of their radio flux can be estimated to be such that without the magnification, these sources would not have been included in the 3CR catalog. Hence, for this class of objects, the amplification bias works at least at the 10% level; however, this value can be considered to be a fairly conservative lower limit. If one considers the fact that these radio sources are fairly extended (and therefore, magnification can occur only by fairly massive lenses), and that the luminosity function of these sources is probably flatter than that of QSOs, the above value for the amplification bias is quite surprising and suggests that the source counts of QSOs are severely affected by the amplification bias.

8.4 Light propagation in inhomogeneous universes

As mentioned before, the inhomogeneity of the universe on scales much smaller than the horizon affects the light propagating from distant sources to us. The most obvious consequence of this effect is the occurrence of multiply imaged QSOs, arcs and rings. For these impressive systems to occur, a massive, compact mass distribution must be close to the line of sight to the source (here, 'compact' means that the central surface mass density is at least of the order of $\Sigma_{\rm cr}$). Since the density of such compact objects is probably very small (i.e., the probability, or 'optical depth', for image splitting on scales of an arcsecond is very small), most light bundles from distant sources will be subject to much subtler effects. Nevertheless, every light bundle will be affected by large-scale matter inhomogeneities in the universe.

The description of light propagation in an inhomogeneous universe can proceed in different ways, depending on which effects one wants to investigate. We will distinguish

between two major descriptions which we term 'clumpy universe' and 'LSS universe', in which the Large Scale Structure of the matter distribution is taken into account.

The clumpy universe description is the one mostly used in gravitational lens theory. It is constructed in the following way: a fraction $(1-\tilde{\alpha})$ of the total matter density of the universe is concentrated in compact clumps, whereas the remaining fraction $\tilde{\alpha}$ is homogeneously distributed (thus the term 'smoothness parameter' $\tilde{\alpha}$). The basic assumption underlying this construction is that the large-scale (or smoothed-out) metric of such a universe is that of a smooth Friedmann–Lemaître universe, and the metric deviations are only local. Support for the validity of this assumption has been given by Futamase & Sasaki (1989). In the LSS universe, the mass distribution is constructed from density fluctuations which evolve owing to gravitational instability. This is done either from sophisticated simulations of the large-scale structure (e.g., N-body codes, Zeldovich approximation) or by using some kind of analytic approximation, such as linear perturbation theory. Superposed on these smooth inhomogeneities, one can account for the presence of compact lenses, such as galaxies. Also in this model, one has to assume that the average metric of the universe is a smooth Friedmann–Lemaître metric, and that metric perturbations are local.

The latter condition which applies in both models can best be described as the 'area-fitting' condition: the two-dimensional surface of sources with constant redshift must have the same area as the corresponding sphere in a smooth Friedmann–Lemaître universe. The immediate consequence of this condition is flux conservation: consider a population of sources all with the same luminosity. Since photons are neither destroyed nor created by the process of gravitational light deflection (Sect. 3), the mean flux of all such sources with constant redshift must equal the flux that one would measure in a smooth Friedmann–Lemaître universe (see SEF, Sect. 4.4.5 for a more detailed discussion of this point).

The Dyer & Roeder (1973) description of light propagation in a clumpy universe proceeds as follows: light bundles which propagate far from all clumps are assumed to be unaffected by the tidal forces produced by the inhomogeneities. This assumption, which is at best approximately valid, defines regions in the universe termed 'empty cones'; a light bundle propagating through such an empty cone is affected only by the local matter density in the beam, given by $\tilde{\alpha}$ times the mean cosmic density. If a source is observed through an empty cone, its flux is smaller than it would be if it was observed in a smooth Friedmann–Lemaître universe, since the matter in the light bundle is reduced, i.e., the light bundle experiences less Ricci focusing. The gravitational deflection by clumps for those light bundles which pass by closely is taken into account explicitly using the lens equation. For example, if a single clump is close to the line of sight to a distant QSO, it is assumed that the light bundles propagate through empty cones (i.e., are unaffected by shear from other clumps) from the source to the lens and from the lens to the observer. Hence, the angular diameter distances entering the lens equation are those obtained from the empty cone – the Dyer–Roeder distances (see SEF, Sect. 4.5) – and therefore, the magnification obtained from the lens equation via (34) is the magnification relative to the case in which the source is observed through an empty cone, *and not* relative to the case were it is observed through a corresponding smooth Friedmann–Lemaître universe. The magnification theorem stated in Sect. 3 is to be interpreted such that no source

can be demagnified relative to the case where the source is seen through an empty cone.

The assumption underlying the empty cone approximation can be studied numerically: a clumpy universe is modeled on the computer, and the propagation of light bundles through such a matter distribution is studied, either using the optical scalar equations or multiple deflection gravitational lens theory (the latter approach is much more effective and provides a superb approximation for all relevant light bundles!); see Sect. 11.4 of SEF for references. One of the main results of such studies is that if the density of clumps in the universe approaches the critical (or closure) density, empty cones are rare, but for the probably more realistic assumption that the clumps do not constitue a major fraction of the critical density, the empty cone approximation seems to be well justified in most cases.

In the case of an LSS model, one cannot define empty cones: along each light bundle, the local matter density will vary as a function of redshift. Therefore, such a model does not allow to define a 'reference' such as the empty cone in a clumpy universe. One thus proceeds to normalize everything with respect to the smooth Friedmann–Lemaître universe, allowing matter inhomogeneities to have either sign. This implies that the Ricci focusing term can have either sign, and thus the magnification can be larger and smaller than unity *relative to the smooth Friedmann–Lemaître universe*, which is the only reference available. Hence, in this context the magnification theorem mentioned in Sect. 4 is no longer valid, since its proof is based on the fact that deflecting matter has nonnegative surface mass density relative to the reference universe (which determines which angular diameter distances have to be used).

Light propagation through clumpy and LSS universe models has been studied by various authors (see BN, Sect. 5.4, and references therein). We mention only a few aspects here.

The number of multiply imaged high-redshift sources yields an estimate of the number density of lenses which can lead to the necessary image splitting. Owing to the amplification bias, which is fairly uncertain to correct for, in most cases only upper bounds on the density of lenses can be given. Analysis of VLBI observations of compact radio sources have set an upper limit to the density of compact objects in the mass range of $10^7 M_\odot \lesssim M \lesssim 10^9 M_\odot$ of $\Omega \lesssim 0.1$, and the VLBA under construction will be able to improve this upper bound considerably (Kassiola, Kovner & Blandford 1991). Lower masses can be probed by higher resolution VLBI observations; it was suggested recently that the comparison of the VLBI maps of the two compact components in 0957+561 could reveal the presence of massive black holes of masses of around $10^6 M_\odot$ and slightly lower (Wambsganss & Paczynski 1992). Smaller masses, of $1 M_\odot$ say, reveal themselves by microlensing, i.e., a time-variable magnification of some compact sources could be detected in principle; unfortunately, however, it will be basically impossible to distinguish this lens induced variability from intrinsic effects.

Weak lensing effects influence the light bundle of every high-redshift source. To first order, a round source will have an elliptical image, caused by the tidal gravitational forces (Weyl focusing, or shear). The faint blue galaxy population mentioned in Sect. 6.3 above provides a potential tool for investigating this effect. If the scale of the matter inhomogeneities causing these tidal distortions of images is much larger than the mean separation between those galaxies, the distortion should be coherent, that is, the direction of the distortion should be the same for neighbouring sources.

However, sources are intrinsically elliptic, and thus for the detection of these coherent effects it is essential to separate intrinsic ellipticity from propagation effects; this can only be done by assuming that the orientation of the intrinsic ellipticity is random. For a detailed study of these effects, see Blandford et al. (1991).

Luminous high-redshift, compact radio sources are spatially correlated with bright (Lick) galaxies, as claimed by Fugmann (1990). If this result is statistically significant, it can most likely only be understood in terms of the amplification bias. However, the correlation takes place on angular scales of tens of arcminutes, much too large for individual galaxies to be responsible for the magnification. In a recent investigation (Bartelmann & Schneider 1992) we have found that only lensing by large scale matter inhomogeneities can be responsible for causing such a correlation: overdensities of matter in a certain region of the sky, which can yield magnification of background sources, is associated with an excess of galaxies, which are assumed to have formed in the density peaks of the cosmic matter distribution. A more detailed investigation of such large scale correlations between galaxies and high-redshift source is planned; it may allow to gain significant information on the shape of the large scale structure density spectrum and on the correlation between luminous and dark matter (the so-called biasing factor).

Finally, we would like to note that if cosmic strings exist, they will be most likely detected by their gravitational lens action (see BN, Sect. 5.4, and SEF, Sect. 13.3.4).

References

Bartelmann, M. & Schneider, P.: 1992, AA submitted.

Bergmann, A.: 1992, PhD Thesis, Stanford University.

Blandford, R.D. & Narayan, R.: 1986, ApJ 310, 568.

Blandford, R.D. & Narayan, R.: 1992, Ann. Rev. Astr. Ap. (in press)(BN).

Blandford, R.D., Saust, A.B., Brainerd, T.G. & Villumsen, J.V.: 1991, MNRAS 251, 600.

Burke, W.L.: 1981, ApJ 244, L1.

Corrigan, R.T. et al.: 1991, Astron. J. 102, 34.

Dyer, C.C. & Roeder, R.C.: 1973, ApJ 180, L31.

Ehlers, J. & Schneider, P.: 1986, AA 168, 57.

Falco, E.E., Gorenstein, M.V. & Shapiro, I.I.: 1991, ApJ 372, 364.

Filippenko, A.B.: 1989, ApJ 338, L49.

Fugmann, W.: 1990, AA 240, 11.

Futamase, T. & Sasaki, M: 1989, Phys. Rev. D 40, 2502.

Gorenstein, M.V. et al.: 1988, ApJ 334, 42.

Gorenstein, M.V., Falco, E.E. & Shapiro, I.I.: 1988, ApJ 327, 693.

Grieger, B.: 1990, in: Mellier et al. 1990, p.198.

Hammer, F. & Le Fevre, O.: 1990, ApJ 357, 38.

Hammer, F. & Le Fevre, O.: 1991, ESO Messenger 63, 59.

Hewitt, J.N. et al.: 1988, Nature 333, 537.

Huchra, J. et al.: 1985, Astron. J. 90, 691.

Jaroszynski, M., Wambsganss, J. & Paczynski, B.: 1992, preprint.

Kayser, R. & Schramm, T.(ed.): 1992, *Proceeding of the Hamburg conference on gravitational lensing*, Springer-Verlag (in press).

Kent, S.M. & Falco, E.E.: 1988, Astron. J. 96, 1570.

Kochanek, C.S.: 1990, MNRAS 247, 135.

Kochanek, C.S.: 1991a, ApJ 382, 58.

Kochanek, C.S.: 1991b, ApJ 379, 517.

Kochanek, C.S., Blandford, R.D., Lawrence, C.R. & Narayan, R.: 1988, MNRAS 238, 43.

Kochanek, C.S. & Narayan, R.: 1992, preprint.

Langston, G.I. et al.: 1990, Nature 344, 43.

McKenzie, R.H.: 1985, J. Math. Phys. 4, 1194.

Mellier, Y., Fort, B. & Soucail, G.(ed.): 1990, *Gravitational Lensing*, Lecture Notes in Physics 360, Springer-Verlag (Berlin).

Miralda-Escude, J.: 1991, ApJ 370, 1.

Moran, J.M., Hewitt, J.N. & Lo, K.Y.(ed.): 1989, *Gravitational Lenses*, Lecture Notes in Physics 330, Springer-Verlag (Berlin).

Narayan, R.: 1992, in: Kayser & Schramm 1992.

Narayan, R. & Schneider, P.: 1990, MNRAS 243, 192.

Ostriker, J.P. & Vietri, M.: 1985, Nature 318, 446.

Paczynski, B.: 1986, ApJ 304, 1.

Paczynski, B.: 1991, ApJ 371, L63.

Perlick, V.: 1990, Class. Quantum Grav. 7, 1319.

Press, W.H., Rybicki, G.B. & Hewitt, J.N.: 1992a, ApJ 385, 404.

Press, W.H., Rybicki, G.B. & Hewitt, J.N.: 1992b, ApJ 385, 416.

Rauch, K. & Blandford, R.D.: 1991, ApJ 381, L39.

Roberts, D.H. et al.: 1985, ApJ 293, 356.

Schneider, P.: 1984, AA 140, 119.

Schneider, P., Ehlers, J. & Falco, E.E.: 1992, *Gravitational Lenses*, Springer-Verlag (New York)(SEF).

Schneider, P. & Wambsganss, J.: 1990, AA 237, 42.

Seitz, S. & Schneider, P.: 1992, AA (submitted).

Sommerfeld, A.: 1959, *Vorlesungen über Theoretische Physik, Bd.IV. Optik*, Geest & Portig (Leipzig.

Smette, A. et al.: 1992, ApJ 389, 39.

Stickel, M., Fried, J. & Kühr, H.: 1988a, AA 198, L13.

Stickel, M., Fried, J. & Kühr, H.: 1988b, AA 206, L30.

Stickel, M., Fried, J. & Kühr, H.: 1989, AA 224, L27.

Stockton, A.: 1980, ApJ 242, L141.

Turner, E.L., Cen, R. & Ostriker, J.P.: 1992, AJ 103, 1427.

Tyson, J.A.: 1988, Astron. J. 96, 1.

Tyson, J.A., Valdes, F. & Wenk, R.A.: 1990, ApJ 349, L1.

Walsh, D.: 1989, in: Moran et al. 1989, p.11.

Wambsganss, J. & Paczynski, B.: 1992, preprint.

Wambsganss, J., Paczynski, B. & Schneider, P.: 1991, ApJ 358, L33.

Whitney, H.: 1955, Ann. Math. 62, 374.

Wu, X.P. & Hammer, F.: 1992, preprint.

Yee, H.K.C.: 1988, Astron. J. 95, 1331.

Young, P. et al.: 1980, ApJ 241, 507.

THE GENERAL RELATIVISTIC N-BODY PROBLEM

Thibault Damour [1], Michael Soffel [2] and Chongming Xu [3]

[1]Institut des Hautes Etudes Scientifiques,91440 Bures-sur-Yvette, and DARC-CNRS Observatoire de Paris, 92195 Meudon, France
[2]Theoretical Astrophysics, Uni. Tübingen, Auf der Morgenstelle 10, W-7400 Tübingen, Germany and DANOF, Observatoire de Paris, 61, avenue de l'Observatoire, 75014 Paris
[3]Department of Applied Mathematics, Univ. of Cape Town, Rondebosch, South Africa

1 Introduction

The development of a "post-Newtonian (PN) formalism" to deal with the general relativistic N-body problem began as early as 1915-1916 in papers by Einstein, Droste and De Sitter. This early work was motivated by the goal to derive observational predictions from General Relativity. Consequences of this historical PN approach are:

- the approach itself is conceived within the conceptual framework of the Newtonian N-body problem;
- General Relativity is compared and contrasted to Newton's theory. This led to the concept of "relativistic effects" = (Einstein) minus (Newton);
- some Newtonian theorems are taken over, without proof, to the general relativistic context;
- the then available observational accuracy motivated the choice of, and the way of solving, the theoretical problems (such as e.g. secular effects in the motion of N "mass points").

Usually this classical PN-approach to the N-body problem is based upon the following ingredients:

- one single coordinate grid for the whole system

$$(x^\mu) = (xt, x, y, z);$$

- one single PN-expansion;
- the use of basic variables or quantities introduced by analogy with the Newtonian ones, e.g. a "matter density" $\rho(t,x,y,z)$, a "center of mass", "multipole moments" or the "Newtonian potential" $U(t,\mathbf{x})$,

$$U(t,\mathbf{x}) \equiv \int d^3x' \frac{\rho(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}$$

- a PN-expansion of the metric in the form

$$g_{00} = -1 + \frac{2}{c^2}U + \frac{1}{c^4}\left[\Phi - 2U^2\right] + O\left(\frac{1}{c^6}\right),$$

$$g_{0i} = -\frac{4}{c^3}U_i + O\left(\frac{1}{c^5}\right),$$

$$g_{ij} = \left(1 + \frac{2}{c^2}U\right)\delta_{ij} + O\left(\frac{1}{c^4}\right),$$

where Φ represents a complicated "relativistic potential". The conceptual danger related with this traditional PN-approach lies in the fact that the "post-Newtonian framework" is misused as a kind of "neo-Newtonian framework". This misuse happened frequently in the literature, e.g. by an implicit identification of $t = x^0/c$ with "absolute Newtonian time", of $(x^i) = (x,y,z)$ with absolute Newtonian space, of $\rho(t,\mathbf{x})$ with "physical mass density" etc.; i.e. by a reduction of Einstein's theory to the Procustean Bed of Newton's theory (Fig. 1). Consequences of this misuse are: errors (e.g. in equations of motion), illusions and confusions.

Motivations for formulating a new PN framework come from two facts: (i) the recent development of new methods for treating the motion of strongly self-gravitating bodies, and (ii) the great improvement in measuring accuracy brought by modern technology. Whereas the dimensionless gravitational potential (GM/c^2R) everywhere in the solar system is smaller than about 10^{-6}, it can reach values of order unity in systems of compact bodies (neutron stars or black holes). E.g. $GM/c^2R \simeq 0.2$ at the surface of a neutron star. Therefore, the theoretical treatment of such compact objects and their motion in a binary system required the introduction of new methods (D'Eath 1975, Damour 1983, Thorne and Hartle 1985). On the other hand new techniques have been developed for the precise determination of i) the global mechanics of the solar system, ii) the local gravitational environment of the Earth and iii) the fitting of the global and local structures. As

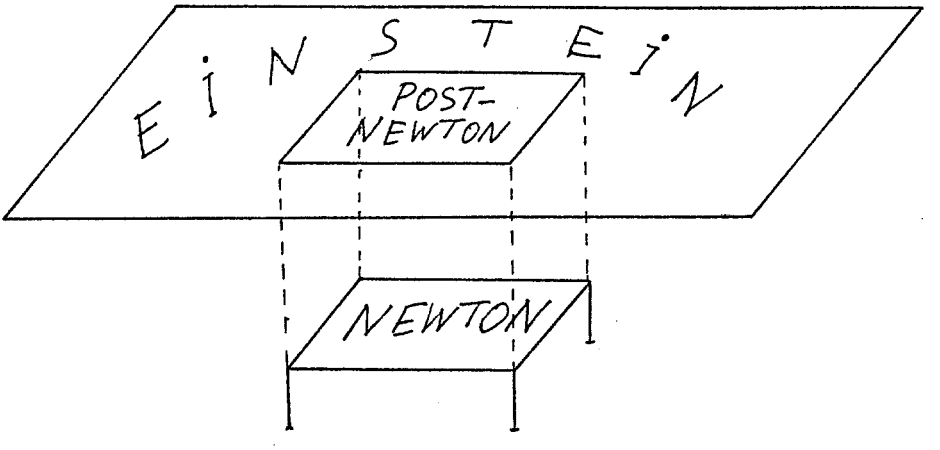

Fig. 1. The misuse of the PN-framework as reduction to the Procustean Bed of Newton's theory

far as the global mechanics of the solar system is concerned, radar ranging to planets now achieve accuracies of a few meters, laser ranging to the Moon a few centimeters and timing of millisecond pulsars less than a microsecond. The determination of the local gravitational environment of the Earth takes advantage of atomic clocks with stabilities of the order of $10^{-14} - 10^{-15}$, the Global Positioning System (GPS) or Satellite Laser Ranging with cm precision. Finally, the relation between the local and the global structures is studied by Very Long Baseline Interferometry (VLBI).

Motivated by these developments we have worked out a new approach (more "Einsteinian") to general relativistic celestial mechanics (Damour-Soffel-Xu 1991, 1992a,b). This new formalism is based upon three basic ingredients:

1. a systematic use of local reference systems with a new way of freezing down the coordinate freedom;
2. the use of new field variables and new matter variables with an associated "transformation theory" and
3. the use of new definitions of "multipole moments", "tidal moments", "center of mass" etc.

The convenience of using local reference systems had already been emphasized in Fokker (1920), Synge (1960), in the theory of motion of strong-

ly self-gravitating bodies (D'Eath 1975, Damour 1983, Thorne and Hartle 1985) and in other articles on the weak field problem (Ashby and Bertotti 1984, 1986; Brumberg and Kopejkin 1988, 1989. The new field and matter variables had been introduced for the problem of the generation of gravitational waves by an isolated system by Blanchet and Damour (1989) and Blanchet et al. (1990). The new definitions of "multipole moments" resulted from a generalization of the Blanchet-Damour (BD) moments (Blanchet and Damour 1989) to the gravitational N-body system. In the following we shall outline the main features of our new framework. For details and proofs one should consult Damour et al. 1991, 1992a,b.

2 First basic ingredient: local reference systems

2.1 Notation

We consider an isolated N-body system where rotating bodies of arbitrary shape and internal structure move under the influence of their mutual gravitational interaction. To describe such a system we employ $N+1$ different coordinate systems (or charts or reference systems): one global system of coordinates x^μ covering the entire system and which might extend to infinity and N local coordinate systems $X_1^\alpha, X_2^\alpha, \ldots, X_N^\alpha$, one for each of the N bodies which is (in a sense which will be defined later in the formalism) attached to the body and co-moving with it (see Fig. 2).

If possible we try to stick to a systematic notation: in the global system quantities will be denoted by small letters, space-time indices are chosen from the second half of the greek alphabet (e.g. $\mu,\nu = 0,1,2,3$), spatial indices from the second half of the roman alphabet (such as $i,j = 1,2,3$); in each of the local systems quantities will be denoted by capital letters and the indices are chosen from the first half of the alphabets (e.g. $\alpha,\beta = 0,1,2,3; a,b = 1,2,3$). Letters $A,B,C,\ldots$ are used as body labels.

2.2 Relativistic kinematics at the 1PN approximation

An important point in our scheme is the way in which the gauge problem (the choice of coordinates) is treated. Usually in the literature coordinates are fixed by certain *differential* gauge conditions, such as the harmonic gauge or the standard PN gauge in the first PN approximation. In our framework, however, in each of the $N+1$ charts, the spatial coordinates are fixed by *algebraic* conditions, whereas we allow for some freedom in the choice of the time coordinate. Note, that differential gauge conditions are not able to fix coordinates in the local accelerated systems where boundary conditions are not available. On the other hand leaving the time gauge open leads to a gauge freedom of the framework which is similar to the one in classical

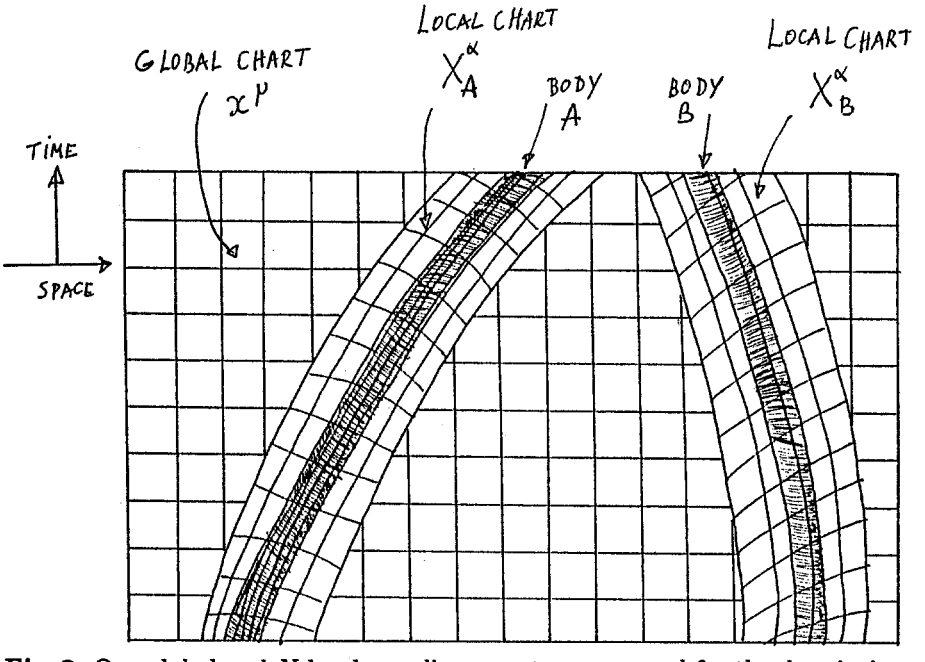

Fig. 2. One global and N local coordinate systems are used for the description of the gravitational N-body system

electrodynamics. As a consequence gauge invariant quantities can be introduced and the gauge freedom can serve as a useful tool for checks of involved calculations.

The logic of our fixing of spatial coordinates goes as follows: if we consider the spatial Einstein equations

$$ {}^4R^{ij}({}^4g) - \frac{1}{2}{}^4Rg^{ij} = \frac{8\pi G}{c^4}T^{ij} \tag{1} $$

plus the usual post-Newtonian assumptions:

$$ g_{00} = -1 + O(c^{-2}), \quad g_{0i} = O(c^{-3}), \quad g_{ij} = \delta_{ij} + O(c^{-2}) \tag{2a} $$

$$ T^{00} = O(c^{+2}), \quad T^{0i} = O(c^{+1}), \quad T^{ij} = O(c^0) \tag{2b} $$

and

$$ \partial_0 = O(c^{-1})\partial_i, \tag{2c} $$

then one finds that

$$ {}^3R_{ij}({}^3\gamma_{kl}) = O(c^{-4}), \tag{3} $$

where the following 3-dimensional metric has been introduced,

$$ \gamma_{ij} \equiv -g_{00}g_{ij} + g_{0i}g_{0j}, \tag{4} $$

and where the superscript 3 refers to the corresponding 3-dimensional Ricci tensor. But in a 3-dimensional manifold the vanishing of the Ricci tensor implies the vanishing of the Riemann curvature tensor. Hence, the metric γ_{ij} describes some 3-space which is *flat* modulo terms of order c^{-4}. This implies that there exist $\gamma-$Cartesian spatial coordinates such that

$$-g_{00}g_{ij} = \delta_{ij} + O(4) \tag{5a}$$

$$-G^A_{00}G^A_{ab} = \delta_{ab} + O(4) \tag{5b}$$

for each of the local systems A. Here, we wrote

$$O(n) \equiv O(c^{-n}). \tag{6}$$

We now face the following mathematical problem: what is the most general structure of coordinate transformations

$$x^\mu = f^\mu_A(X^\alpha_A) \tag{7}$$

from the global to the local A-system which preserve

1. the PN assumptions (2) and
2. the spatial isotropy conditions (5).

The answer is

$$x^\mu(X^\alpha) = z^\mu(X^0) + e^\mu_a(X^0)\left[X^a + \frac{1}{2c^2}A_a\mathbf{X}^2 - \frac{(\mathbf{A}\cdot\mathbf{X})}{c^2}X^a\right] + \eta^\mu, \tag{8}$$

where we have omitted a label A on all quantities on the right-hand side and where

$$e^\mu_0 = \frac{dz^\mu(S)}{dS}, \tag{9a}$$

$$e^0_a = e^i_a\frac{dz^i}{dS} + \frac{1}{c^3}\epsilon_a(S), \tag{9b}$$

$$e^0_0 e^i_a = \left(1 + \frac{\mathbf{v}^2}{2c^2}\right)\left(\delta^{ij} + \frac{v^i v^j}{2c^2}\right)R^j_a(S). \tag{9c}$$

In these equations $z^\mu(S)$ is the global-frame representation of the central worldline $(X^0, \mathbf{X}) = (S, \mathbf{0})$, A_a is given by

$$A_a \equiv c^2 e^i_a\frac{d^2z^i}{dS^2} + O(2) \tag{10}$$

and $\eta^\mu = O(\mathbf{X}^2)$ as $X^a \to 0$ with

$$\eta^0 = \frac{\xi(X^\alpha)}{c^3} = O(3), \quad \eta^i = O(4). \tag{11}$$

The velocity $\mathbf{v}$ of body A is given by

$$v^i = \frac{dz^i}{dS} + O(2) \tag{12}$$

and $R^j_a(S)$ is a slowly changing orthogonal matrix.

In other words a solution of the question raised above requires the following arbitrary elements (s. Fig.3)

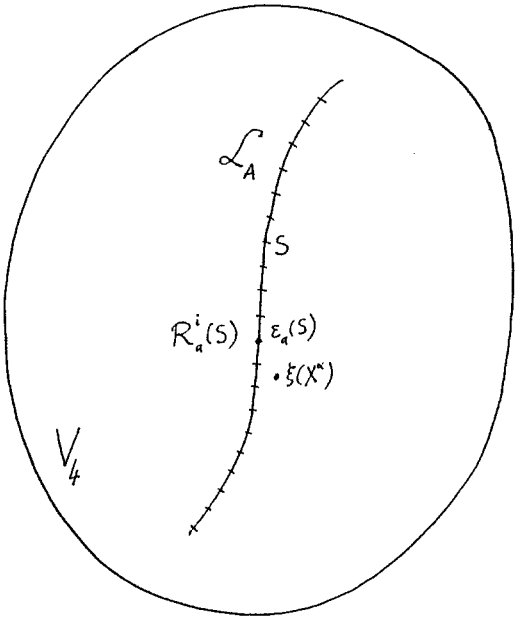

Fig. 3. Elements needed to describe the relation between the global and some local reference system

- a worldline $\mathcal{L}_A$ in a differentiable manifold V_4,
- a special parametrization S along $\mathcal{L}_A$,
- the global x^μ-representation of $\mathcal{L}_A$, $x^\mu = z^\mu_A(S)$,
- along $\mathcal{L}_A$:
  - a slowly changing orthogonal 3×3 matrix

$$R^i_a R^i_b = \delta_{ab}, \quad dR^i_a/dS = O(3) \tag{13}$$

  - three quantities $\epsilon_a(S)$
- around $\mathcal{L}_A$: a function $\xi(X^\alpha)$ with $\xi = O(\mathbf{X}^2)$ when $X^a \to 0$.

Note that this answer is independent of the value of the curved spacetime metric. We would like to indicate the main steps of the proof: the PN assumptions and spatial isotropy conditions lead to the relation

$$f_{\mu\nu}\frac{\partial x^\mu}{\partial X^a}\frac{\partial x^\nu}{\partial X^b} = \left(2 + f_{\mu\nu}\frac{\partial x^\mu}{\partial X^0}\frac{\partial x^\nu}{\partial X^0}\right)\delta_{ab} + O(4), \tag{14}$$

independent of the metric $g_{\mu\nu}$. Here, $f_{\mu\nu}$ refers to the flat space metric in global coordinates, i.e.

$$f_{\mu\nu} = \mathrm{diag}(-1, +1, +1, +1). \tag{15}$$

Without loss of generality we can write

$$x^\mu = f^\mu(X^\alpha) = z^\mu(X^0) + e^\mu_a(X^0)X^a + \xi^\mu(X^0, X^a), \tag{16}$$

where the first term is independent of X^a, the second linear in X^a and the third at least quadratic in X^a when $X^a \to 0$. Let us define

$$\xi^i \equiv e^i_a \Xi^a, \quad A_a \equiv f_{\mu\nu}e^\mu_a\frac{d^2 z^\nu}{dS^2} \tag{17}$$

then (14) leads to

$$\frac{\partial \Xi^b}{\partial X^c} + \frac{\partial \Xi^c}{\partial X^b} = -\frac{2}{c^2}(A_a X^a)\,\delta_{bc} + O(4). \tag{18}$$

This equation is easily recognized as the equation for conformal Killing vectors of Euclidean 3-space and since $\Xi^a = O(\mathbf{X}^2)$ when $X^a \to 0$ it has a unique solution ("inverted translations") in the form

$$\Xi^a = \frac{1}{c^2}\left[\frac{1}{2}A_a\mathbf{X}^2 - X^a(\mathbf{A}\cdot\mathbf{X})\right] + O(4). \tag{19}$$

The uniqueness of Ξ^a then leads to the uniqueness of our PN kinematics. Fig. 4 shows the various formal similarities between our relativistic and the Newtonian kinematics.

3 Second basic ingredient: new field and matter variables (and associated "transformation theory")

3.1 New field and matter variables

Essential for our new framework is the use of new *field* and *matter* variables. If we write the metric tensor in the form

$$g_{00} = -\exp\left(-\frac{2}{c^2}w\right) \tag{20a}$$

$$g_{0i} = -\frac{4}{c^3}w_i \tag{20b}$$

$$g_{ij} = +\gamma_{ij}\exp\left(+\frac{2}{c^2}w\right), \tag{20c}$$

RELATIVISTIC	NEWTONIAN
$x^\mu = z^\mu + e^\mu_a[X^a + \Xi^a] + \eta^\mu$	$\vec{x} = \vec{z}(t) + \mathcal{R}(t)\vec{X}$
1PN spatial rigidity	spatial rigidity
free central world line $\mathcal{L}_A$ $z^\mu_A(s)$	free translation of origin $\vec{z}(t)$
free 1PN choice of "time graduation" along $\mathcal{L}_A$	temporal rigidity
free choice of slowly changing $R^i_a(s)$	free choice of rotation matrix $\mathcal{R}(t)$

Fig. 4. Formal similarities and contrasts between our relativistic and the Newtonian kinematics

then the ten degrees of freedom are represented by one scalar field w, one vector field w_i with 3 degrees of freedom and one second rank symmetric tensor field γ_{ij} with 6 independent components. We have already learned that modulo $O(4)$ γ_{ij} represents the metric of some flat 3-space; therefore there exists preferred spatial coordinates x^i, such that

$$\gamma_{ij} = \delta_{ij} + O(4).$$

The spatial Einstein equations then are automatically fulfilled at the first PN level. The remaining Einstein equations

$$R^{00} = \frac{8\pi G}{c^4}\left(T^{00} - \frac{1}{2}Tg^{00}\right) \tag{21a}$$

$$R^{0i} = \frac{8\pi G}{c^4}\left(T^{0i} - \frac{1}{2}Tg^{0i}\right) \tag{21b}$$

then take the form

$$\Delta w + \frac{3}{c^2}\partial_t^2 w + \frac{4}{c^2}\partial_t\partial_i w_i = -4\pi G\sigma + O(4) \tag{22a}$$

$$\Delta w_i - \partial_{ij}^2 w_j - \partial_t\partial_i w = -4\pi G\sigma^i + O(2) \tag{22b}$$

with the new matter variables

$$\sigma \equiv \frac{T^{00} + T^{ss}}{c^2} \tag{23a}$$

$$\sigma^i \equiv \frac{T^{0i}}{c}. \tag{23b}$$

The variable σ plays the role of an "active gravitational mass density", σ^i the role of an "active mass current density". Note, that neither for the definition of these matter variables nor for the formulation of the two Einstein equations (22) had we to assume any specific structure for the matter distribution (such as a perfect fluid energy-momentum tensor). In fact, using our new field and matter variables, the exponential representation of the metric as in (20) and leaving the structure of matter completely open leads to a PN framework which is much more compact and simpler than the classical one. As can be seen from (22) the field equations even become *linear*! Moreover, the freedom in the time coordinate leaves the following electromagnetic-like gauge invariance: if (w, w_i) is a solution of Einstein's equations (22) then so is

$$w' = w - \frac{1}{c^2}\partial_t\lambda \tag{24a}$$

$$w_i' = w_i + \frac{1}{4}\partial_i\lambda, \tag{24b}$$

where $\lambda(x^\mu)$ is some arbitrary differentiable function. This gauge transformation corresponds to a change of the time coordinate according to

$$t' = t - \frac{1}{c^4}\lambda(t, \mathrm{x}). \tag{25}$$

This gauge invariance is formally useful in many respects and the formalism for the metric potential $w_\mu \equiv (w, w_i) \equiv (w, \mathrm{w})$ becomes very similar to classical electrodynamics. E.g. we can introduce gauge-invariant gravito-electric and gravito-magnetic fields e and b in the form

$$\mathbf{e} \equiv \nabla w + \frac{4}{c^2}\,\partial_t \mathbf{w} \tag{26a}$$

$$\mathbf{b} \equiv -4\nabla \times \mathbf{w} \tag{26b}$$

which obey Maxwell-like equations:

$$\nabla \cdot \mathbf{b} = 0 \tag{27a}$$

$$\nabla \times \mathbf{e} = -\frac{1}{c^2}\partial_t \mathbf{b} \tag{27b}$$

$$\nabla \cdot \mathbf{e} = -\frac{3}{c^2}\partial_t^2 w - 4\pi G\sigma + O(4) \tag{27c}$$

$$\nabla \times \mathbf{b} = 4\partial_t \mathbf{e} - 16\pi G\sigma + O(2). \tag{27d}$$

Note that in each of the local systems one can introduce local metric potentials ($W_\alpha^A = (W^A, W_a^A)$) by

$$G_{00}^A(X) = -\exp\left(-\frac{2}{c^2}W^A(X)\right) \tag{28a}$$

$$G_{0a}^A(X) = -\frac{4}{c^3}W_a^A(X) \tag{28b}$$

$$G_{ab}^A(X) = \delta_{ab}\exp\left(+\frac{2}{c^2}W^A(X)\right) + O(4). \tag{28c}$$

These local potentials satisfy field equations of exactly the same form as equations (22) above, for example

$$\Delta_X W + \frac{3}{c^2}\frac{\partial^2}{\partial T^2}W + \frac{4}{c^2}\frac{\partial}{\partial T}\frac{\partial}{\partial X^a}W_a = -4\pi G\Sigma(X) \tag{29}$$

etc. Here, X refers to the corresponding local coordinates and $\Sigma(X)$ is the active gravitational mass density as recorded in the local coordinate system

$$\Sigma(X) = \left(\frac{T^{00} + T^{aa}}{c^2}\right)_X. \tag{30}$$

We now come to the structure of the PN metric. For simplicity we will choose the temporal harmonicity condition which reads

$$0 = \Box_g x^0 = -\frac{4}{c^2}(\partial_t w + \partial_i w_i) + O(5) \tag{31}$$

in the global system. Einstein's equations then take the form (note the compactness and simplicity of the field equations!)

$$\Box w_\mu = -4\pi G\sigma^\mu + O(4,2). \tag{32}$$

Here, $O(4,2) = O(4)$ for $\mu = 0$ and $O(2)$ for $\mu = i$. Equations (32) can immediately be solved with the boundary conditions that the metric approaches the Minkowski metric at (past null) infinity (i.e. the metric potentials w_μ vanish at infinity). Because of the linearity of the field equations the solution can be written as a linear superposition of the contributions of the N bodies in the form

$$w_\mu(x) = \sum_{A=1}^{N} w_\mu^A(x) + O(4,2), \tag{33}$$

where

$$w_\mu^A(x) = G \int_A \frac{d^3x'}{|\mathbf{x} - \mathbf{x}'|} \sigma^\mu \left(t \mp \frac{|\mathbf{x} - \mathbf{x}'|}{c}, \mathbf{x}' \right). \tag{34}$$

Here, the $\mp$ sign stands for the time-symmetric average, i.e.

$$f\left(t \mp \frac{r}{c}\right) \equiv \frac{1}{2}\left[f\left(t - \frac{r}{c}\right) + f\left(t + \frac{r}{c}\right)\right]. \tag{35}$$

Indeed, we can use the time-symmetric solution instead of the retarded one because they lead to different physical predictions (in the near zone) at higher post-Newtonian orders (2.5PN). Eqs. (33) and (34) describe the full solution of the field equations in the global system in harmonic time gauge at the 1PN approximation.

3.2 Transformation theory

In each of the N local systems the field equations

$$\Box_X W^A + \frac{4}{c^2} \partial_T \left(\partial_T W^A + \partial_b W_b^A \right) = -4\pi G \Sigma^A + O(4) \tag{36a}$$

$$\Delta_X W_a^A - \partial_a \left(\partial_T W^A + \partial_b W_b^A \right) = -4\pi G \Sigma_a^A + O(2) \tag{36b}$$

are linear but we cannot impose any boundary conditions as we did in the global system. We now decompose the local metric potentials $W_\alpha^A(X)$ as a sum of two contributions

$$W_\alpha^A(X) = \overset{+}{W}{}_\alpha^A + \overline{W}_\alpha^A, \tag{37}$$

where $W^+{}_\alpha^A$ ("self potentials") is the piece that is locally generated by body A as seen in its corresponding local frame and $\overline{W}_\alpha^A$ ("external potentials") is a remainder which is generated by all the other bodies in the system and also by inertial effects due to the accelerated motion of the local system. We define the self-part $W^+{}_\alpha^A$ by

$$\overset{+}{W}{}_\alpha^A(X_A^\beta) = G \int_A \frac{d^3X'_A}{|\mathbf{X}_A - \mathbf{X}'_A|} \Sigma^\alpha \left(T_A \mp \frac{|\mathbf{X}_A - \mathbf{X}'_A|}{c}, \mathbf{X}'_A \right). \tag{38}$$

Together with the solution (33) in the global system and the transformation laws of the metric potentials that we will discuss now this also fixes the external part of the metric $\overline{W}{}^{A}_{\alpha}$. In other words, in each of the $N+1$ reference systems the solution of the field equations are now available in explicit form given the transformation laws for the metric potentials between the global and the local systems. Remember that the metric in the global system is represented by (20) (i.e. by $w_\mu(x)$), whereas in the local A-system it is represented by (28) (i.e. by $W^A_\alpha(X)$). The relation between the various metric potentials is simply derived from the usual tensorial transformation rule for the metric tensor

$$g^{\mu\nu}\left[x^\lambda(X^\alpha_A)\right] = \frac{\partial x^\mu(X)}{\partial X^\alpha_A}\frac{\partial x^\nu(X)}{\partial X^\beta_A}\, G^{\alpha\beta}_A(X^\gamma_A) \tag{39}$$

with

$$x^\mu(X) = z^\mu(X^0) + e^\mu_a(X^a + \Xi^a) + \eta^\mu.$$

From this one gets

$$w_\mu(x) = \mathcal{A}^A_{\mu\alpha}(X^0)W^A_\alpha(X) + \mathcal{B}^A_\mu(X) \tag{40}$$

with

$$[\mathcal{A}_{\mu\alpha}] = \begin{bmatrix} \mathcal{A}_{00} & \mathcal{A}_{0a} \\ \mathcal{A}_{i0} & \mathcal{A}_{ia} \end{bmatrix} = \begin{bmatrix} 1 + 2\mathbf{V}^2/c^2 & 4V_a/c^2 \\ v^i & R^i_a \end{bmatrix}, \tag{41}$$

where

$$\mathcal{B}_0 = \frac{c^2}{2}\ln(A^0_0A^0_0 - A^0_aA^0_a) \tag{42a}$$

$$\mathcal{B}_i = \frac{c^3}{4}(A^0_0A^i_0 - A^0_aA^i_a), \tag{42b}$$

where

$$A^\mu_\alpha = \frac{\partial x^\mu(X^\beta)}{\partial X^\alpha} \tag{42c}$$

is the Jacobian matrix of the coordinate transformation. Note the affine character of the transformation law for the metric potentials. Let us now consider the decompositions

$$w_\mu(x) = \overset{+}{w}{}^A_\mu(x) + \overline{w}^A_\mu(x) \tag{43a}$$

$$W^A_\alpha(X) = \overset{+}{W}{}^A_\alpha(X) + \overline{W}^A_\alpha(X) \tag{43b}$$

in which $w^{+\,A}_{\ \mu} \equiv w^A_\mu$, as defined in harmonic gauge by (34),

$$\overline{w}^A_\mu \equiv \sum_{B\neq A} w^B_\mu(x)$$

and where $W^{+\,A}_{\alpha}(X)$ is given, in harmonic gauge, by (38). In order to derive an expression for $\overline{W}^{A}_{\alpha}(X)$ we need *detailed transformations laws* for W^{+} and $\overline{W}$ separately. These detailed transformation laws read:

$$\overset{+}{w}{}^{A}_{\mu}(x) = \mathcal{A}^{A}_{\mu\alpha}(X^0)\,\overset{+}{W}{}^{A}_{\alpha}(X) + O(4,2) \tag{44a}$$

$$\overline{w}^{A}_{\mu}(x) = \sum_{B\neq A} w^{B}_{\mu}(x) = \mathcal{A}^{A}_{\mu\alpha}\overline{W}^{A}_{\alpha}(X) + \mathcal{B}^{A}_{\mu}(X) + O(4,2). \tag{44b}$$

Note how remarkably simple this result is: the self-part of the metric potential transforms just with the homogeneous part of the full transformation (40). The Newtonian analogues of these transformation rules would be

$$\overset{+}{u}{}^{A}(x) = \int_{A} d^3x' \frac{\rho(t,\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} = \int_{A} d^3X' \frac{\rho(T,\mathbf{X}')}{|\mathbf{X}-\mathbf{X}'|} \equiv \overset{+}{U}{}^{A}(X) \tag{45a}$$

$$\sum_{B\neq A} u^{B}(x) = \overline{U}^{A}(X) + C(t) + \frac{d^2\mathbf{z}}{dt^2}\mathbf{X}, \tag{45b}$$

where $\overline{U}^{A}$ represents the additional (effective) potential which must be added to the locally generated one (45a) when describing the gravitational dynamics in a local accelerated frame, i.e.

$$\overline{U}^{A}(X) = \sum_{B\neq A} u^{B}(x) - C(t) - \frac{d^2\mathbf{z}}{dt^2}\mathbf{X}. \tag{46}$$

From this we see that the inhomogeneous part $\mathcal{B}^{A}_{\mu}$ on the r.h.s. of (40) contains inertial terms due to the acceleration of the local A-system. Equations (44) constitute a central point in our framework. We concocted three independent proofs for (44); an elegant proof which employs certain invariance properties of the time-symmetric Green's function for the d'Alembertian can be found in Damour et al. (1991). The detailed transformation laws (44) lead to a consistent and complete formulation of our method in the "continuous" case, i.e. without skeletonizing the bodies. This "continuous" formulation is outlined in Fig.5.

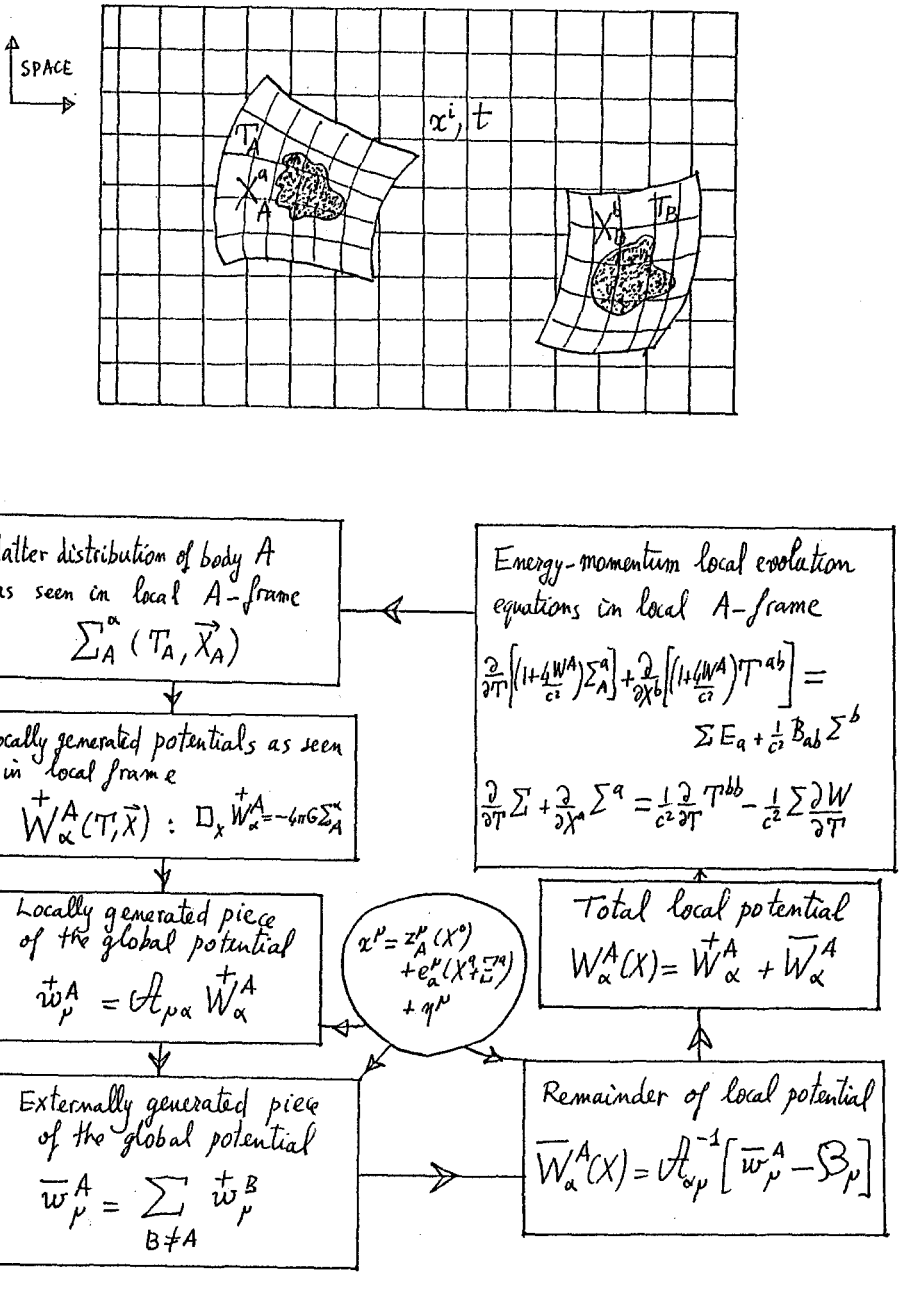

Fig. 5. The "continuous" formulation of the method

4 Third basic ingredient: multipole and tidal moments or the "skeletonized" formulation

In our framework we use new definitions of collective variables such as "multipole moments", "center of mass" or "tidal moments". Characteristics of the new "relativistic multipole moments" (of body A) are:

- they are expressed in terms of the matter distribution of body A as seen in the local A-frame: $\Sigma^A_\alpha(T,\mathbf{X})$;
- the locally generated relativistic potentials, $W^{+\,A}_{\ \alpha}(T,\mathbf{X})$, can be expressed in terms of them (Blanchet and Damour 1989).

Characteristics of the new "relativistic tidal moments" are:

- they are expressed in terms of the local-frame "external" gravitational potentials: $\overline{W}^A_\alpha(T,\mathbf{X})$;
- they are gauge invariant in the sense described above;
- the local evolution equations for the matter distribution of A can be expressed in terms of $W^{+\,A}_{\ \alpha}$ and of the tidal moments.

The (BD) mass and spin multipole moments of body A are defined by

$$\begin{aligned} M^A_{a_1a_2\ldots a_l}(T) \equiv & \int_A d^3X\, X^{<a_1}X^{a_2}\ldots X^{a_l>}\left[\frac{T^{00}+T^{bb}}{c^2}\right]_{\text{A-frame}} \\ & + \frac{1}{2(2l+3)c^2}\frac{d^2}{dT^2}\int_A d^3X\,\mathbf{X}^2X^{<a_1}X^{a_2}\ldots X^{a_l>}\left[\frac{T^{00}}{c^2}\right]_A \\ & - \frac{4(2l+1)}{(l+1)(2l+3)c^2}\frac{d}{dT}\int_A d^3X\,X^{<b}X^{a_1}\ldots X^{a_l>}\left[\frac{T^{0b}}{c}\right]_A \end{aligned} \tag{47a}$$

and

$$S^A_{a_1\ldots a_2}(T) \equiv \int_A d^3X\,\epsilon_{bc<a_l}X_{a_1}\ldots X_{a_{l-1}>}X_b\left[\frac{T^{0c}}{c}\right]_{\text{A-frame}}. \tag{47b}$$

Here, the brackets $<>$ indicate a symmetric and trace-free (STF) projection. Hence, the mass and spin moments are Cartesian STF tensors. The relativistic tidal moments also have this property; they are defined by

$$G^A_{a_1a_2\ldots a_l}(T) \equiv \partial_{<a_1a_2\ldots a_{l-1}}\overline{E}^A_{a_l>}|_{\mathbf{X}=0} \tag{48a}$$

$$H^A_{a_1a_2\ldots a_l}(T) \equiv \partial_{<a_1a_2\ldots a_{l-1}}\overline{B}^A_{a_l>}|_{\mathbf{X}=0}, \tag{48b}$$

where $\overline{E}$ and $\overline{B}$ are the external gravito-electric and gravito-magnetic fields belonging to the local A-system:

$$\overline{E}^A_a = \partial_a \overline{W}^A + \frac{4}{c^2}\partial_T \overline{W}^A_a \tag{49a}$$

$$\overline{B}^A_a = \epsilon_{abc}\partial_b(-4\overline{W}^A_c). \tag{49b}$$

The multipole mass and spin moments have the remarkable property that the self-potentials can be expressed in terms of them in a very simple manner. The following relativistic multipole expansions hold everywhere outside body A:

$$\overset{+}{W}{}^A(T,\mathbf{X}) = G\sum_{l\geq 0}\frac{(-)^l}{l!}\partial_L\left(\frac{M^A_L(T\mp R/c)}{R}\right)$$

$$+\frac{1}{c^2}\partial_T(\Lambda^A-\lambda)+O(4) \tag{50a}$$

$$\overset{+}{W}{}^A_a(T,\mathbf{X}) = -G\sum_{l\geq 1}\frac{(-)^l}{l!}\left[\partial_{L-1}\left(\frac{dM^A_{aL-1}/dT}{R}\right)\right.$$

$$\left.+\frac{l}{l+1}\epsilon_{abc}\partial_{bL-1}\left(\frac{S^A_{cL-1}}{R}\right)\right]$$

$$-\frac{1}{4}\partial_a(\Lambda^A-\lambda)+O(2), \tag{50b}$$

where $L\equiv a_1a_2\ldots a_l$, $\partial_L\equiv\partial^l/\partial X^{a_1}\ldots\partial X^{a_l}$ and $M_L\equiv M_{a_1a_2\ldots a_l}$ (the $\Lambda^A-\lambda$-terms are pure gauge). This means that outside body A the self-field generated by body A can be skeletonized by means of the mass and spin-moments, M_L, S_L, of body A itself. In other words, the self-field outside body A is completely determined by its mass $M(T)$, its dipole moment $M_a(T)$, quadrupole moment $M_{ab}(T)$ etc. and by its spin dipole $S_a(T)$, spin quadrupole $S_{ab}(T)$ etc.

In a dual manner, the relativistic tidal moments, G_L, H_L, completely determine (modulo gauge terms) the *external* potentials in the local A-system ($\dot{} = d/dT$):

$$\overline{W}^A(T_A,\mathbf{X}_A) = \sum_l\frac{1}{l!}\left[\hat{X}G^A_L(T)+\frac{1}{2(2l+3)c^2}\mathbf{X}^2\hat{X}^L\ddot{G}^A_L(T)\right]+$$

$$+\frac{1}{c^2}\partial_T\bar{\Lambda}^A+O(4) \tag{51a}$$

$$\overline{W}^A_a(T_A,\mathbf{X}_A) = \sum_l\frac{1}{l!}\left[-\frac{2l+1}{(l+1)(2l+3)}\hat{X}^{aL}\dot{G}^A_L(T)\right.$$

$$+\frac{l}{4(l+1)}\epsilon_{abc}\hat{X}^{bL-1}H^{A}_{cL-1}(T)\Big]$$

$$-\frac{1}{4}\partial_a\bar{A}^A+O(2), \tag{51b}$$

where

$$\hat{X}_L\equiv X^{<L>}\equiv \mathrm{STF}_L X^L. \tag{52}$$

This means that the combined tidal-inertial forces experienced by body A can be "skeletonized" by means of some gravito-electric tidal monopole $G(T)$, dipole $G_a(T)$, quadrupole $G_{ab}(T)$ etc. and some gravito-magnetic tidal dipole $H_a(T)$, quadrupole $H_{ab}(T)$ etc. In the Newtonian limit only the gravito-electric tidal moments survive (see the discussion in Sec. VI A of Damour et al. 1991).

5 Equations of motion

5.1 D'Alembert's approach to the problem of motion

We now come to the problem of equations of motion. There are several principles which might be used to derive equations of motion. Let us first describe them within the familiar framework of Newtonian dynamics. In Newton's viewpoint we might consider some body from an inertial frame (x) and compute the acceleration of each mass element from Newton's law

$$(dm)\mathbf{a}=d\mathbf{F}.$$

Integration of this equation over body A then leads to the equation of translational motion in the inertial frame. However, equally well we may derive this equation of motion in the manner of d'Alembert as an equilibrium condition of the form

$$\int_A [d\mathbf{F}+d\mathbf{F}_{\mathrm{inertial}}]=0 \tag{53}$$

formulated in the accelerated system that is comoving with body A. Now, the Euler equations in the local accelerated system read

$$\frac{\partial\rho_A(X)}{\partial t}+\frac{\partial}{\partial X^i_A}\left[\rho_A V^i_A\right]=0 \tag{54}$$

$$\frac{\partial(\rho_A V^i_A)}{\partial t}+\frac{\partial}{\partial X^j_A}\left[\rho_A V^i_A V^j_A+t^{ij}\right]=\rho_A\frac{\partial}{\partial X^i_A}U^{\mathrm{eff}}_A, \tag{55}$$

where U^{eff}_A is the total effective potential (including inertial effects) to be used in the local accelerated frame, i.e.

$$U^{\mathrm{eff}}_A(X)=U(\mathbf{z}_A+\mathbf{X})-C(t)-\frac{d^2\mathbf{z}_A}{dt^2}\mathbf{X}=\overset{+}{U}{}^A+\overline{U}^A, \tag{56}$$

in terms of the notation (45), (46). By integrating equations (54) and (55) one gets

$$\frac{d}{dt}m^A(t) = 0 \tag{57a}$$

$$\frac{d^2}{dt^2}m_i^A(t) = \int_A d^3X_A\,\rho\frac{\partial\overline{U}^A}{\partial X_A^i} \tag{57b}$$

$$\frac{d}{dt}s_i^A(t) = \epsilon_{iab}\int_A d^3X_A\,\rho X_A^a\frac{\partial\overline{U}^A}{\partial X_A^b}, \tag{57c}$$

where

$$m_i^A(t) \equiv \int_A d^3X\,\rho X_A^i \tag{58}$$

$$s_i^A(t) \equiv \epsilon_{iab}\int_A d^3X\,\rho X_A^a V_A^b. \tag{59}$$

Let us now introduce some (Newtonian) tidal moments by

$$\overline{U}^A(X) = \sum_l \frac{1}{l!}g_L^A(t)\hat{X}^L \equiv \sum_l \frac{1}{l!}g_{i_1\ldots i_l}^A(t)X^{<i_1}\ldots X^{i_l>}. \tag{60}$$

Equations (57) then take the form

$$\frac{dm^A(t)}{dt} = 0 \tag{61a}$$

$$\frac{d^2m_i^A(t)}{dt^2} = \sum_l \frac{1}{l!}m_L^A g_{iL}^A = m^A g_i^A + m_j^A g_{ij}^A + \ldots \tag{61b}$$

$$\frac{ds_i^A(t)}{dt} = \sum_l \frac{\epsilon_{iab}}{l!}m_{aL}^A g_{bL}^A = \epsilon_{iab}m_a^A g_b^A + \epsilon_{iab}m_{aj}^A g_{bj}^A + \ldots. \tag{61c}$$

Let us now attach the spatial origin of the accelerated local A-frame to the matter distribution of body A. We do this by choosing this origin as the center of mass of body A, which means imposing the vanishing of the mass dipole moment

$$0 = m_i^A(t) \equiv \int_A d^3X\,\rho X_A^i = \int_A d^3x\,\rho\left[x^i - z_A^i(t)\right]. \tag{62}$$

Since this equation should be valid for all times we also have

$$\frac{d^2m_i^A(t)}{dt^2} = 0. \tag{63}$$

Inserting the information (63) into (61b) we get the translational equations of motion in the form of an *equilibrium condition* in the local accelerated frame:

$$0 = m^A g_i^A + 0 + \frac{1}{2} m_{jk}^A g_{ijk}^A + \dots \ . \tag{64}$$

To transform the d'Alembert-like equation of motion (64) (equilibrium between "real" and "inertial" forces) into a more usual Newtonian-like inertial-frame equation of motion it is sufficient to note, from (56) and (60), that

$$g_i^A = \left. \frac{\partial \overline{U}^A}{\partial X^i} \right|_{\mathbf{z}_A} = \sum_{B \neq A} \partial_i u^B(\mathbf{z}_A) - \frac{d^2 z_A^i}{dt^2}. \tag{65}$$

Separating out the inertial term $-\ddot{z}^i_{\rm CMA}$ (where $z^i_{\rm CMA}$ denotes the center-of-mass of body A) gives the inertial-frame translational equation of motion of body A:

$$m^A \frac{d^2 z^i_{\rm CMA}}{dt^2} = \sum_{B \neq A} \left[m^A \partial_i u^B(\mathbf{z}_{\rm CMA}) + \frac{1}{2} m_{jk}^A \partial_{ijk} u^B(\mathbf{z}_{\rm CMA}) + \dots \right]. \tag{66}$$

5.2 Relativistic tidally expanded equations of motion

The above d'Alembert approach to the problem of motion can conveniently be transferred to our PN framework. The local evolution equations in the local frame are obtained from

$$0 = \nabla_\beta T^{\alpha\beta} = \frac{\partial}{\partial X^\beta} T^{\alpha\beta} + \Gamma^\alpha_{\sigma\beta} T^{\sigma\beta} + \Gamma^\beta_{\sigma\beta} T^{\alpha\sigma} \tag{67}$$

and read

$$\frac{\partial}{\partial T} \left[\left(1 + \frac{4W}{c^2} \right) \Sigma^a \right] + \frac{\partial}{\partial X^b} \left[\left(1 + \frac{4W}{c^2} \right) T^{ab} \right] = F^a(T, \mathbf{X}) + O(4) \tag{68a}$$

$$\frac{\partial}{\partial T} \Sigma + \frac{\partial}{\partial X^a} \Sigma^a = \frac{1}{c^2} \frac{\partial}{\partial T} T^{bb} - \frac{1}{c^2} \Sigma \frac{\partial W}{\partial T} + O(4) \tag{68b}$$

with the local force density

$$F^a(T, \mathbf{X}) = \Sigma E_a + \frac{1}{c^2} B_{ab} \Sigma^b = \left(\Sigma \mathbf{E} + \frac{1}{c^2} \boldsymbol{\Sigma} \times \mathbf{B} \right)_a . \tag{69}$$

Note the simple Lorentz form of this relativistic force density! Here, the local gravito-electric and gravito-magnetic fields are described by

$$E_a(W) = \partial_a W + \frac{4}{c^2} \partial_T W_a \tag{70a}$$

$$B_{ab}(W) = \partial_a(-4W_b) - \partial_b(-4W_a). \tag{70b}$$

As introduced above, we can now decompose the full potentials W into self- and external parts and use the expansions (51) of $\overline{W}$ into tidal moments. Using the relativistic definitions for M^A and M_a^A as in (47a) one finds that two remarkable things happen: i) all self-effects due to the self-potentials W^+ cancel and ii) all "bad expressions" which cannot be combined to yield BD moments also cancel. Finally we get equations of the form

$$\frac{dM^A}{dT_A} = \bar{F}_0^A \left[M_L^A, G_L^A\right] + O(4) \tag{71a}$$

$$\frac{d^2 M_a^A}{dT_A^2} = \bar{F}_a^A \left[M_L^A, S_L^A, G_L^A, H_L^A\right] + O(4) \tag{71b}$$

$$\frac{dS_a^A}{dT_A} = \bar{L}_a^A \left[M_L^A, S_L^A, G_L^A, H_L^A\right] + O(4/2). \tag{71c}$$

In (71c) the symbol $O(4/2)$ means that the usual definition (47b) appearing in the 1PN framework gives the spin vector and therefore the spin motion only to Newtonian order, but that a post-Newtonian definition for the spin vector exists such that the torque on the r.h.s. of (71c) can be expressed in terms of our internal and external moments (Damour et al. 1992b). The r.h.s. of equations (71) read explicitly

$$\bar{F}_0 = -\frac{1}{c^2}\sum_l \frac{1}{l!}\left\{(l+1)M_L\dot{G}_L + l\dot{M}_L G_L\right\} \tag{72a}$$

$$\begin{aligned}\bar{F}_a = \sum_l \frac{1}{l!}\Bigg\{ & M_L G_{aL} + \frac{l}{c^2(l+1)} S_L H_{aL} + \frac{\epsilon_{abc}}{c^2(l+2)} M_{bL}\dot{H}_{cL} \\ & + \frac{\epsilon_{abc}}{c^2(l+1)}\dot{M}_{bL}H_{cL} - \frac{4(l+1)}{c^2(l+2)^2}\epsilon_{abc}S_{bL}\dot{G}_{cL} \\ & - \frac{4}{c^2(l+2)}\epsilon_{abc}\dot{S}_{bL}G_{cL} - \frac{2l^3+7l^2+15l+6}{c^2(l+1)(2l+3)}M_{aL}\ddot{G}_L \\ & - \frac{2l^3+5l^2+12l+5}{c^2(l+1)^2}\dot{M}_{aL}\dot{G}_L - \frac{l^2+l+4}{c^2(l+1)}\ddot{M}_{aL}G_L + O(4)\end{aligned} \tag{72b}$$

$$\bar{L}_a = \sum_l \frac{1}{l!}\epsilon_{abc}M_{bL}G_{cL} + \frac{1}{c^2}\cdots . \tag{72c}$$

Having in hand the results (71), (72) we can now generalize to the relativistic context the d'Alembert approach to the problem of motion. Let us attach the spatial origin of the local A-frame to the matter distribution of body A by requiring the vanishing of the BD mass dipole moment:

$$M_a^A(T) = 0. \quad \text{(attachment condition)} \tag{73}$$

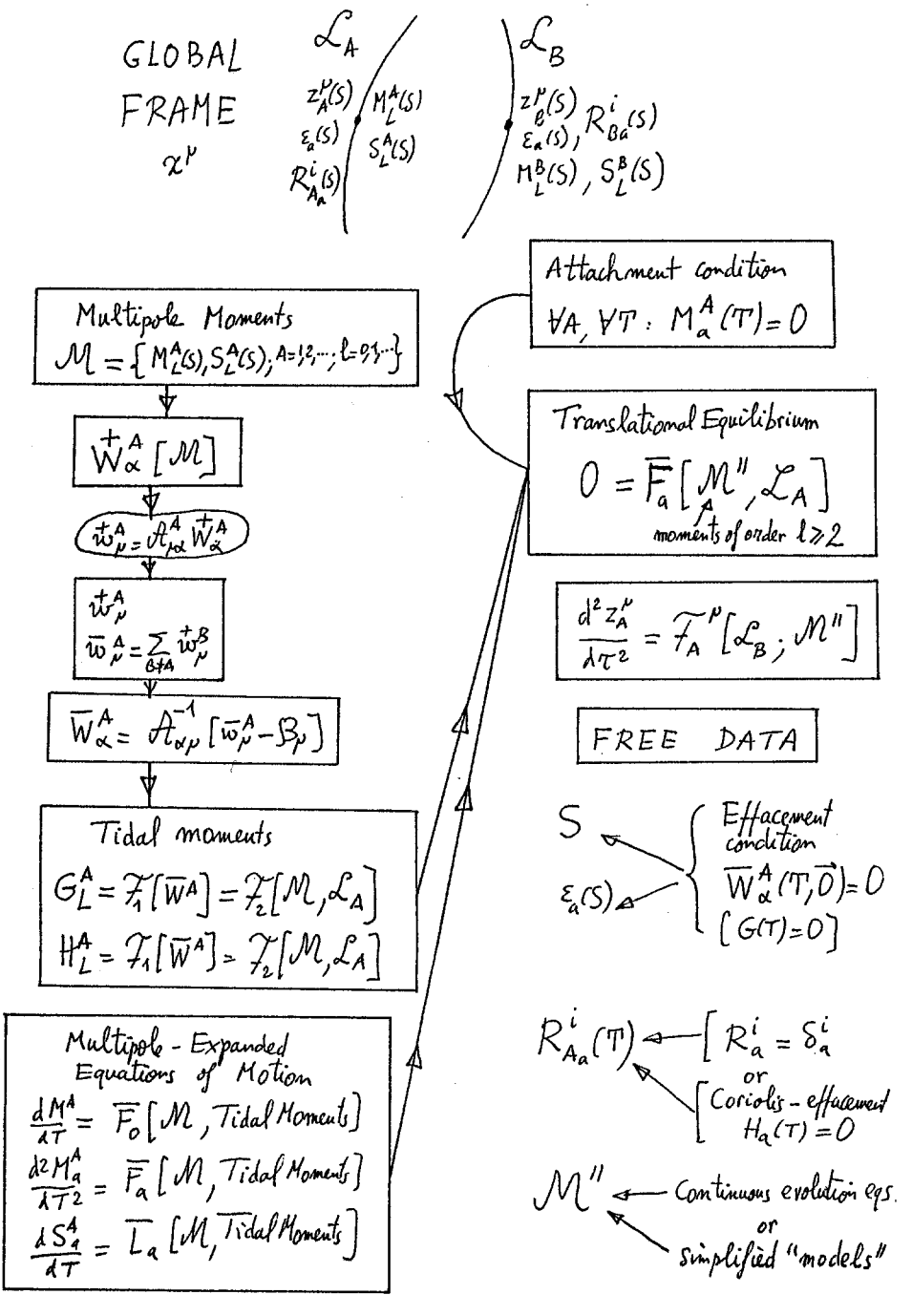

Fig. 6. Skeletonized formulation of the method

Note that this is a new, specific way of defining a relativistic center of mass for body A. Combining (71b) and (73), we then get the full PN translational equations of motion in the form of a local equilibrium condition

$$\bar{F}_a = 0. \tag{74}$$

Then, the Newton-like form of the PN equations of motion, i.e. the relativistic analogue of (66) giving now the acceleration of the global-system coordinate position of the PN center of mass of body A, z^i_A, is obtained by separating out the inertial term, $-\ddot{z}^i_A + O(c^{-2})$, hidden in the PN tidal dipole moment G_a appearing in the right-hand side of (72b). Finally, to make these equations fully explicit we have to do four things:

1. We take advantage of the freedom of choice of the special parametrization S along $\mathcal{L}_A$ to set the monopole gravito-electric tidal moment $G(T) \equiv \overline{W}(T, \mathbf{0})$ to zero (weak effacement condition).
2. We also fix the freedom of choice in the orthogonal matrix. For example the choice $R^A_{ia} = \delta_{ia}$ suggests itself as a particularly simple one.
3. We must compute the expressions for the remaining tidal moments, $G^A_a, G^A_{ab}, \ldots, H^A_a, H^A_{ab}, \ldots$ in terms of the PN multipole moments of the other bodies $B \neq A$ (see Damour et al. 1992a).
4. Finally, to obtain a closed system of equations of motion, we need to specify somehow the time evolution of the multipole moments of order $l \geq 2 : \mathcal{M}'' = \{G^A_{ab}, H^A_{ab}, G^A_{abc}, \ldots\}$. The logic of the derivation of completely explicit PN equations of motion is summarized schematically in Fig. 6. For the detailed implementation of this method to the simplest "monopole approximation" (Lorentz-Droste-Einstein-Infeld-Hoffman equations of motion) see Sec. VII C of Damour et al. (1991).

References

Ashby, N., Bertotti, B. (1984): Phys.Rev.Lett. **52**, 485

Ashby, N., Bertotti, B. (1986): Phys. Rev. **D34**, 2246

Blanchet, L., Damour, T. (1989): Ann. Inst. Henri Poincaré **50**, 377

Blanchet, L., Damour, T., Schäfer, G. (1990): Mon. Not. R. astr. Soc. **242**, 289

Brumberg, V.A., Kopejkin, S.M. (1988): in: "Reference Systems", eds. J. Kovalevsky, I.I. Mueller and B. Kolaczek, Reidel, Dordrecht

Brumberg, V.A., Kopejkin, S.M. (1989): Nuovo Cim B **103**, 63

Damour, T. (1983): in "Gravitational Radiation", eds N. Deruelle and T. Piran, North-Holland, Amsterdam, pp. 59-144

Damour, T., Soffel, M., Xu, C. (1991): Phys. Rev. D **43**, 3273

Damour, T., Soffel, M., Xu, C. (1992a): Phys. Rev. D, in press

Damour, T., Soffel, M., Xu, C. (1992b): to be submitted to Phys. Rev. D

D'Eath, P.D. (1975): Phys. Rev. **D11**, 2183

Fokker, A.D. (1920): Kon. Akad. Weten. Amsterdam, Proc. **23**, 729
Synge, J.L. (1960): "Relativity: The General Theory", North-Holland, Amsterdam
Thorne, K.S., Hartle, J.B. (1985): Phys. Rev. **D31**, 1815

OBSERVABLE RELATIVISTIC EFFECTS IN THE SOLAR SYSTEM

Michael H. Soffel

Uni. Tübingen, Theor. Astrophysics, Auf der Morgenstelle 10, 7400 Tübingen, FRG

1 Introduction

This article contains a brief overview over measurable general relativistic effects in the solar system. Unfortunately, because of time constraints, several important topics, especially those related with specific tests of Einsteinian gravity were not discussed during the Bad Honnef Meeting. Among the more interesting topics not discussed are e.g. new developments for improving the test of the (weak) equivalence principle such as a Satellite Test of the Equivalence Principle (STEP). The goal of STEP is a test of the equivalence principle at a level of 10^{-17}, roughly a six orders of magnitude improvement over previous tests by a free-fall experiment in a drag-free satellite (see e.g. Worden et al., 1990, Barlier et al., 1991).

2 Gravitational redshift, clock rates etc.

To classify the various experiments related with General Relativity it is useful to start from some explicit form for the space-time metric. E.g. in the vicinity of the Earth using local geocentric coordinates (cT, X^a) the metric, satisfying Einstein's equation in the first post-Newtonian approximation might be written in the form (Damour et al., 1991, 1992)

$$G_{00} = -\exp(-2W/c^2) = -1 + 2W/c^2 - 2W^2/c^4 + O(6) \qquad (1a)$$

$$G_{0a} = -4W_a/c^3 \qquad (1b)$$

$$G_{ab} = \delta_{ab}\exp(+2W/c^2) + O(4) = \delta_{ab}(1 + 2W/c^2) + O(4). \qquad (1c)$$

Here, $O(n) \equiv O(c^{-n})$ stands for the orders of neglected terms. We see that at the 1PN level the metric can be represented by some scalar potential W,

generalizing the usual Newtonian potential and some vector potential W_a describing gravito-magnetic effects. It can be shown that these potentials can be written as a sum of two parts:

$$W = \overset{+}{W} + \overline{W} \tag{2}$$

where W^+ is the self-part due to the Earth itself and $\overline{W}$ is the external part originating from all other bodies in the solar system. To a good approximation W^+ is given by (Damour et al., 1991, 1992)

$$\overset{+}{W} = G \sum_{l \geq 0} \frac{(-)^l}{l!} \partial_L \left(R^{-1} M_L^{\oplus}(T \pm R/c) \right), \tag{3}$$

where the $\pm$-sign indicates a time symmetric average,

$$M_L(T \pm R/c) = \frac{1}{2} \left[M_L(T + R/c) + M_L(T - R/c) \right],$$

and L stands for a multi-index of spatial coordinates, i.e. $L = a_1 \dots a_l$ (e.g. $\partial_L = \partial^l / (\partial X^{a_1} \dots \partial X^{a_l})$). $M_L^{\oplus}$ are the relativistic Cartesian mass multipole moments of the Earth. To Newtonian order $\overline{W}$ is given by

$$\overline{W} = \overline{w} - \overline{w}(\mathbf{z}_{\oplus}) - \mathbf{a}_{\oplus} \cdot \mathbf{X} + \dots \ . \tag{4}$$

Here, $\overline{w}$ is the external potential resulting from Sun, Moon and the planets apart from the Earth in the global, barycentric system, $\mathbf{z}_{\oplus}$ denotes the geocenter and $\mathbf{a}_{\oplus}$ the acceleration of the geocenter. For the gravito-magnetic vector potential W_a we may neglect the external part and the self-part is essentially determined by the spin-vector $\mathbf{S}_{\oplus}$ of the Earth

$$\overline{W}_a \simeq -\frac{G}{2} \frac{\mathbf{X} \times \mathbf{S}_{\oplus}}{R^3} + \dots \ . \tag{5}$$

The dominant general relativistic effects are related with the problem of time. For that reason I would like to discuss some aspects of great practical importance (e.g. for high precision navigation in the solar system) related with the realization of time scales in the vicinity of the Earth in some detail. To this end we consider some clock whose elapsed proper time interval $d\tau$ is given by

$$-c^2 d\tau^2 = dS^2 = -\left(1 - \frac{2W}{c^2}\right) c^2 dT^2 + d\mathbf{X}^2 + O(2) \tag{6}$$

or

$$\dot{\tau} \equiv \frac{d\tau}{dT} = \left(1 - \frac{W}{c^2} - \frac{1}{2} \frac{\mathbf{V}^2}{c^2}\right) + O(4). \tag{7}$$

Here, the W-term indicates the (1st order) gravitational redshift effect. The best experimental verification of it is still the Vessot-Levine Gravity Probe-B experiment where a hydrogen maser clock on board a Scout rocket (Fig.

1) was flown to an altitude of about 10 000 km and its frequency compared to a similar clock on the ground (Vessot and Levine 1979, Vessot et al., 1980). The flight path was carefully monitored using tracking data and highly precise corrections for Doppler and atmospheric shifts were applied.

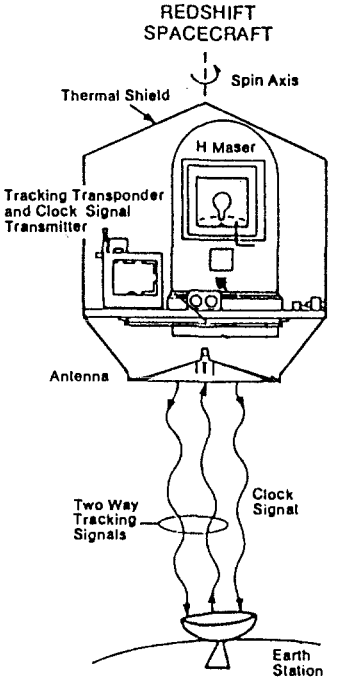

Fig. 1. The Gravity Probe-A redshift experiment

Writing the gravitational redshift of frequencies in the form

$$\frac{\Delta\nu}{\nu} = (1+\alpha)\frac{\Delta W}{c^2} \tag{8}$$

the experiment gave

$$|\alpha| < 2 \times 10^{-4}.$$

For the problem of terrestrial clocks we can transform the metric to coordinates co-rotating with the Earth (Soffel 1989). In these coordinates the space-time component of the metric has a $\mathbf{V}_\oplus/c = (\boldsymbol{\Omega} \times \mathbf{X})/c$ term and in the time-time part W is replaced by the geopotential,

$$W \to W_{\text{geo}} = W + (\boldsymbol{\Omega} \times \mathbf{X})^2/2. \tag{9}$$

E.g. for earthbound clocks ($d\mathbf{X} = 0$) we get

$$\dot{\tau} \simeq (1 - W_{\text{geo}}/c^2) \tag{10}$$

and

$$\frac{(d\tau)_1}{(d\tau)_2} = \frac{f_2}{f_1} = \frac{1-(W_{\text{geo}}/c^2)_1}{1-(W_{\text{geo}}/c^2)_2} \simeq \frac{[1+g(\psi)\cdot h/c^2]_1}{[1+g(\psi)\cdot h/c^2]_2}. \tag{11}$$

Here, g is the latitude ψ dependent gravity acceleration and h the height of the clock above the geoid (some equipotential surface at mean sea level). This implies e.g. that the readings of the reference clocks in Braunschweig (PTB), FRG and in Boulder (NBS), Colorado, differ by some 5.4μs/y and one should keep in mind that the "accuracy" (frequency stability) of present clocks is of order $\simeq 10^{-14} - 10^{-15}$ and improvements in the art of metrology are very rapid indeed. Not only for the comparison of clock readings is the gravitational redshift of practical importance but also for the dissemination of time, which nowadays is usually performed by means of the Global Positioning System (GPS). The GPS system comprises presently about 15 operating satellites with cesium or rubidium clocks on board flying at an altitude of about 20 000 km and emitting time signals which can be used by GPS receivers for positioning and navigational purposes.

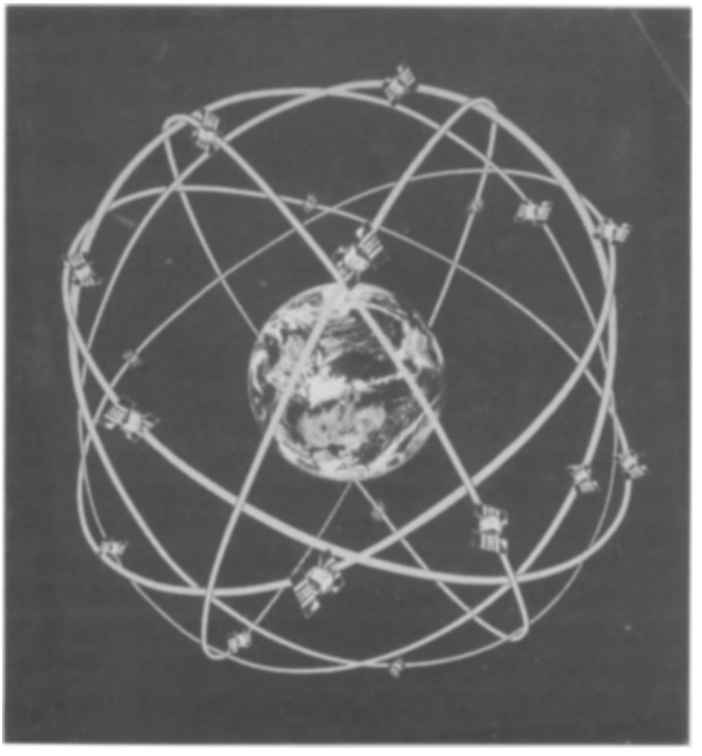

Fig. 2. The Global Positioning System (GPS)

Let me now describe how some geocentric time scale can be realized (e.g. Soffel and Brumberg 1992). First, let me note that the definition of the SI-second is of local nature, i.e. it depends upon the clock's position:

Def. (SI-second) = duration of 9 192 631 770 periods of the radiation that corresponds to the two hyperfine levels of the ground state of the Cs 133 atom.

Because of this local nature of the SI-second the times scales TT (or TDT) and the international atomic time TAI have been introduced w.r.t. the geoid:

"TT be a time scale differing from T = TCG uniquely by a constant rate, its unit of measurement being chosen such that it agrees with the SI second on the geoid".

Here, T = TCG is the geocentric coordinate time. Because of (2) and (7) we get

$$\tau = T - S(T)/c^2 + O(4) \tag{12a}$$

$$\dot{S} = W_{\text{geo}} + \overline{W} \simeq (W_0 - g \cdot h) + \overline{W} \tag{12b}$$

and we can split $S(T)$ into a secular and some remaining part

$$S(T) = S^* T + S_R(T) \tag{13}$$

with

$$S^* = W_0 = \text{const.} \tag{14}$$

Therefore, starting from the geocentric coordinate time T the time scale TT might be defined by

$$\text{TT} = k_E\, T \tag{15}$$

with

$$k_E = 1 - c^{-2} S^* = 1 - 6.97 \times 10^{-10}. \tag{16}$$

Now, because of their stochastic and non-ideal nature real clocks might *not* indicate (ideal) proper time. For that reason international atomic time is defined by averaging over the readings of several reference clocks (i) in the world after corrections for the height above the geoid (and tidal terms) have been applied. We may write

$$\text{TAI} = \text{mean}\left(\tau_{\text{act}} + c^{-2} S_R(T)\right)_i \tag{17}$$

and one realization of terrestrial time TT is given by

$$\text{TT(realized)} = \text{TAI} + 32.184\,\text{s}. \tag{18}$$

For all of this to work clocks are synchronized by means of coordinate time, i.e. two clocks showing proper times τ_1 and τ_2 are called synchronous, if the corresponding T-values agree:

$$\tau_1 \quad \text{syn} \quad \tau_2 \qquad \leftrightarrow \qquad T_1 = T_2. \tag{19}$$

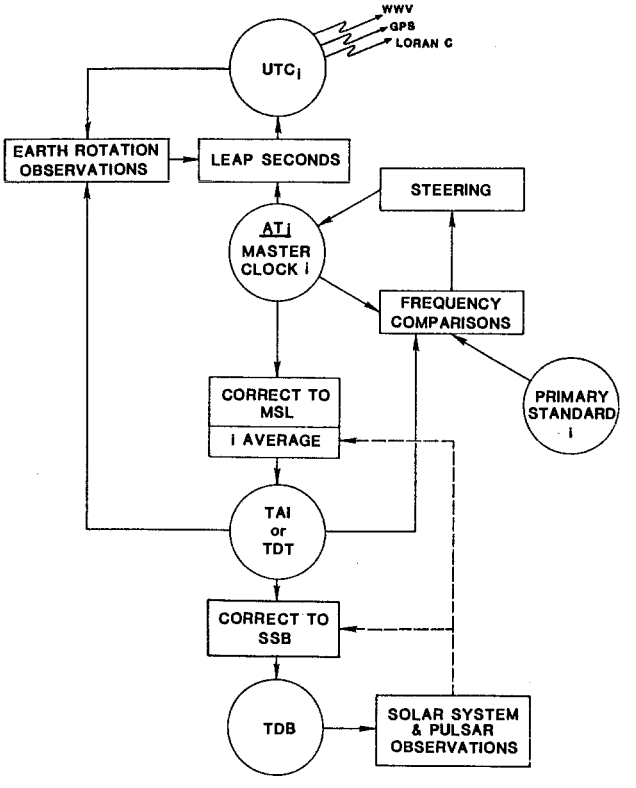

Fig. 3. Various operationally realized time scales and their connection (from Backer and Hellings 1986)

Fig. 3 shows the connection between various time scales used for practical purposes (from Backer and Hellings 1986). Here, TDT is the old notation for TT. MSL stands for mean sea level, UTC for Coordinated Universal Time, SSB for the solar system barycenter and TDB is barycentric dynamical time.

3 Anomalous perigee advances

Often the Eddington-Robertson parameters β and γ are included in the metric according to

$$G_{00} = -1 + 2W/c^2 - 2\beta W^2/c^4 + O(6) \qquad (20a)$$

$$G_{ab} = \delta_{ab}\left(1 + 2\gamma W/c^2\right) + O(4). \tag{20b}$$

β gives a measure of the non-linearity of the metric theory of gravity and γ is the space-curvature parameter. In Einstein's theory of gravity $\beta = \gamma = 1$. With these parameters the relativistic perigee advance of some satellite or planetary orbit in the central field of mass M is given by

$$\Delta\omega = 2\pi(2\gamma + 2 - \beta)\frac{GM}{c^2 a(1-e^2)} \quad \text{rad/rev.} \tag{21}$$

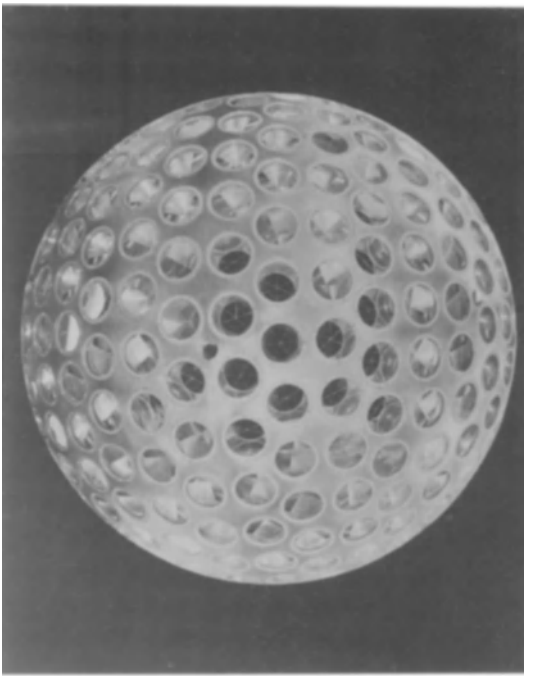

Fig. 4. The LAser GEOdynamical Satellite LAGEOS

Let me start with a discussion of satellite orbits. At this place I would like to mention that we are now in a fortunate position to have the most general post-Newtonian equations for satellite motion e.g. in the local geocentric system. In the DSX-framework this geodesic equation can be obtained from the Lagrangian

$$\mathcal{L} = 1 - \frac{d\tau}{dT} = \frac{1}{2}\mathbf{v}^2 + W + \frac{1}{c^2}\left[\frac{1}{8}\mathbf{v}^4 - \frac{1}{2}W^2 + \frac{3}{2}W\mathbf{v}^2 - 4W_a v^a\right] \tag{22}$$

and the complete PN expressions for W and W_a can be found in (Damour et al., 1991, 1992). Measurements of certain satellite orbits can be achieved by means of Satellite Laser Ranging (SLR). To this end lasergeodynamical satellites have to be used such as LAGEOS (Fig.4), a completely passive satellite with a radius of 30 cm and a weight of 410 kg. Its surface is completely covered with 426 laser retroreflectors which reflect incoming laser

light back in the same direction as that of the incoming pulses. One typically works with laser pulses of about 200 ps or 6 cm length at pulse rates of about 4 Hz. Presently achieved accuracies are at the cm level. Such SLR measurements are routinely used for the determination of Earth's rotation parameters and the the Earth's mass multipole moments. Mark Vincent (1984) has tried to determine the PPN parameters β and γ from LAGEOS data; however his derived values were never officially published because at that time the numerical program used (UTOPIA from the University of Texas) had still serious problems with relativity, which meanwhile have been solved. An italian group around A. Milani (1992) has tried in vain to determine values for β and γ just from LAGEOS data with an independent satellite program. The problem is that the LAGEOS orbit essentially determines the value for $GM_\oplus$, and since the eccentricity of the LAGEOS orbit is very small the relativistic deviations from Newton's $1/r^2$-law are mainly absorbed in the $GM_\oplus$-value. Other relativistic effects like the anomalous perigee advance might be absorbed by the higher multipole moments of the Earth. Likely one has to use the data of a second satellite with a different value for the semi-major axis to get values for the relativistic parameters.

As far as the relativistic advances of planetary perihelia is concerned we have to address the old solar quadrupole controversy. In the PPN framework the anomalous advance of planetary perihelia is given by

$$\Delta\varpi = 2\pi(2\gamma + 2 - \beta)\frac{GM}{c^2 a(1-e^2)} + \frac{3\pi R_\odot^2}{a^2(1-e^2)^2} J_2^\odot, \qquad (23)$$

where $J_2^\odot$ is the quadrupole moment of the Sun. Probably the most reliable determination of $J_2^\odot$ results from helioseismological measurements of the 5 min oscillations of the Sun (Brown et al. 1989). These measurements indicate that $J_2^\odot$ is of order 2×10^{-7} and therefore can be neglected for the problem of relativistic advance of planetary perihelia. In that case data from planetary radar ranging (mainly to Mercury) gave

$$\frac{2\gamma + 2 - \beta}{3} = 1.00 \pm 0.002$$

and if the value for γ is taken from gravitational time delay measurements (see below) one obtains (Shapiro 1990)

$$\beta = 1. \pm 0.003.$$

Radar measurements determine the anomalous perihelion shift of Mercury's orbit (42.98″/cen.) with an accuracy of about 0.5%. It is interesting to compare these numbers with those for certain binary pulsar systems, such as PSR 1913+16 or PSR 1534+12. For these pulsars the times of arrival (TOAs) of radio pulses can be determined extremely accurately (at the 3 μs level in fortunate cases; Taylor et al. 1992) from which high-precision system-parameters can be deduced. The anomalous periastron advance for

PSR 1913+16 [PSR 1534+12] was found to be 4.23 [1.76]°y^{-1} and the agreement with the prediction from Einstein's theory is better than about 1% (Taylor et al. 1989).

4 "Lense-Thirring" effects (gravitomagnetism)

Gravitomagnetic effects arise from the time-space components of the metric tensor or of the gravito-magnetic potential W_a. There is indirect evidence for the occurence of gravito-magnetism in Nature (Nordtvedt 1988). If one writes the metric tensor in barycentric coordinates then one will find a term

$$g_{0i} = \Delta \frac{GM_\oplus v^i_\oplus}{c^2 r}, \tag{24}$$

where $\Delta = 4$ in Einstein's theory of gravity. If we then transform the metric to the geocentric system this term is precisely canceled in Einstein's theory in accordance with the equivalence principle; if we keep the parameter Δ, however, we will find anomalous distance oscillations of satellite orbits proportional to $\Delta - 2\gamma - 2$. Assuming an accuracy of about 10 cm for the LAGEOS orbit one finds (Nordtvedt 1988)

$$\Delta - 2\gamma - 2 \leq 0. \pm .004.$$

As is well known gravitomagnetism leads to an additional spin precession ("dragging of inertial frames") of some (torque free) gyroscope w.r.t. "fixed stars", often referred to as the Schiff (or Lense-Thirring) precession. In Einstein's theory to PN order the total relativistic precession rate in the local frame of the gyro is given by

$$-\frac{1}{2}\frac{\mathbf{v}\times\mathbf{a}}{c^2} + \frac{3}{2}\frac{\mathbf{v}\times\nabla W}{c^2} + 2\frac{\nabla\times\mathbf{W}}{c^2}. \tag{25}$$

Here, the first term describes Thomas precession, the second term the geodetic (or de Sitter-Fokker) precession and the last term is the gravito-magnetic Schiff precession. Measuring this precession of spin axes is one goal of the well known Stanford gyroscope experiment (NASA'a Gravity Probe B, GP-B; see e.g. Everitt 1974, Lipa et al., 1974, Worden et al., 1974, Lipa et al., 1978, Bardas et al., 1989). The goal of the experiment is to measure the geodetic effect to better than 0.01% and the frame-dragging effect to better than 1%.

According to the latest design a drag free satellite will house two pairs of gyros and a 4 m focal length reference telescope with folded optics, all made of fused quartz and attached to a rigid quartz block. The quartz block is cooled by liquid helium to a temperature of 1.6 degrees K. The four gyros bear coatings of superconducting niobium, are electrostatically levitated and are spun by jets of helium gas until they rotate with about

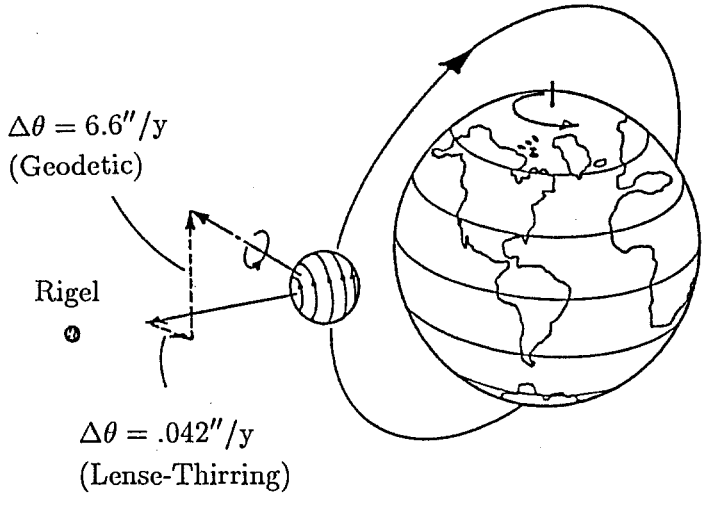

Fig. 5. NASA's Gravity Probe B (GP-B) experiment. The relativistic precession rates of gyroscope whose spin vector is parallel to the line of sight to a guide star and located in the orbital plane of a 650 km polar orbit. The geodetic precession is in the plane of the orbit and has a predicted value in Einstein's theory of 6.6 arc sec/y. The Lense-Thirring precession is in the plane of the celestial equator and in the same sense as the Earth's rotation; its predicted value is 0.042 arc sec/y

2000 rev./s. After the gyros rotate with their nominal speed the helium gas is evacuated so that they will lose only a quarter of a percent of their spin rate over the period of a year. the readout of the gyros' spin axes takes advantage of the magnetic (London) moment of rotating superconductors aligned with the spin axes. A tilt of the spin axes can then be monitored with SQUID magnetometers.

The present status of GP-B is that the construction of major ground-based test facilities have been completed and first tests have been successful. There always has been a problem with the reference star for the pointing of the telescope axis, since the proper motion of the originally selected star Rigel is not known with sufficient accuracy. For that reason one is presently looking for other star candidates with sufficient radio brightness such that it can be monitored also with VLBI.

In the motion of artificial satellites gravitomagnetism leads to a secular nodal drift of satellite orbits (the satellite's orbital angular momentum represents the spin vector in this case). However, this gravito-magnetic drift of the nodes of satellite orbits is very small; e.g. for the LAGEOS orbit it is roughly comparable with the effect of the Newtonian $l = 12$ mass multipole

moment of the Earth. Because these multipole moments cannot be measured with such high accuracy presently it seems impossible to recover the gravito-magnetic precession from the analysis of one single satellite orbit. However, in first order perturbation theory the nodal drift due to multipole moments is of the form

$$[\Delta\Omega(Y_l^m)]_{\text{sec.}}^{(1)} = \cos I \cdot \Psi(a, e, \text{even powers of } \sin I), \tag{26}$$

where I, a, e are inclination, semi-major axis and eccentricity of the orbit. This implies that if we consider a second satellite wit the same values for a and e but with an inclination $I' = 180° - I$, then in the sum of the two nodes the contributions from the multipole moments cancel (to a very high degree). This ides goes back to Ignazio Ciufolini (1986a,b; see also Casotto et al., 1990) and is now called the LAGEOS III problem (the orbit of LAGEOS II, expected to be launched in October 1992, will have a different inclination). Two assessment studies (one from the Italian group and one from the University of Texas in Austin) and an evaluation by some NASA advisory panel came to the conclusion that $\dot{\Omega}_{\text{LT}}$ might be measurable in such a LAGEOS I+III experiment with $\sim 10\,\%$ accuracy, given three years of data. Presently the launch of LAGEOS III is scheduled for 1994 with a Thor-δ rocket, whereby the precise orbit requirements might be fulfilled.

5 Geodetic precession

The geodetic precession in Einstein's theory of gravity should lead to an anomalous precession of the lunar node and perigee of the order of $2''$ per century. Writing

$$\dot{\omega}_{\text{M}} = 2''/100\text{y} \times (1 + h), \tag{27}$$

Shapiro et al. (1988) were first able to measure h by means of Lunar Laser Ranging (LLR) data. LLR data obtained between the years 1970 and 1986 were analyzed and used for the simultaneous estimation of 335 parameters out of which about 250 are time dependent Earth's rotation parameters. It is amusing to note that only from these data (i.e. only from measuring the distance to certain laser reflectors on the lunar surface) the mass of Jupiter could be determined with an accuracy of about 10%. The analysis of Shapiro et al. (1988) gave

$$h = 0.019 \pm .010,$$

where systematic errors had been taken into account. Two more recent measurements of the lunar geodetic precession by the JPL group (Dickey et al., 1989) and by the german group (Müller et al., 1991) essentially confirmed the Shapiro et al. valué for h. In Müller et al. (1991) about 6300 LLR data from the period 1969 – 1990 were used to determine the value for h (among

other parameters relevant for tests of metric theories of gravity); they find $h = 0.002 \pm 0.01[0.002]$, were the formal 1σ error is given in square brackets.

6 Light deflection, signal retardation

In Einsteinian gravity light rays are geodesics w.r.t. $g_{\mu\nu}$ and along light rays $ds = 0$ (light rays are null geodesics). From these conditions the equation for some light ray to PN order (with parameter γ) in the spherical field of the Sun takes the form

$$\mathbf{x}(t) = \mathbf{x}_0 + \hat{\mathbf{n}}c(t - t_0) - \frac{(\gamma+1)GM_\odot}{c^2}\left[\hat{\mathbf{n}}\ln\left(\frac{r + \mathbf{x}\cdot\hat{\mathbf{n}}}{r_0 + \mathbf{x}_0\cdot\hat{\mathbf{n}}}\right) + \frac{\mathbf{d}}{d^2}[r - r_0 - c(t - t_0)]\right]. \tag{28}$$

Here, $\hat{\mathbf{n}}$ is some "Euclidean" unit vector in the direction of the ray at $\mathbf{x}_0$ and $\mathbf{d}$ connects the center of the Sun with the point of closest approach of the unperturbed light ray. E.g. the angle of light deflection of starlight by the Sun is given by

$$\Delta\varphi \simeq 1.75'' \left(\frac{1+\gamma}{2}\right)\frac{R_\odot}{d}. \tag{29}$$

The accuracy of present ground based optical measurements is fairly poor. However, instead of measuring the light deflection in space one can also measure the corresponding effect in the time domain, called the (Shapiro) gravitational time delay. Time delay experiment involving the VIKING spacecrafts (Reasenberg et al., 1979) gave

$$\gamma = 1. \pm 0.002.$$

The gravitational time delay plays an important role in Verly Long Baseline Interferometry (VLBI).[1] In VLBI the radio signal of distant radio sources (quasars) are received by several VLBI antennas. These antennas usually have a great spatial separation from each other and no direct broadband connection. At each of the stations the radio signal is transformed from the GHz to the MHz region and together with time tags from some local oscillator (typically H-masers) serving as phase reference they are recorded on video tape. An interference signal is then produced by cross-correlation of the tape data in a correlator. VLBI is one of the most precise astrometric techniques. With a modern Mark III system residuals now ly in the 30 – 50 ps region, i.e. for angles one achieves sub mas accuracies. This implies that

[1] Very accurate PN VLBI theories have been published by Kopejkin (1990) and Soffel et al., (1991).

the light deflection (more precisely the time delay) in the field of the Sun is measurable even for sources lying close to 180° from the Sun. A recent analysis of VLBI data by Robertson et al., (1991) gave

$$\gamma = 1. \pm .002[.00096].$$

In this study 342 810 VLBI observations from the geodetic programs POLARIS (the polar motion analysis), IRIS and the Crustal Dynamics Programme (CDP) were analyzed. Fig. 6 shows the relativistic deflection in mas for the 26 sources that are currently observed in the IRIS (International Radio Interferometric Surveying) program.

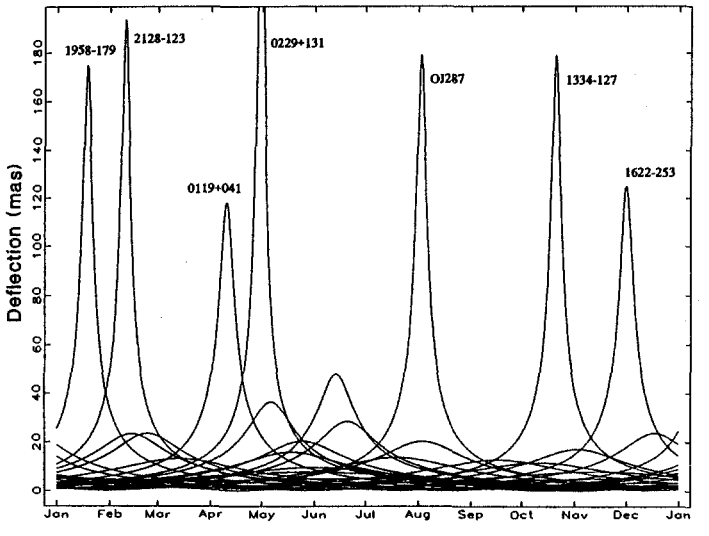

Fig. 6. Light deflection angles in mas for the 26 sources that are currently observed in the IRIS program (from Robertson et al., 1991)

Finally, I would like to mention that the Shapiro effect plays a substantial role in any relativistic timing model for the analysis of times of arrival from (binary) radio pulsars.

7 Einsteinian dynamics of the solar system

Numerical ephemerides (such as the DE-programs from JPL) for the solar system or the motion of the Moon with a certain number of fit-parameters might be used together with certain data sets to get information about relativistic effects in the motion of the solar system (such modern ephemerides usually comprise the relativistic equations of motion for point masses (the EIH equations) in addition to the Newtonian equations for extended bodies). Using such PPN-ephemerides with LLR and radar data to Mercury, Venus and Mars (VIKING) Ron Hellings (1983) obtained

$$\gamma - 1 = (-0.7 \pm 1.7) \times 10^{-3}$$
$$\beta - 1 = (-2.9 \pm 3.1) \times 10^{-3}$$
$$J_2^{\odot} = (-1.4 \pm 1.5) \times 10^{-6}.$$

The more recent analysis by Müller et al. (1991) using LLR data gave

$$\gamma - 1 = (-0.08 \pm 2.0[0.46]) \times 10^{-3}$$
$$\beta - 1 = (-0.21 \pm 1.5[0.24]) \times 10^{-3}.$$

Measurements of the Nordtvedt effect (the strong equivalence principle) are discussed in Shapiro et al. (1976), Dickey et al. (1989) and Müller et al. (1991); for a treatment of anomalous PPN-effects the reader is referred to Will (1981, 1984). Finally, the time variation of the gravitational constant G was determined with a variety of methods. Radar measurements to VIKING gave an upper limit of $3 \times 10^{-11}\,\mathrm{yr}^{-1}$ by Reasenberg (1983), $(0.2 \pm 0.4) \times 10^{-11}\,\mathrm{yr}^{-1}$ by Hellings (1983); lunar laser ranging data lead to (Müller et al. 1991)

$$\dot{G}/G = (0.01 \pm 1.04) \times 10^{-11}\,\mathrm{yr}^{-1}.$$

A more recent determination using data from PSR 1913+16 by Damour et al. (1988) gave $(1.0 \pm 2.3) \times 10^{-11}\,\mathrm{yr}^{-1}$.

8 Further OUTLOOK into the future

So far we have been dealing with measurements which have already been performed or are in preparation. In this chapter I would like to present some outlook into the future and ask about the prospects for highly precise improved measurements of general relativistic effects in the solar system.

8.1 Optical interferometry in space

To measure the gravitational light deflection in the field of the Sun with great precision future optical interferometry in space looks very promising. One envisaged experiment of that kind is POINTS, standing for "high Precision Optical INTerferometer in Space", where one tries to achieve accuracies of the order of $5 \times 10^{-6''}$ with a Michelson interferometer in orbit (Reasenberg et al., 1982, 1989; Reasenberg and Chandler 1989). Such an accuracy can only be achieved with ultrastable and thermally controlled materials and a real time metrology of the apparatus with laser interferometry to 10^{-12} m. Expected light deflection angles at the limb of the Sun due to post-post-Newtonian effects, due to the quadrupole moment of the Sun and the angular momentum of the Sun are $(\Delta\varphi)_{\mathrm{PPN}} \sim 11\,\mu$as, $(\Delta\varphi)_{\mathrm{J}_2^\odot} \sim 0.2\,\mu$as and $(\Delta\varphi)_{\mathrm{L}_\odot} \sim 0.7\,\mu$as (Epstein and Shapiro 1980). These numbers imply, that POINTS certainly will not be able to measure the $J_2^\odot$ effect and the gravito-magnetic light bending. Probably, the main goal will be a major improvement of the measurement of the space curvature parameter γ.

8.2 Improved perihelia advance measurements

There is some chance to improve the measurement of the perihelion advance of Mercury. One possibility would be by means of a Mercury lander equipped with a radio transponder. There is some chance to improve the topography model for the surface of Mercury by observing same areas on Mercury from different orbital positions by means of the Arecibo or Goldstone anntennas. For more details see Shapiro (1990).

8.3 Improved redshift experiments

New improved redshift experiments might be carried out in the future. An old idea is to send some atomic clock (H-maser) in a spacecraft very close to the Sun and to compare the readings of this clock with those of terrestrial clocks. Another possibilty is to compare very distant clocks. e.g. pulsar clocks, with terrestrial clocks and to analyse the annual deviation resulting from the eccentricity e of the Earth's orbit

$$\Delta\tau \simeq \frac{2}{c^2}\sqrt{GM_\odot a}\, e \sin E \simeq 1.6 \times 10^{-3} \sin E\,(\mathrm{sec}) \qquad (30)$$

An ideal candidate for such an astronomical clock is the ms pulsar PSR 1937+21, whose pulsar clock is stable to $\sim 10^{-14}$ or better over a year or more! Fig. 7 shows a comparison between the pulsar clock PSR 1937+21 and the best clocks of the world. Here, the proper motion problem of the pulsars could be solved with VLBI positioning.

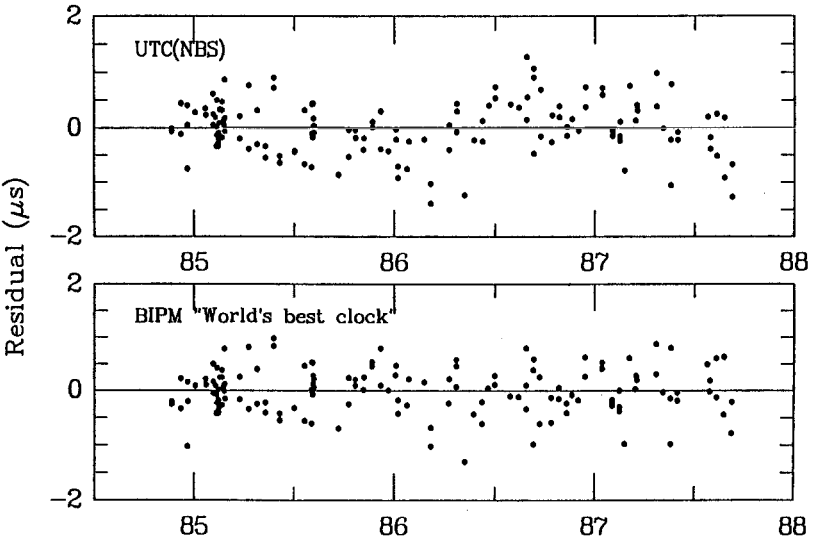

Fig. 7. Post-fit arrival time residuals for PSR 1937+21 relative to UTC (NBS) (top) and to the BIPM "World's best clock" (bottom) (from Taylor 1989)

References

Backer, D.C., Hellings, R.W. (1986): Ann. Rev. Astron. Ap. **24**, 537

Bardas, D., et al. (1989): in "Proc. of the 5th Marcel Grossmann Meeting on General Relativity", ed. by D.Blair, M.Buckingham and R.Ruffini, World Scientific

Barlier, F. et al. (1991): STEP, Satellite Test of the Equivalence Principle, Assessment Study Report, ESA – NASA, SCI(91)4, January 1991

Brown, T.M. et al. (1989): Astrophys. J. **343**, 526

Casotto, S., Ciufolini, I., Vespe, F., Bianco, G. (1990): Nouvo Cim. **105B**, 589

Ciufolini, I. (1986a): Phys. Rev. Lett. **56**, 278

Ciufolini, I. (1986b): Found. of Physics **16**, 259

Damour, T., Gibbons, G.W., Taylor, J.H. (1988): Phys. Rev. Lett. **61**, 1151

Damour, T., Soffel, M.H., Xu, C. (1991): Phys. Rev. D **43**, 3273

Damour, T., Soffel., M.H., Xu, C. (1992): Phys. Rev. D **45**, 1017

Dicke, J.O. et al. (1989): JPL preprint

Epstein, R., Shapiro, I.I. (1980): Phys. Rev D **22**, 2947

Everitt, C.W.F. (1974): in: "Experimental Gravitation: Proc. of Course 56 of the Intern. School of Physics Enrico Fermi", ed. B. Bertotti, Academic Press, New York

Hellings, R.W. (1983): Proc. 10th Intern. Conf. General Relativity and Gravitation, (Padova), ed. B. Bertotti et al. (Dordrecht: Reidel)
Kopejkin, S.M. (1990): Astron. J. (USSR) **67**, 10
Lipa, J.A., Everitt, C.W.F. (1974): in: "Experimental Gravitation: Proc. of Course 56 of the Intern. School of Physics Enrico Fermi", ed. B. Bertotti, Academic Press, New York
Lipa, J.A., Everitt, C.W.F. (1978): Acta Astronautica **5**, 119
Milani, A. (1992): private commmunication
Müller, J., Schneider, M., Soffel, M.H., Ruder, H. (1991): Astrophys. J. **L101**, 382
Nordtvedt, K. (1988): Phys. Rev. Lett. **61**, 2647
Reasenberg, R.D., Shapiro, I.I. (1982): Acta Astronautica **9**, 103
Reasenberg, R.D. (1983): Phil. Trans. R. Soc. Lond. **A310**, 227
Reasenberg, R.D. (1989): in "Proc. of the 5th Marcel Grossmann Meeting on General Relativity", ed. by D.Blair, M.Buckingham and R.Ruffini, World Scientific
Reasenberg, R.D., Chandler, J.F, (1989): in "Proc. of the 5th Marcel Grossmann Meeting on General Relativity", ed. by D.Blair, M.Buckingham and R.Ruffini, World Scientific
Robertson, D.S., Carter, W.E., Dillinger, W.H. (1991): Nature **349**, 768
Shapiro, I.I. (1990): in "General Relativity and Gravitation", ed. Ashby, Bartlett and Wyss, Cambridge University Press
Shapiro, I.I., Counselman, C.C., King, R.W. (1976): Phys. Rev. Lett. **36**, 555
Shapiro, I.I., Reasenberg, R.D., Chandler, J.F., Babcock, R.W. (1988): Phys. Rev. Lett. **61**, 2643
Soffel, M.H. (1989): "Relativity in Astrometry, Celestial Mechanics and Geodesy", Springer, Berlin
Soffel, M.H., Müller, J., Wu, X., Xu, C. (1991): Astron. J. **101**, 2306
Soffel, M.H., Brumberg, V.A. (1992): Celest. Mech., in press
Taylor, J.H. (1989): in "Timing Neutron Stars", ed. H.Ögelman and E.P. van den Heuvel, NATO ASI Series, Kluwer Academic Publishers, Dordrecht
Taylor, J.H., Weisberg, J.M. (1989): Astrophys. J. **345**, 434
Taylor, J.H., Wolszczan, A., Damour, T., Weisberg, J.M. (1992): Nature **355**, 132
Thomas, J., et al. (1989): Phys. Rev. Lett. **63**, 1902
Vessot, R.F., Levine, M.W. (1979): Gen. Rel. and Grav. **10**, 181
Vessot, R.F. et al. (1980): Phys. Rev. Lett. **45**, 2081
Vincent, M.A. (1984): "The Determination of the Post-Newtonian Parameters in Gravitational Theory Using Laser Ranging to the LAGEOS Satellite", Ph.D. Thesis, Uni. of Texas, Austin
Worden, P.W., Everitt, C.W.F. (1974) in: "Experimental Gravitation: Proc. of Course 56 of the Intern. School of Physics Enrico Fermi", ed. B. Bertotti, Academic Press, New York
Worden, P.W., Everitt, C.W.F., Bye, M. (1990): Satellite Test of the Equivalence Principle, Science Requirements Documents, Stanford University, May 1990
Will, C. (1981): "Theory and Experiment in Grvaitational Physics", Cambridge University Press, Cambridge
Will, C. (1984): Phys. Rep. **113**, 345

New Results for Relativistic Parameters from the Analysis of LLR Measurements

Jürgen Müller [1], Manfred Schneider [1],

Michael Soffel [2], Hanns Ruder [2]

[1]Forschungseinrichtung Satellitengeodäsie, Technische Universität München, D-8000 München, Fed. Rep. of Germany
[2]Lehr- und Forschungsbereich Theoretische Astrophysik, Universität Tübingen, D-7400 Tübingen, Fed. Rep. of Germany

Abstract: The theoretical model for the analysis of Lunar Laser Ranging (LLR) data is described and results of a least-squares fit are presented. Using observations of more than 21 years we determined parameters concerning the Earth-Moon system (e.g. the mass of the system, the lunar tidal acceleration etc.), the station-reflector geometry (e.g. coordinates, Earth's orientation parameters etc.), and other physical parameters like the solar quadrupole moment $J_2^\odot$. Besides these parameters others suitable for testing metric theories of gravity in the first post-Newtonian approximation can be determined with great accuracy. These are the PPN (parametrized post-Newtonian) parameters γ and β, the Nordtvedt parameter η, the geodetic precession Ω_{GP} of the lunar orbit, as well as a possible time variation of the gravitational constant $\dot{G}/G$. These relativistic parameters are discussed in detail including the derivation of realistic errors.

1 Introduction

During the Apollo 11 mission (July, 1969) the first retroreflector was deployed on the surface of the Moon. In the following years four other reflector arrays (two American: Apollo 14 and Apollo 15; and two French transported by unmanned Soviet landers: Luna 17 and Luna 21) were set up. Unfortunately measurements to one reflector (Luna 17) cannot be carried out, because it became dusty by the restart of the unmanned lander.

Since that time continuing laser range measurements from the Earth to the Moon are performed. 'Measurements' means recording the round-trip travel times of laser pulses from the laser station on the Earth to some

reflector on the Moon and back to the receiver. This value valid for a certain epoch is stored together with the actual weather data and some informations about possible measurement errors. The present-day accuracy of the LLR observations amounts to 0.1 nanoseconds corresponding to 3 centimeter in the Earth-Moon distance.

The LLR observations which are averaged to normal points (consisting of 3 to 400 lunar returns) are taken from five sites: the 2.7 m telescope of the McDonald observatory; the McDonald Laser Ranging Station (MLRS, situated in separate locations before and after a move in 1988); the Haleakala observatory on Maui, Hawaii; and the CERGA system in Grasse, France. At the moment Haleakala gets no financial support to continue Lunar Laser Ranging. At the McDonald station the lunar observation program is classified to the lowest priority. Fortunately the Australian laser station Orroral and the German laser station Wettzell improved their equipment for Lunar Laser Ranging. Soon they will be able to range to the Moon (first successful observations in Wettzell exist) and to close the gap from the loss of the American stations.

Today more than 7100 normal points obtained during the period from 1969 to 1991 are available. For each epoch the observed travel time can be compared with the theoretically computed travel time. With suitable analysis methods one can determine a lot of parameters describing the dynamics of the Earth-Moon system. Besides these parameters others suitable for testing metric theories of gravity in the first post-Newtonian approximation can be determined. However, the basis for a promising analysis is a consistent theoretical model.

2 Model

For about ten years our group has been active to develop the theoretical and practical prerequisites for the analysis of Lunar Laser Ranging measurements (Gleixner 1986; Schastok et al. 1989; Bauer 1989; Soffel 1989; Schneider 1989). In the last years efforts have been concentrated to revise and extend the model upon which the LLR data analysis is based and to determine relativistic parameters with high accuracy (Müller 1991).

Based upon Einstein's theory of gravity a fully consistent model for the LLR measuring process has been worked out at the first post-Newtonian level. This model employs essentially three coordinate systems: one global (barycentric) system in which the motion of the solar system is computed and the pulse propagation is described and two local, accelerated systems comoving with the Earth and the Moon respectively. In the geocentric (selenocentric) system the station (reflector) coordinates are defined. The relation between the global and the local coordinate systems can be obtained e.g. by means of the Brumberg-Kopejkin approach (Brumberg et al. 1988, 1989;

Kopejkin 1988). It allows to formulate equations for the coordinate transformations between different reference frames, to derive relativistic equations of motion (including spin-orbit and spin-spin coupling terms) and to describe the observations in a consistent manner.

The choice of this formalism has historical reasons. Equivalently the Damour-Soffel-Xu formalism (Damour et al. 1991) could be used to obtain all equations necessary for the LLR data analysis.

In the Brumberg-Kopejkin scheme Einstein's field equations

$$R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R = \kappa T^{\mu\nu} \tag{1}$$

are solved in the harmonic gauge in the first post-Newtonian approximation. Using the perfect fluid approximation for the description of the matter of the bodies and regarding the solar system as an isolated system the solution of Einstein's equations in the global system is obtained in the usual way.

For the local metric one starts with an ansatz describing the gravitational influence of the central body itself (self-part) and of all the other bodies of the system (external part). For the external part one starts with an expansion in terms of powers of the local coordinate, such that external bodies will only generate tidal effects in the local frame. The formalism is then completed by means of the asymptotic matching technique, i.e. in the region of common validity the global and the local metric can be transformed into each other by some suitable coordinate transformation and some condition which relates the origin of the local system with the matter distribution of the central body (e.g. with the center of mass of this body).

The resulting equations are used in computer programs for the LLR data analysis. In detail we have programs for the

1. computation of the ephemerides,
2. computation of the partials of the lunar and solar orbital elements with respect to the solve-for parameters,
3. computation of the partials of the Earth-Moon distance with respect to the solve-for parameters, calculation of the residuals and determination of the solve-for parameters (including the $1\,\sigma$ errors).

3 The Ephemeris Program

Starting with known initial conditions (position and velocity vectors as well as the masses of the bodies), in the ephemeris program the ephemerides of the solar system bodies are computed by numerical integration. The initial conditions are taken from the ephemeris DE200 which is produced by a combined analysis of different observation data at the Jet Propulsion Laboratory (JPL). DE200 is available for everyone.

Our ephemeris program for the major bodies of the solar system (Sun, planets, Moon, some asteroids) in harmonic barycentric coordinates contains the following elements:

- with respect to the translational motion of bodies:
  - the Einstein-Infeld-Hoffmann (EIH) equations for point masses;
  - the classical Newtonian equations of motion for the description of the anisotropy of the bodies (the lowest mass multipole moments of Earth and Moon up to degree and order 4 are included; for the Sun a quadrupole moment is taken into account);
  - the lunar tidal acceleration;
- with respect to the rotational motion of the Moon:
  - the Euler equations of rotational motion (torques ared produced by the Sun and the Earth including the influence of the quadrupole moment of the Earth);
  - a modification of these equations because of elastic behaviour and because of energy dissipation of the Moon;
  - the geodetic precession of the lunar spin;
- with respect to the rotational motion of the Earth[1]:
  - the 1980 IAU[2] nutation series for $\Delta\psi$ and $\Delta\epsilon$, supplemented by the two out-of-phase terms of the 18.6 year nutation period; the annual coefficients are corrected by a 2 mas (milliarcseconds) bias. The four coefficients of the 18.6 year nutation period can be estimated from LLR data by introducing well-known constraints for some ot these;
  - the precession angles of Lieske et al. (1977), but the luni-solar precession constant is fitted to the observations; the geodetic precession of the geocentric system is already considered in these angles.
  - the Earth orientation parameters are taken from BIH/IERS[3]; they are used as approximate values and can be determined in the LLR analysis.

The program for the computation of the partials is similiar to the ephemeris program; indeed, it is used in a simplified version.

In the program for the determination of the solve-for parameters all elements which are necessary for a least-squares fit (residuals, information matrix etc.) are computed, and finally the fit is performed.

---

[1] The angles describing the rotational motion of the Earth are not computed by integrating the corresponding Euler equations, but they are introduced from analytical models; they are especially needed for the coordinate transformations between barycentric and geocentric system.

[2] International Astronomical Union.

[3] Bureau International de l'Heure/International Earth Rotation Service.

Before the calculation of the residuals can be done, the station and reflector coordinates as well as the observed round-trip time have to be corrected by some systematic influences caused by the non-rigidity of the Earth, the definition of the local reference systems and the propagation of laser signals.

4 Corrections of the Station resp. Reflector Coordinates and the Light Travel Time

Tidal effects of the solid Earth (which cannot be regarded as a rigid body) on the station coordinates are considered and effects for plate motion are taken into account using the AM0-2 model by Minster and Jordan (1978). The rates of plate motion for the Pacific, the North American and the Eurasian plate can be estimated.

By means of the general coordinate transformation (e. g. Brumberg et al. 1988; Damour et al. 1991) geocentric station coordinates are related with corresponding barycentric quantities. A similiar transformation relates the selenocentric reflector coordinates with corresponding barycentric ones. In our program the transformation between the geocentric time TDT (terrestrial dynamical time) and the barycentric time TDB (barycentric dynamical time) is achieved by means of the Hirayama et al. (1987) series.

For the propagation of laser signals the light time equation containing the gravitational time delay (Shapiro effect) is employed. Here, both the influence of the Sun and the Earth are taken into account in the barycentric coordinate system. The model of Marini and Murray (1973) is introduced to correct for atmospheric influences.

5 Modifications of the Model

The EIH-equations are introduced in the form containing the metric parameter γ and β (e.g. Will 1981). Corresponding to the PPN framework further parts of the LLR model have to be modified when parametrizing γ and β, e.g. the light time equation or the equations for coordinate transformations. These changes also have to be done when affecting the model accuracy. Both metric parameters equal 1 in the general theory of relativity.

The gravitational constant is allowed to vary with time by introducing

$$G = G_0 \left(1 + \frac{\dot{G}}{G_0} \times \Delta t \right) . \tag{2}$$

The linearization (2) is only valid for weakly gravitating bodies which is true for the solar system. $\dot{G}/G$ equals 0 in Einstein's theory of gravity and belongs to its basic assumptions.

In contrast to Einstein's theory some other metric theories of gravity (e.g. the Brans-Dicke theory) involve a violation of the strong equivalence principle (Will 1981). If, according to the difference in gravitational self-energy, Earth and Moon would fall towards the Sun at different rates the lunar orbit about the Earth would be polarized in the direction of the Sun. This effect, called Nordtvedt effect can be accurately investigated by analysing LLR data. The additional contribution to the Earth-Moon distance being given by (Will 1981)

$$\Delta r_{\mathrm{EM}} = 9.2\,\eta\,\frac{\mathbf{x}_{\mathrm{ES}}\cdot\mathbf{x}_{\mathrm{EM}}}{r_{\mathrm{ES}}\,r_{\mathrm{EM}}}\ \mathrm{m}\,. \tag{3}$$

Similiary another term was added for the determination of the geodetic precession of the lunar orbit. Following a method similiar to that of Shapiro et al. (1988), we introduced a solve-for parameter h by an additional acceleration term of the form

$$\mathbf{a}_{\mathrm{GP}} = 2\,h\,\boldsymbol{\Omega}_{\mathrm{GP}}\times\mathbf{v}_{\mathrm{ME}} \tag{4}$$

in the equations of translational motion of the Moon. h indicates the deviation of the observed geodetic precession from the value predicted in Einstein's theory of gravity.

6 Solve-for Parameters

With such a model one can compute for a certain epoch the light travel time from some station to some laser reflector on the lunar surface and back to the receiver which can be compared with the observed value. For the results reported here we have performed a least-squares fit to 7100 LLR observations (normal points), spanning the period from 1969 to 1991.

We estimated two groups of parameters. The splitting in two groups is done to achieve a faster convergency for the parameters of interest. The first group includes all so-called main parameters of the Earth-Moon system (a total number of about 880 parameters, the most are introduced to model the Earth's orientation), the second group consists of parameters, with which one can test Einstein's theory of gravity in the first post-Newtonian approximation.

To the first group belong:

- geocentric coordinates of the ground stations which define a terrestrial reference frame;
- optional: rates of plate motion deviating from the predicted AM0-2 values;
- Earth's rotation (variations in the length of day);

- position of the rotational axis with respect to the solid Earth (polar motion on the Earth's surface);
- the luni-solar precession constant and the four coefficients of the 18.6 year nutation period; they describe the spatial shift of the rotational axis and indicate lacks in the analytical models for the description of the Earth's rotation in space introduced in the analysis;
- selenocentric coordinates of four retroreflectors which define a reference frame on the Moon;
- the rotation of the Moon for one initial epoch (physical libration angles);
- position and velocity of the Moon for this epoch;
- the mass of the Earth-Moon system times the gravitational constant G; the accurate knowledge of the lunar position and mass allows a more accurate determination of positions in the solar system;
- the lowest mass multipole moments of the Moon up to degree and order 3, i.e. the gravitational field of the Moon which deviates from a spherical mass distribution;
- the lunar Love number; it models the static perturbation of a homogeneous elastic sphere;
- the rotational energy dissipation parameter describing a delay time for the Moon to react upon gravitational perturbations;
  These two parameters give informations about the inner composition of the Moon;
- the lag angle indicating the lunar tidal acceleration $\dot{n}_{\mathrm{M}}$ resulting from the tidal bulge raised on the Earth.

To the second group belong:

- the quadrupole moment of the Sun $J_2^{\odot}$ (indicating the static flattening and the shape of the Sun);
- the space curvature parameter γ entering the light time equation and the equations of translational motion as well as the formulae of coordinate transformations;
- the non-linearity parameter β involved in the equations of motion and the formulae for coordinate transformations;
- the Nordtvedt parameter η (strong equivalence principle);
- the scale factor h for an additional (deviating from Einstein's theory) geodetic precession Ω_{GP} of the lunar orbit;
- the time variation of the gravitational constant $\dot{G}/G$; this parameter is important e.g. for the unification of the fundamental forces (gravitational, electromagnetic, weak and strong nuclear) or in cosmology for the description of the evolving universe.

7 Investigation of Realistic Errors

Results for the second group of parameters will be presented below. Once the main parameters from the first group have been estimated values for the six parameters of the second group are derived by means of a least-squares fit to weighted observations. Besides the new values for the solve-for parameters the fit gives formal $1\,\sigma$ errors (indicated below in square brackets) and the correlation matrix. Derived correlations for our six parameter of interest are given in Table 1.

Table 1. Derived correlations for our six parameters of interest

	$J_2^\odot$	γ	β	η	$\Omega_{\rm GP}$	$\dot{G}/G$
$J_2^\odot$	*	2	3	1	5	5
γ	2	*	7	2	3	4
β	3	7	*	2	3	5
η	1	2	2	*	1	1
$\Omega_{\rm GP}$	5	3	3	1	*	2
$\dot{G}/G$	5	4	5	1	2	*

The various integers indicate the correlations between the six parameters of the second group; 0: correlation $\leq$ 5%; 1: 5% –15%; ...; 9: 85% –95%; *: $\geq$ 95%.

In most cases realistic errors will be much greater than formal ones. We have estimated the various realistic errors given below by a combination of techniques: we have analysed the various correlations of parameters and estimated the influence of insufficient modelling and inaccurate model parameters. Main error sources for most parameters involve inaccurate Earth's orientation parameters (especially those of the early years) and nutation coefficients. For some parameters the insufficient modelling of plate tectonics or unmodelled perturbations caused by asteroids play a role. Further error sources are e.g. the lunar tidal acceleration, the inaccurate solar mass or the orbital elements of the Earth which might not sufficiently accurate be determined. Table 2 shows various error sources for our six parameters. Column 2 gives estimated errors resulting from weakly known Earth's orientation parameters. Column 3 shows errors resulting from inaccurate nutation coefficients. In column 4 differences to the estimated values are listed if one

neglects the asteroids in the ephemeris program. Finally, in column 5 corresponding differences are given for the two cases i) the AM0-2 model is used for the description of plate motion and ii) the rates of plate motion are estimated from LLR data.

Table 2. Various erros sources for our six parameters

Parameter	ΔEOP	ΔNut.	ΔAster.	ΔPl. Mo.
$J_2^\odot$	1.2×10^{-6}	0.5×10^{-6}	0.4×10^{-6}	0.2×10^{-6}
γ	0.6×10^{-3}	1.8×10^{-3}	0.1×10^{-3}	0.2×10^{-3}
β	0.1×10^{-3}	0.8×10^{-3}	0.1×10^{-3}	0.1×10^{-3}
η	0.5×10^{-3}	0.05×10^{-3}	0.01×10^{-3}	0.02×10^{-3}
Ω_{GP} [as/cy]	0.003	0.001	0.003	0.0002
$\dot{G}/G$ [yr^{-1}]	5.5×10^{-13}	0.9×10^{-13}	2.6×10^{-13}	0.3×10^{-13}

as/cy = arcseconds/century

The estimation of realistic errors finally has been checked by means of a "modified worst case analysis" (Bender 1990), where the total error of the observations is partly transfered to the solve-for parameters.

8 Results

We obtained for

- the quadrupole moment of the Sun:

$$J_2^\odot = (2.0 \pm 1.5\,[0.5]) \times 10^{-6}$$

where the estimated realistic error is given and the value in square brackets indicates the formal $1\,\sigma$ error.

We would like to remark that our value for $J_2^\odot$ has the same order as that derived from an analysis of the global five-minute oscillations of the Sun (Hill et al. 1982; Gough 1982; Campbell et al. 1983). A more recent investigation of the solar oscillations, however, seems to indicate that the value of $J_2^\odot$ might, in fact, be smaller and of the order of 2×10^{-7} (Brown et al. 1989).

$J_2^\odot$ obtained by analysing LLR data is slightly affected by some (hitherto unknown) systematic influences which can be reduced by using observations over a longer period of time. One error source which is still not considered in the above error value is a possible insufficient knowledge of the Earth's orbit. Would be found that the orbital elements taken from an analysis of different measurements (especially radar observations) would inaccurately be determined and had to be estimated from the LLR data the formal error of the solar quadrupole moment would increase to 3.1×10^{-6}. That would mean $J_2^\odot$ could not significantly be determined from LLR observations. However, it would be possible to indicate an upper limit of the order of 6×10^{-6} for $J_2^\odot$. At the moment we investigate that issue to abolish the uncertainties.

The value for the quadrupole moment of the Sun enters the expression for the anomalous advance of a planet. In the PPN framework the change of ω per revolution is given by

$$\dot{\omega}_{\text{theo}} = 6\,\pi\,\frac{GM_\odot}{a\,(1-e^2)\,c^2}\,\lambda \tag{5}$$

with

$$\lambda = \frac{2+2\,\gamma-\beta}{3} + \frac{R_\odot^2\,c^2}{2\,GM_\odot\,a\,(1-e^2)}\,J_2^\odot\,. \tag{6}$$

For $\beta = \gamma = 1$ (Einstein case) we find for Mercury

$$\begin{aligned}\dot{\omega}_{\text{theo}} &= 42.98\,\left(1+3\times10^{-3}\,J_2^\odot\right)\ \text{as/cy}\\ &= 43.24 \pm 0.2\ \text{as/cy}\,.\end{aligned}$$

This has to be compared with the observed value of

$$\dot{\omega}_{\text{obs}} = 43.11 \pm 0.21\ \text{as/cy}$$

resulting from planetary radar measurements (Shapiro et al. 1976). The observed and the theoretical value are in good agreement within the given errors.

Besides the correlations of $J_2^\odot$ with the other five parameters given in Table 1 this value is highly correlated with the station coordinates and slightly with the lunar tidal acceleration and the $\Delta\psi$-coefficients of the 18.6 year nutation period, which has to be taken into account in the error analysis.

- the space curvature parameter γ:

$$\gamma - 1 = (-0.1 \pm 2.0\,[0.4]) \times 10^{-3}\,.$$

Other measurements of $\gamma-1$ gave: $(0. \pm 2.0)\times10^{-3}$ (Reasenberg et al. 1979), $(-0.7 \pm 1.7) \times 10^{-3}$ (Hellings 1983) and $(0.2 \pm 2.0) \times 10^{-3}$ (Robertson et al. 1991).

• the non-linearity parameter β:

$$\beta - 1 = (-0.2 \pm 1.5\,[0.2]) \times 10^{-3}\,.$$

Hellings (1983) obtained $\beta - 1 = (-2.9 \pm 3.1) \times 10^{-3}$. β is strongly correlated with γ and slightly with the orbital parameters of the Moon, the lunar tidal acceleration and $\dot{G}/G$.

The error of both metric parameter γ and β increases when estimating the orbital elements of the Earth simultaneously and would be obtained to 4×10^{-3} for γ and 3×10^{-3} for β.

• the Nordtvedt parameter η:

$$\eta = 0.0001 \pm 0.0015\,[0.0006]\,,$$

corresponding to a variation of the Earth-Moon distance of one millimeter with an error of 15 millimeter. Other measurements of η gave $0.001 \pm 0.015\,[0.004]$ by Shapiro et al. (1976) where the value in square brackets indicates again the formal $1\,\sigma$ error; Dickey et al. (1989) obtained a value of 0.003 with an estimated total error of 0.004. The Nordtvedt parameter is strongly correlated with the mass of the Earth-Moon system (about 70 %) and the X- and Y-components of the lunar orbit.

• the scale factor for an additional geodetic precession Ω_{GP} of the lunar orbit ($h = 0$ in Einstein's theory):

$$h = 0.002 \pm 0.010\,[0.002]\,.$$

This value for h means that the rate of geodetic precession predicted in Einstein's theory of gravity (≈ 2 as/cy) is confirmed with an error of 0.002 as/cy. In comparison Shapiro et al. (1988) obtained $h = 0.00 \pm 0.02$.

Besides the correlation given in Table 1 Ω_{GP} is weakly correlated with the station coordinates.

• the time variation of the gravitational constant:

$$\frac{\dot{G}}{G} = (0.003 \pm 1.04\,[0.004]) \times 10^{-11}\ \mathrm{yr}^{-1}\,.$$

Here the realistic error is essentially determined by that of the lunar tidal acceleration $\dot{n}_{\mathrm{M}}$. The error estimation for $\dot{n}_{\mathrm{M}}$ is a subtle issue that will be discussed elsewhere; here we have used a rather conservative (large) value for the error of $\dot{n}_{\mathrm{M}}$ ($0.9\,\mathrm{as/cy}^2$) and assumed a correlation of about 100% between both parameters to derive the above realistic error for $\dot{G}/G$.

Other determinations for $\dot{G}/G$ based upon radar measurements to VIKING gave an upper limit of $3 \times 10^{-11}\ \mathrm{yr}^{-1}$ by Reasenberg (1983) resp.

$(0.2 \pm 0.4) \times 10^{-11}$ yr^{-1} by Hellings et al. (1983); a more recent determination using data from the binary pulsar PSR 1913+16 by Damour et al. (1988) gave $(1.0 \pm 2.3) \times 10^{-11}$ yr^{-1}.

9 Conclusion

We have shown some theoretical and practical aspects of the LLR data analysis and gave some insights into the estimation of realistic errors.

In conclusion, LLR data analysis provides an excellent method for determing parameters of the Earth-Moon system, including Earth's orientation parameters. After having obtained a good set of starting values, parameters related with metric theories of gravity can be determined with great accuracy. Our parameter fit to about 7100 LLR observations does not indicate any violation of Einstein's theory of gravity in the first post-Newtonian approximation, rather confirms the general theory of relativity impressively.

At the moment we investigate the possibility to determine further relativistic parameters from the LLR data; effects of interest are e.g. the Lorentz contraction or the proportionality of inertial and gravitational mass.

Support from the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

References

Bauer, R. (1989): *Bestimmung von Parametern des Erde-Mond-Systems – Ein Beitrag zur Modellerweiterung und Bewertung, Ergebnisse* – (Veröffentlichung der Deutschen Geodätischen Kommission, Reihe C, Nr. 353, München)

Bender, P.L. (1990): private communication

Brown, T.M., Christensen-Dalsgaard, J., Dziembowski, W.A., Goode, P., Gough, D.O., Morrow, C.A. (1989): "Inferring the Sun's Internal Angular Velocity from Observed p-Mode Frequency Splittings", in Astrophys. Journal, Vol. 343, pp. 526–546

Brumberg, V.A., Kopejkin, S.M. (1988): "Relativistic Theory of Celestial Reference Frames", in *Reference Systems*, ed. by J. Kovalevsky, I.I. Mueller, B. Kolaczek (Reidel, Dordrecht), pp. 115–140

Brumberg, V.A., Kopejkin, S.M. (1989): "Relativistic Reference Systems and Motion of Test Bodies in the Vincinity of the Earth", in Nuovo Cim., Vol. 103 B, pp. 63–98

Campbell, L., McDow, J.C., Moffat, J.W., Vincent, D. (1983): In Nature, Vol. 305, pp. 508

Damour, T., Gibbons, G.W., Taylor, J.H. (1988): "Limits on the Variability of G Using Binary-Pulsar Data", in Phys. Rev. Lett., Vol. 61, No. 10, pp. 1151–1154

Damour, T., Soffel, M., Xu, C. (1991): "General-relativistic celestial mechanics. I. Method and definition of reference systems", in Phys. Rev. D, Vol. 43, No. 10, pp. 3273–3307

Dickey, J.O., X. X., Williams, J.G. (1989): "Investigating Relativity Using Lunar Laser Ranging: Geodetic Precession and the Nordtvedt Effect", JPL Geodesy and Geophysics Preprint, No. 173

Gleixner H. (1986): *Ein Beitrag zur Ephemeridenrechnung und Parameterschätzung im Erde-Mond-System* (Veröffentlichung der Deutschen Geodätischen Kommission, Reihe C, Nr. 319, München)

Gough D.O. (1982): In Nature, Vol. 298, pp. 334

Hellings, R.W. (1983): In Conference Papers of the 10th International Conference on General Relativity and Gravitation

Hellings, R.W., Adams, P.J., Anderson, J.D., Keesey, M.S., Lau, E.L., Standish, E.M., Canuto, V.M., Goldman, I. (1983): In Phys. Rev. Lett., Vol. 51, pp. 1609

Hill, H.A., Bos, R.J., Goode, P.R. (1982): In Phys. Rev. Lett., Vol. 49, pp. 1794

Hirayama, Th., Kinoshita, H., Fujimoto, M.-K., Fukushima, T. (1987): "Analytical Expression of TDB–TDT", in Proceedings of the IUGG General Assembly (Vancouver, Canada)

Kopejkin, S.M. (1988): "Celestial Coordinate Reference Systems in Curved Space-Time", in Celestial Mechanics, Vol. 44, pp. 87–115

Lieske, J.H., Lederle, T., Fricke, W., Morando, W. (1977):"Expressions for the Precession Quantities Based upon the IAU (1976) System of Astronomical Constants", in Astronomy and Astrophysics, Vol. 58, pp. 1

Marini, J.W., Murray, C.W. (1973): "Correction of Laser Range Tracking Data for Atmospheric Refraction at Elevations above 10 Degrees", NASA, Techn. Report, X-591-73-531

Minster, J.B., Jordan, T.H. (1978): "Present-Day Plate Motions", in Journal of Geophys. Res., Vol. 83, pp. 5531–5354

Müller, J., (1991): *Analyse von Lasermessungen zum Mond im Rahmen einer post-Newton'schen Theorie* (Veröffentlichung der Deutschen Geodätischen Kommission, Reihe C, Nr. 383, München)

Reasenberg, R.D. (1983): In Philos. Trans. Roy. Soc. London, Vol. 310 A, pp. 227

Reasenberg, R.D., Shapiro, I.I., MacNeil, P.E., Goldstein, R.B., Breidenthal, J.C., Brenkle, J.P., Cain, D.L., Kaufman, T.M., Zygielbaum, A.I. (1979): In Astrophys. Journal, Vol. 234, L219

Robertson, D.S., Carter, W.E., Dillinger, W.H. (1991): "New Measurement of Solar Gravitational Deflection of Radio Signals Using VLBI", in Nature, Vol. 349, pp. 768–770

Schastok, J., Gleixner, H., Soffel, M.H., Ruder, H., Schneider, M. (1989): In Comp.-Phys. Commun., Vol. 54, pp. 167

Schneider, M. (ed.) (1989): *Satellitengeodäsie: Ergebnisse aus dem gleichnamigen Sonderforschungsbereich der Technischen Universität München* (VCH Verlagsgesellschaft, Weinheim)

Shapiro, I.I., Counselman, C.C., III., King, R.W. (1976): In Phys. Rev. Lett., Vol. 36, pp. 555

Shapiro, I.I., Reasenberg, R.D., Chandler, J.F., Babcock, R.W., (1988): "Measurement of the de Sitter Precession of the Moon: A Relativistic Three-Body Effect", in Phys. Rev. Lett., Vol. 61, pp. 2643–2646

Soffel, M. (1989): *Relativity in Astrometry, Celestial Mechanics and Geodesy* (Springer, Berlin)

Very-Long-Baseline Interferometry in Astro-, Geo-, and Gravitational Physics

Eugen Preuss [1], James Campbell [2]

[1]Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69,
D-5300 Bonn 1, Germany
[2]Geodätisches Institut der Universität Bonn, Nussallee 17,
D-5300 Bonn 1, Germany

Abstract: Very-Long-Baseline Interferometry (VLBI) is routinely used at dm/cm-wavelengths on baselines ranging up to $\approx$ 10000 km. Global VLBI networks of large radio telescopes equipped with low noise receivers, atomic clocks driving independent local oscillators, and broad-band recorders (data rates $\gtrsim$ 100 Mbit/s) allow the measurement of coherence functions of radio wave fields with high sensitivity (detection limits $\gtrsim 10^{-29}$ W/(Hz m^2)) and differenced group delays of radio signals with high accuracy (errors $\lesssim 5 \times 10^{-11}$ s). We discuss basic principles, current limits, and important developments of the method in view of its use in astro-, geo-, and gravitational physics.

1 Introduction and Summary

In this lecture we will discuss some important aspects of VLBI (Very-Long-Baseline Interferometry) and its use in astro-, geo-, and gravitational physics. This introductory chapter summarizes the characteristic features of VLBI. We will then discuss some basic concepts of radio interferometry (Chap. 2), the VLBI signal path (Chap. 3), high-precision interferometry based on the group delay observable (Chap. 4), and finally important areas of scientific application, major current activities and developments (Chap. 5).

For more thorough study we refer to the following

- monographs on radio interferometry: Thompson, Moran, Swenson (1986); (ed.) Meeks (1976); Wohlleben, Mattes, Krichbaum (1991), on relativity in astrometry and geodesy: Soffel (1989);
- lecture notes on synthesis imaging (eds.): Perley, Schwab, Bridle (1989), on VLBI: Felli and Spencer (1989);
- conference proceedings on radio interferometry (eds.): Cornwell and Perley (1991), on the impact of VLBI on astro- and geophysics: Reid and Moran

(1988), on parsec-scale radio jets: Zensus and Pearson (1989), on geodetic VLBI: Carter (1991), on space- and mm-VLBI: Hirabayashi, Inoue, Kobayashi (1992).

1.1 VLBI: a Summary

All types of radiointerferometer currently used in radioastronomy and geodesy work according to the same principles. They measure primarily the complex coherence function $\Gamma_{12} \propto \langle v_1 v_2^* \rangle$ between two space-time points 1,2 of a radio-wavefield by cross correlating the fluctuating antenna signals v_1, v_2 sampled prior to quadratic detection; i.e. the quantities being correlated are voltages and ***not*** intensities. This is an essential prerequisite for achieving the required sensitivity, measuring accuracy, and imaging capability. For an incoherent and stationary radio source $\Gamma_{12} = \Gamma(\boldsymbol{d}_\perp/\lambda, \tau)$ depends only on the projected antenna spacing $\boldsymbol{d}_\perp/\lambda$ (perpendicular to the line of sight) measured in wavelengths and on the time difference τ of the two antenna signals with respect to an incoming plane wave front (parallel to $\boldsymbol{d}_\perp$). The potential of interferometry as a high precision/ high angular resolution method is due to the fact that for non-monochromatic, incoherent sources (bandwidth B, diameter φ) Γ is zero everywhere (due to destructive interference) except for a small range of values of τ and φ fulfilling the temporal and spatial 'coherence conditions' respectively: $\tau < 1/\mathrm{B}$ and $\varphi < \lambda/\mathrm{d}_\perp$. For a continuum source Γ has a sharp maximum around $\tau = 0$ depending on B. This allows an accurate measurement of the VLBI delay τ_g, i.e. the difference in arrival time of a radio wave group at two antennae. The quantity τ_g is the most important VLBI observable for measuring angular positions of radio sources, vector spacings of antennae, Earth rotation parameters, and any other quantity affecting the path of the radio signal directly or indirectly.

The ratio $\lambda/\mathrm{d}_\perp$ determines the scale of angular resolution, and $\Gamma(\boldsymbol{d}_\perp/\lambda, \tau \to 0)$ is proportional to the Fourier transform of the intensity distribution (image) of the radio source. The image can be restored if Γ is measured at a sufficient number of points in the $\boldsymbol{d}_\perp/\lambda$-plane or 'aperture plane'. Adequate sampling of the aperture plane is achieved with an array of, say, 10 radio telescopes, by taking advantage of the Earth rotation which changes the length and aspect angle of the projected baselines as the observing goes on.

From a technical point of view, one basically distinguishes two types of instrument: a) the short-baseline ($\lesssim$ 150 km) 'connected-element interferometers', and b) the VLB interferometers (baselines up to $\approx$ 10000 km) which have no real-time coherent links between the receiving elements. In VLBI the signals received at each interferometer element are recorded digitally during the observing run and correlated at a later time at a central processing facility. Specific technical requirements for this kind of operation are fast signal recorders, compact storage media, multi-station correlators, and high precision frequency standards (atomic clocks) which control the independent local oscillators and the signal recording at each station. For sensitivity reasons low noise microwave receivers and large antennae as interferometer elements are required.

The first succesful VLBI experiments were carried out in Canada at 75 cm and in the USA at 50 cm wavelength in 1967. Since then VLBI has developed rapidly. The sceptics who predicted that VLBI would never be able to produce proper images and measure intercontinental distances with cm-accuracy were refuted by the progress of the last two decades. Even now the potential of VLBI is far from being exhausted. Subcm-accuracy, for example, is now the declared goal of geophysical VLBI.

The scientific motivation and justification for the development of VLBI have come from astrophysics (quasars, active galactic nuclei (AGNs), interstellar masers (H_2O, OH), stars), geophysics (variation of Earth rotation parameters, crustal motion, tides of the solid Earth), and gravitational physics (light deflection).

VLBI networks of continental and global diameters, routinely operated at dm/cm wavelengths, are effectively the largest telescopes which have ever looked into the depths of the universe. Images can be obtained with an angular resolution of $\gtrsim 0\overset{''}{.}0002$ (see Fig. 13), still unrivalled in any other branch of astronomy. VLBI observations have produced many important discoveries including the common occurrence of collimated outflow (beams, jets) from AGNs on the $\lesssim$pc scale often extending up to $\lesssim$1 Mpc from the origin, and the structural variability of compact radio sources many of which show component separating with apparent speeds $\gtrsim$c (superluminal motion).

VLBI is the only method for measuring intercontinental distances ***directly***, i.e. independently of a model for the Earth's gravitational field, with cm accuracy. VLBI has made possible for the first time the direct measurement of the relative motion of tectonic plates predicted by Wegener's continental drift hypothesis (Göttingen 1912).

VLBI has become the method with the highest precision and the highest time resolution for monitoring the Earth's rotation parameters (polar motion, UT1 variations). The current IERS (International Earth Rotation Service, Paris, formerly BIH) Earth rotation bulletin is based primarily on VLBI measurements.

The accuracy level of geophysical VLBI observing campaigns routinely requires taking into account the effect of gravitational light deflection by the sun for ***all*** observations. This in turn has allowed the determination of the parameter γ_{PPN} of the Parametrized Post-Newtonian formulation of gravitation theories, as 1 ± 0.002.

VLBI is international for obvious reasons and interdisciplinary by its very nature.

Some important ongoing developments are:

- the build up of dedicated VLBI arrays of $\geqq 10$ stations with multi-frequency observing capability for astronomical radio sources (most notably the Very-Long-Baseline Array (VLBA), New Mexico, USA),
- the organisation of long term programs for monitoring geodynamical effects (IRIS project and NASA's Crustal Dynamics Program) with the declared goal of advancing into the sub-cm accuracy regime,
- the development of mm-VLBI, presently at 43, 100, and 230 GHz,
- the planning and preparation for space VLBI: projects Radioastron (Interkosmos, Russia, CIS) and VSOP (ISAS, Japan). The idea is to operate ground-

based VLBI networks in conjunction with space borne antennae. Such space systems, by utilizing both Earth rotation and the orbital motion of the satellite antennae, will effectively create telescopes of diameter $\gtrsim 25000\,\mathrm{km}$.

- the development of advanced VLBI systems (recording, playback, and correlation equipment) based on high-density recording techniques. Recording rates of up to 2 Gbit/s have already been demonstrated to be possible.

The main reasons which have enabled radio interferometry to advance so swiftly toward higher and higher resolution are

a) the favorable properties of the atmosphere at dm/cm wavelengths: radio imaging is diffraction limited even at resolutions $\lesssim 0\farcs0001$ at these wavelengths.

b) the possibility of coherent reception in the radio regime. The frequencies of the signals received are 'downconverted' by mixing with a oscillator signal. This feature, referred to as 'heterodyne frequency conversion', is crucial in making radio interferometry such a powerful method. It enables the major part of the signal processing to be performed at intermediate frequencies that are most appropriate for amplification, transmission, filtering, recording, and other processes.
For a summary of technical characteristics of VLBI, see also Tab. 1.

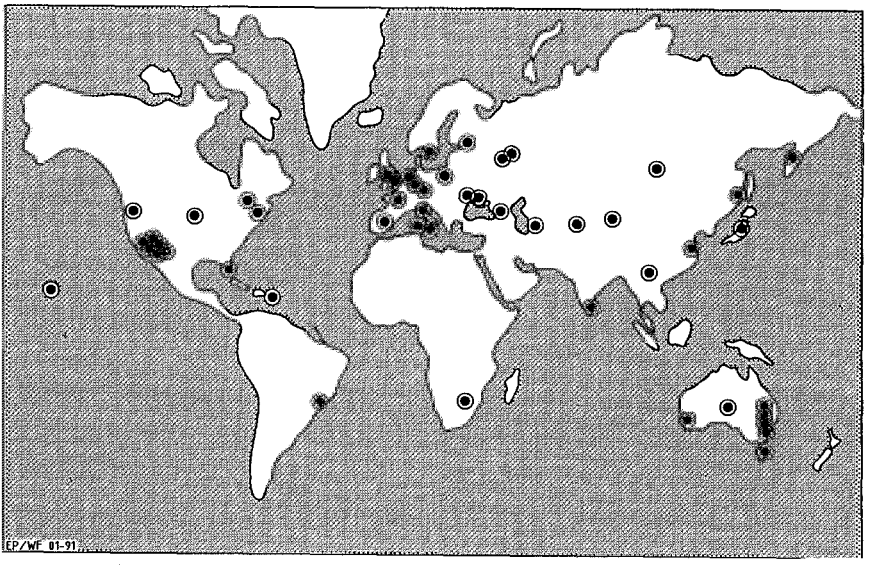

Fig. 1. About 55 major VLBI stations expected to be operational by the mid 1990's. The radio telescopes are located in Australia (7), Brazil, Canada, China (3), CIS (former USSR) (11), England (2), France, Germany (2), Italy (4), India, Japan (2), The Netherlands, Poland, South Africa, Spain (2), Sweden, USA (13), and (not shown here) the German station O'Higgins in Antarctica

2 Basic Concepts: Temporal and Spatial Coherence

In order to understand the high precision, the high angular resolution, and the imaging capability of VLBI measurements it is necessary to consider the interference process taking place in an interferometer. We make the following assumptions:

a) the radio source of diameter φ is spatially incoherent,
b) the wave field is stationary,
c) the signal is quasi-monochromatic, i.e. spectral bandwidth B $\ll$ observing frequency ν,
d) the wave field can be described by plane waves.

These assumptions are necessary to make the problem tractable; they hold in almost all practical cases. For the sake of simplicity we assume here:

e) polarisation match, i.e. we neglect the vector character of the wave field,
f) 'white noise', i.e. negligible ν-dependence of the spectral flux density within B, which is normally the case for continuum sources in contrast to line sources,
g) absence of the atmosphere and a noiseless and perfectly coherent interferometer. Both assumptions are never fulfilled in reality but are helpful for understanding the principle.

As all interferometric arrays are, from an operational point of view, an ensemble of two-element-interferometers, it is sufficient in the following to consider a 'twin interferometer'. This measures the mutual coherence function Γ_{12} between two space-time points 1,2. Assumptions (a) and (b) imply that Γ depends only on the 'spatial frequency' $\boldsymbol{d}_\perp/\lambda$ (see Fig. 3) and delay τ of a (wave group) signal with respect to the reference- or aperture plane (spanned by $\boldsymbol{d}_\perp/\lambda$) :

$$\Gamma(\boldsymbol{d}_\perp/\lambda, \tau) = \Gamma_{12} \equiv \langle V_1 V_2^* \rangle_{t(acc)} \tag{1}$$

where V_1, V_2 are complex amplitudes (Mandel and Wolf 1965) associated with the fluctuating field strengths of the radio wave field and which are measured by the induced antenna voltages; the brackets denote a time average over an accumulation time t(acc) $\lesssim$ 2 s.

Assumption (c) factorizes the temporal and spatial interference effects so that we can consider them as decoupled in a first approximation.

It is now important to note that Γ is only measurably non-zero under fairly restrictive 'coherence conditions'. This is because for given instrumental parameters B and $\boldsymbol{d}_\perp/\lambda$ the superposition of spectral components with $\tau > B^{-1}$ and/or the superposition of directional components from a cone angle $\varphi > \lambda/d_\perp$ leads to destructive interference. The temporal (longitudinal) and spatial (lateral) coherence conditions are therefore

$$\tau < 1/B \quad \text{and} \quad \varphi < \lambda/d_\perp . \tag{2}$$

$1/B$ determines the order of the spectral coherence time and $\lambda/d_\perp$ the scale of angular resolution. A source with $\varphi \ll \lambda/d_\perp$ is called a 'point source'. $\varphi \gtrsim \lambda/d_\perp$(max) means the source is 'resolved'.

The temporal coherence condition means in practice that only signals from incoming plane wave fronts should be crosscorrelated. This again means that the data streams coming from two antennae must be shifted before the correlation by the difference in arrival time τ_g of a plane wave front at two interferometer elements. The 'VLBI delay' τ_g is dominated by the light travel time corresponding to the longitudinal baseline component $\boldsymbol{d} \cdot \boldsymbol{s}$ (Fig. 3). τ_g has therefore to be known a priori for each moment of the observation with an accuracy $\leqq 1/B$, so that it can be continuously compensated before the correlation ('delay tracking') and the correlation is always performed near the interference maximum ($\tau = 0$). The sharpness of this maximum on the other hand enables τ_g to be determined with higher accuracy, a posteriori. This is the basis for high precision interferometry and its use in astro-, geo-, and gravitational physics.

To see this more quantitatively consider the normalized 'delay characteristic' $\tilde{F}(\tau)$ which describes the interferometer response as a function of τ for a sufficiently strong and compact continuum source. The superposition of harmonic components in the downconverted observing band (= 'base band' 0 MHz $\rightarrow$ B), lead after crosscorrelation with delay τ, to the following expression for $\tilde{F}(\tau)$:

$$\tilde{F}(\tau) = e^{-i\pi B\tau} \cdot \int_{-B/2}^{B/2} F(\nu')\, e^{-i2\pi\nu'\tau}\, d\nu' \;/\; \int_{-B/2}^{B/2} F(\nu')\, d\nu' \tag{3}$$

$$= e^{-i\pi B\tau} \cdot \frac{\sin(\pi B\tau)}{\pi B\tau} \quad \text{for rectangular bandpasses} \tag{4}$$

where $F(\nu')$ is the product of the bandpasses (frequency filter characteristic), here assumed to be identical; the origin of ν' is taken to be at the midpoint of the base band, i.e. the center frequency B/2. The half width of the central peak of the sinc function around $\tau = 0$ is proportional to B^{-1} and so is the accuracy with which the position of the peak can be determined. Here continuous coverage of the frequency band was assumed. This condition can be relaxed, that is, the bandwidth may be spread without the need to continuously cover the range between the minimum and maximum frequency. In this way the measuring accuracy of τ_g is further increased (Rogers 1970):

$$\delta_{\mathrm{rms}}\tau_g \approx (2\pi\,(\nu_{\max} - \nu_{\min}) \cdot \mathrm{SNR})^{-1} \quad , \tag{5}$$

where δ_{rms} means 'rms error', SNR Signal-to-Noise ratio, and $B_s = (\nu_{\max} - \nu_{\min})$ is the 'spanned bandwidth'.

τ_g is in practice the most important VLBI observable for astrometric and geodetic applications, i.e. for measuring source positions or directional vectors ($\boldsymbol{s}; |\boldsymbol{s}| = 1$), baseline vectors ($\boldsymbol{d}$), variations of the Earth rotation vector ($\boldsymbol{\Omega}_\oplus$), and all other quantities which influence τ_g. The vectors $\boldsymbol{s}$, $\boldsymbol{d}$, and $\boldsymbol{\Omega}_\oplus$ are related by (6) and (7):

$$\tau_g = -\frac{\boldsymbol{d}}{c} \cdot \boldsymbol{s} + \text{smaller terms} \tag{6}$$

$$\nu_f = \nu \cdot \dot{\tau}_g = \frac{\boldsymbol{d}}{\lambda} \cdot (\boldsymbol{\Omega}_\oplus \times \boldsymbol{s}) + \text{smaller terms} \tag{7a}$$

$$\approx 7.3\,\mathrm{kHz} \cdot (\mathrm{d}_\perp\,(\mathrm{E\text{–}W})/1000\,\mathrm{km})/(\lambda/\mathrm{cm}) \cdot \cos\delta \quad . \tag{7b}$$

The 'fringe frequency' ν_f or the 'delay rate' $\dot{\tau}_g$ describe the relative phase drift $2\pi\nu_f t$ between signals V_1 and V_2 induced by the differential Doppler effect due to the Earth rotation. "E–W" indicates the East-West component of the projected baseline; δ is the source declination.

Consider spatial coherence. What is the relation between Γ and the brightness distribution or 'image' $I(\sigma)$ of an extended radio source with $\sigma \equiv s - s_0$ (Fig. 3) denoting the relative position of a picture element? This can easily be seen thanks to assumptions (a) and (d).

In the case of an extended radio source the signal received at each interferometer element is a superposition $\int d\sigma\, V(\sigma)$ of contributions from all source elements, and

$$\Gamma_{12} = \langle V_1 V_2^* \rangle = \langle \iint V_1(\sigma) V_2^*(\sigma')\, d\sigma\, d\sigma' \rangle \qquad (8)$$

Incoherence of the radio source means that the time average of all cross products of contributions from different source elements disappear so that Γ_{12} is reduced to

$$\Gamma_{12} = \int \langle V_1(\sigma) V_2^*(\sigma) \rangle\, d\sigma \quad . \qquad (9)$$

The signals $V_1(\sigma)$ and $V_2(\sigma)$ of a plane harmonic wave (assumption d) passing through the aperture plane ($\tau = 0$) at a small angle σ to the reference direction are identical except for a phase difference $2\pi(d_\perp/\lambda)\sigma$ (Fig. 3):

$$V_2(\sigma) = V_1(\sigma) \cdot e^{i2\pi (d_\perp/\lambda)\cdot\sigma} \quad . \qquad (10)$$

Inserting (10) into (9) and taking into account that $I(\sigma) = \langle V_1(\sigma) V_1^*(\sigma) \rangle$, yields immediately the Fourier relation between I and Γ, the so-called van Cittert-Zernike theorem

$$\Gamma(d_\perp/\lambda; \tau = 0) = \int I(\sigma) \cdot e^{-i2\pi (d_\perp/\lambda)\cdot\sigma}\, d\sigma \quad . \qquad (11)$$

Spatial frequency and relative angular position are usually expressed by their rectangular E/W and N/S components: $(u, v) = (d_\perp/\lambda)$ and $(\xi, \eta) = \sigma$. By setting $Ae^{i\Phi} \equiv \Gamma(u, v, \tau = 0)$ [A,Φ real] we can rewrite (11):

$$A(u,v) \cdot e^{i\Phi(u,v)} = \iint I(\xi,\eta) \cdot e^{-i2\pi (u\xi + v\eta)} d\xi d\eta \quad . \qquad (12)$$

A is called 'fringe amplitude' or 'correlated flux density' S(corr) typically calibrated in units of Jy (Jansky) $= 10^{-26} Wm^{-2} Hz^{-1}$. The normalized quantity S(corr)/S(total) is called 'visibility' (meaning 'fringe visibility').

Relation (12) is the basis for interferometric imaging. Proper image restoration requires A and Φ to be measured at a sufficient number of points (u,v) in the Fourier- or 'aperture' plane. This is achieved with an array of $\gtrsim 10$ antennae observing the radio source for several ($\lesssim 12$) hours. During this time the Earth rotation generates for each binary combination in the array a sequence of different points in the (u,v) plane ("Earth rotation synthesis"). In astronomical interferometry the actual degree of (u,v) plane coverage varies considerably. The extreme cases are, at one end minimal knowledge of $\Gamma(u, v)$, namely at one or a few (u,v)

points only, and at the other end maximal knowledge, i.e. I is completely sampled over an aperture of radius $d_\perp(\max)/\lambda$ on a grid of cell size $\lambda/(2d_\perp(\max))$. In the latter case, which is only achievable in local interferometry, perfect imaging is possible by straightforward Fourier inversion. In intermediate (typical VLBI) cases image restoration requires iterative methods; and the achievable dynamic range depends strongly on the actual coverage of the (u,v) plane and the amount of retrievable phase (Φ) information (Sect. 3.5).

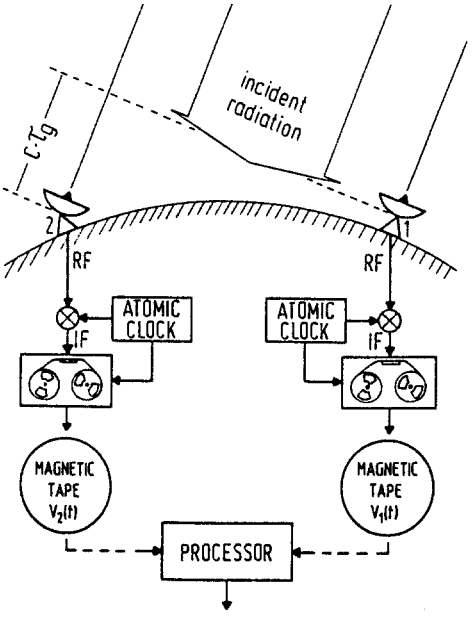

Fig. 2. Schematic diagram of the experimental setup of VLBI

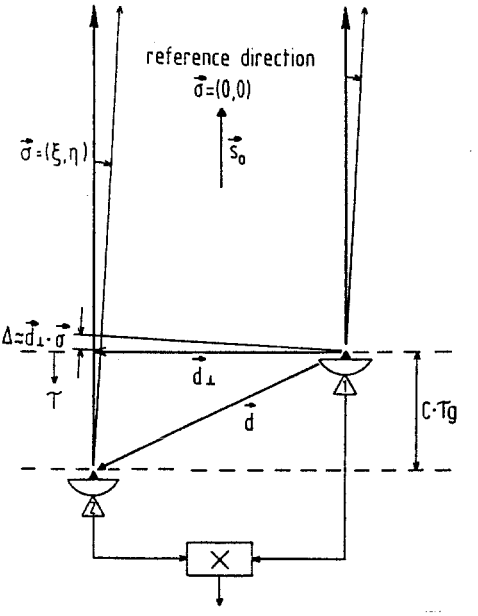

Fig. 3. Basic geometry of a two-element interferometer. τ_g is here the geometric delay, the dominant part of the group delay

3 Along the VLBI Signal Path

3.1 VLBI Specifics

The defining property of VLBI is the lack of coherent links between the interferometer elements and the correlator, i.e. the 'focus' of a VLBI network or 'array' (Fig. 2). The reason for this is the prohibitive cost of such links over long distances. As a consequence, the measuring process is decomposed into four main steps to be briefly described in the following:

1) signal reception and digital recording at N stations,
2) the correlation of the signals for all N(N-1)/2 baselines,

3) the signal analysis yielding the observables proper and calibration of the 'raw' fringe amplitude,
4a) in astronomy: the restoration of the radio source image and/or
4b) in geodetic/astrometric VLBI: the determination of source positions, baseline vectors, Earth rotation parameters and other quantities which can measurably affect the differential light paths.

The main constituents of the interferometric setup [important parameters or effects in brackets] are (Fig. 2, Fig. 3):

- the radio source ['image' $I(\xi, \eta)$, angular position];
- the Earth's atmosphere [fluctuation of path length and attenuation];
- the interferometer elements comprising: antennae [diameter D, baseline vectors **d**], receiving systems [observing wavelength λ, noise temperature T_{sys}], independent frequency standards [stability $\delta\nu/\nu$], signal recorders [sampling rate or bandwidth B];
- the central VLBI processor incorporating control computer, playback system, and correlator [number of stations which can be correlated simultaneously at a given bandwidth];
- the Earth [rotation vector $\Omega_\oplus$, fringe frequency ν_f]

The 'VLBI system' (recording and correlation equipment) most widely used at the moment is the Mark III-System developed at Haystack Observatory, Ma., USA (Rogers et al. 1983). It has a maximal contiguous bandwidth of 112 MHz. It will be succeeded by the next generation Mark IV system in the near future.

The following limitations are typical for VLBI:
a) The interferometer coherence time $t_{coh} \gtrsim 10$ min is much shorter than a full observing run (~ 12 h) due to the limited – albeit high – stability of the independent frequency standards ($\lesssim 10^{-14}$) and by the atmosphere. As t_{coh} is the upper limit to the coherent integration time t_{int}, this limits the sensitivity of the interferometer.
b) It also leads to the loss of phase of the complex coherence function. Fortunately this visibility phase can be recovered later to some extent depending on the amount of measured data.
c) The nonoptimal geographical distribution and/or an insufficient number of VLBI stations result in an incomplete coverage of the Fourier or (u,v) plane. Both limitations (b) and (c) pose problems to image restoration.

The point source detection limit of a two-element interferometer is given by $\lesssim 7$ times the rms noise, $\sigma(S_{corr})$, of the correlated flux density

$$1\sigma(S_{corr}) \approx \frac{7500 \cdot \sqrt{T_{sys,1}[K] \cdot T_{sys,2}[K]}}{\sqrt{B[Hz] \cdot t_{int}[s] \cdot D_1[m] \cdot D_2[m] \cdot \sqrt{\eta_1 \cdot \eta_2}}} \text{ Jy} \quad , \tag{13}$$

with t_{int} coherent integration time and η aperture efficiency, defined as ratio of the effective to the geometric antenna collecting area, typically about 0.5.

3.2 Signal Reception and Recording

Figure 4 shows the main processes taking place at the site of each interferometer element: these are

- signal reception with primary amplification in a low noise amplifier and frequency down-conversion from the radio band in the GHz regime to the 'video' regime $0\,\mathrm{MHz} \leqq \nu \leqq \mathrm{B}$. Low-noise receivers typically used are maser amplifiers, FETs (field- effect transistors), and increasingly HEMTs (high electron mobility transistors)
- sampling with 1 bit per sample (Mk III system). The sampling rate has to be twice the bandwidth;
- recording of the signal on magnetic tape for later processing at the correlator. The data on tape are 1-bit samples of the fluctuating antenna voltage (not intensity!); they are typically dominated by uncorrelated noise from the receiver, the atmosphere, extended radio emission and other sources so that the primary signal , i.e. the uncalibrated correlation coefficient is typically only of the order of 10^{-3} to 10^{-5}.

The whole process is under control of a highly stable atomic frequency standard ($\delta\nu/\nu \lesssim 10^{-14}$ over 1 to 10^5 s). It controls the local oscillator and functions as clock, implicitly by controlling the sampling, and explicitly by marking the data with UTC[1] 200 times per second. The most demanding task of the atomic clock is not the time keeping (tie to UTC) but rather to guarantee the extreme phase stability needed for coherent integration over $\lesssim$ several hundred seconds and the instrumental group delay stability over $\lesssim$24 h. Harmonic signals received at two stations have to keep 'in step' for at least about one minute for constructive interference ("fringes") to be measurable, and preferably for much longer. Hydrogen masers are at the moment the atomic clocks best suited for VLBI owing to their high stability on both short and long time scales from one to several 10^4 seconds. The time synchronisation of the various station clocks has to be on the order of microseconds which can easily be achieved.

3.3 Cross-Correlation

Fig. 4 shows the main functions of the VLBI processor. The 'a priori information' comprises all data which determine the space/time structure of the interferometric setup at each moment of the observation such as source positions, station coordinates, UT1[2]–UTC, clock and local oscillator offsets etc.. These data enable the control computer to calculate during the correlation run (which in a way is a reproduction of the observing run) the running values of the expected delay τ_g^{exp} and expected fringe frequency ν_f^{exp} for all baselines. As we have seen before there is no use in just correlating signals which have arrived at the same UTC at different stations. To meet the temporal coherence condition the signals from two stations

[1] UTC (Universal Time Coordinated) = TAI (Temps Atomique International) + leap second

[2] UT1 = observed mean solar time corrected for polar motion

have to be synchronised w.r.t. an incoming plane wave front ($\| \boldsymbol{d}_\perp ; \tau = 0$). This is done by the 'delay tracker' with submicrosecond accuracy in order to be always close to the interference maximum (width $\approx 1/\mathrm{B}$ on the τ-axis).

The 'fringe rotator' is supposed to remove the differential Doppler effect due to the Earth rotation. Due to the uncertainty in the a priori knowledge of the interferometer geometry, this means in practice that the 'natural fringe frequency' ν_f [10 kHz scale] will not be completely removed but slowed down by several orders of magnitude to the 'residual fringe frequency' $\nu_\mathrm{res} = \nu_\mathrm{f} - \nu_\mathrm{f}^\mathrm{exp}$ [1 to 10 mHz scale]. This is achieved by mixing signal v_1 with a harmonic signal of the expected fringe frequency $\nu_\mathrm{f}^\mathrm{exp}$. The mixing is done twice, using two signals in 'phase quadrature' (90° phase difference) so that at this point a complex number is generated in the digital signal stream used to compute the complex correlation coefficient $\hat{\rho}_{12}$,

$$\hat{\rho}_{12}(\mathrm{u},\mathrm{v},\tau_\mu[\mu\mathrm{s}]\,;\,\mathrm{t[s]}) \;=\; \langle \mathrm{v}_1^\mathrm{sync}\mathrm{v}_2^* \rangle_{\approx 1\mathrm{s}} / \sqrt{\langle \mathrm{v}_1^\mathrm{sync}\mathrm{v}_1^{\mathrm{sync}*} \rangle \cdot \langle \mathrm{v}_2\mathrm{v}_2^* \rangle} \qquad (14a)$$

$$\text{with} \quad \mathrm{v}_1^\mathrm{sync}(\mathrm{t}') = \mathrm{v}_1(\mathrm{t}'[\mu\mathrm{s}] + \tau_\mathrm{g}^\mathrm{exp}[\mathrm{ms}] + \tau_\mu[\mu\mathrm{s}]) \cdot \mathrm{e}^{-\mathrm{i}2\pi\nu_\mathrm{f}^\mathrm{exp}\mathrm{t}'} \,, \qquad (15b)$$

where $\mathrm{v}_1, \mathrm{v}_2$ are antenna voltages, and $\tau_\mu = 0 \pm \mu \cdot 0.25\,\mu\mathrm{s}$ [$\mu = 0, 1, 2, \ldots$]. Note, the correlation is done in a number of 'delay channels' τ_μ spaced by $1/(2\mathrm{B})$ around $\tau_\mu = 0$ to allow for the uncertainty in the a priori knowledge of τ_g. The latter typically causes the peak of ρ_{12} to be located not exactly at $\tau_\mu = 0$ but offset by the 'residual delay' $\tau_\mathrm{res} = \tau_\mathrm{g} - \tau_\mathrm{g}^\mathrm{exp} = \tau - \tau_\mu$ on the μs scale.

Fig. 5 shows, so to speak, part of a running 'interferogram' as it is displayed on the monitor of the VLBI processor during the correlation run. Under assumptions (a) - (f), Sect. 2 (we relax ass. h)), the interferometer response is described approximately by the expression:

$$\hat{\rho}_{12}(\mathrm{u},\mathrm{v},\tau\,;\,\mathrm{t}) \;\propto\; \mathrm{A}(\mathrm{u},\mathrm{v})\,\mathrm{e}^{-\mathrm{i}(\Phi(\mathrm{u},\mathrm{v})+\Psi_{12}(\mathrm{t}))}\,\frac{\sin(\pi\mathrm{B}\tau)}{\pi\mathrm{B}\tau}\,\mathrm{e}^{-\mathrm{i}\pi(2\nu_\mathrm{res}\mathrm{t}+\mathrm{B}\tau)} \,. \qquad (15)$$

This means the correlation signal $\hat{\rho}_{12}$ has a quasi-sinusoidal form as a function of time due to the Earth rotation. The oscillation (frequency ν_res) is modulated by the bandwidth effect (terms containing $\mathrm{B}\tau$, see (3),(4)), by the source structure effect (visibility $\mathrm{A}e^{\mathrm{i}\Phi}$, see (12)), and by the limited phase stability of the instrumentation and/or the atmosphere ($\Psi_{12}(\mathrm{t})$). The signal is detectable only around its maximum at $\tau = \tau_\mu + \tau_\mathrm{res} = 0$ providing the fringe amplitude A is well above the noise (13). Ψ_{12} is constant and therefore without any effect for a perfect interferometer only. In VLBI Ψ_{12} fluctuates irregularly on time scales larger than the interferometer coherence time $\gtrsim$5 min. Note that in general for a resolved source both A and Φ are implicitly time dependent via u(t) and v(t) and vary on the same scale as Φ as observing time goes on. This means that in general the visibility phase Φ is corrupted by the irregular phase fluctuations of the instrumentation and/or the atmosphere and is therefore not directly measurable.

It is instructive to remember some important characteristic time/frequency scales which play a role in the interferometric process. They obey the following inequalities (see also Tab. 1):

$$\nu_{\text{radio}}[\text{GHz}] \gg \text{B}[\text{MHz}] \gg \nu_{\text{f}}[\text{kHz}] \gg \frac{1}{\text{t}_{\text{acc}}[\text{s}]} > \frac{1}{\text{t}_{\text{coh}}[\text{min}]} \gg \frac{1}{\text{t}_{\text{obs}}[12\text{h}]} \quad . \qquad (16)$$

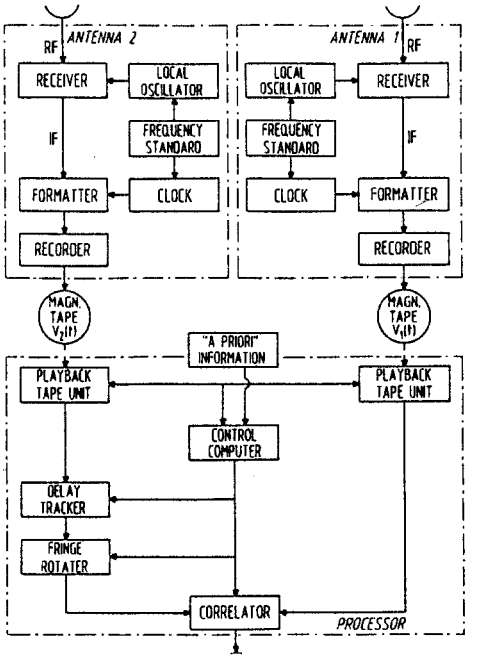

Fig. 4. Functional diagram of the VLBI signal path

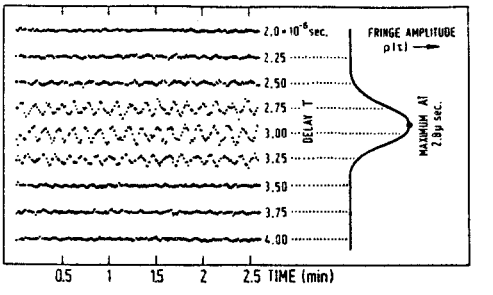

Fig. 5. Online CRT display of a 2 MHz-correlator signal. Shown is the output of 9 delay channels

3.4 Signal Analysis and Primary Observables

The correlation signal is determined by the following measurable quantities or observables

a) the VLBI delay $\tau_{\text{g}} = \tau_{\text{g}}^{\text{exp}} + \tau_{\text{res}}$; $\tau_{\text{res}} = -\tau_{\mu}(\rho = \max)$ is measured by fitting the delay characteristic (4) to the signal and determining the maximum of ρ_{12} on the τ_{μ}-axis.

b) the fringe frequency $\nu_{\text{f}} = \nu_{\text{f}}^{\text{exp}} + \nu_{\text{res}}$; ν_{res} is determined by Fourier analysis of $\hat{\rho}_{12}$ as a function of time over intervals of length $\text{t}_{\text{int}} \lesssim \text{t}_{\text{coh}}$. This, at the same time, provides a "coherent average" of $\hat{\rho}_{12}$.

c) the fringe amplitude $\text{A(u,v)} = \sqrt{\text{Re}^2\hat{\rho}_{12} + \text{Im}^2\hat{\rho}_{12}}$;

d) the fringe or interferometer phase $\Phi + \Psi_{12} = -\text{arctg}(\text{Im}\hat{\rho}_{12}/\text{Re}\hat{\rho}_{12})$.

Fringe amplitude and phase are at this point coherently averaged quantities of the signal maximum.

3.5 VLBI Imaging

Image reconstruction from VLBI data faces the following main problems:

a) incomplete sampling of the Fourier plane ("holes" in the aperture),
b) strongly corrupted phase Φ of the coherence function, and
c) amplitude calibration problems because there are practically no point sources on the milliarcsec scale which would provide a fixed reference amplitude over the whole (u,v) plane.

These drawbacks preclude image restoration by straightforward Fourier inversion of (12). Nevertheless around 1977/78 it was convincingly shown that these problems can be overcome by iterative procedures (Readhead and Wilkinson 1978, Cornwell 1987). Of greatest practical importance are the "self-calibration" (see review by Pearson and Readhead 1984) and the "maximum entropy" methods (see review by Narayan and Nityananda 1986).

Self-calibration is based on the crucial assumption that the N(N-1)/2 (number of baselines) phase or amplitude errors $\Psi_{ij} = \psi_i - \psi_j$ at any given observing time can be reduced to N errors ψ_i attributable to individual interferometer elements. This assumption holds well within a certain accuracy; it implies the validity of the phase closure relation, which states that phase errors cancel when summing interferometer phases around a triangle of baselines, and the resulting sum, the 'closure phase' $\Phi_{12} + \Phi_{23} + \Phi_{31}$ contains true phase information. In a similar way one can at least partially achieve internal amplitude calibration (problem (c)) by taking advantage of the fact that the ratio of measured amplitudes $A_{1234} = A_{12}A_{23}/(A_{13}A_{24})$, the 'closure amplitude' is independent of antenna gain factors (indices 1,2,3,4) and therefore contains true amplitude information.

Problem (a) is overcome by applying the CLEAN method (Högbom 1974) to deconvolve the point spread function of the "synthesized aperture" from the image at each iteration cycle.

In principle there is a 'holographic' method also applicable to VLBI, called 'phase referencing', which allows the full visibility phase to be retrieved. This requires a point source near (in the same isoplanatic patch as) the target source to be observed at nearly the same time. Then the error phases can simply be removed by subtraction (Alef 1989).

See Table 1 for figures of merit concerning VLBI imaging and Figs. 6 and 7 for an example of VLBI imaging. For VLBI polarimetry we refer to Roberts et al. (1990) and references therein.

4 High-Precision Interferometry

The term 'high-precision interferometry' is used here to include all those applications of VLBI that rely on the exploitation of the group delay observable. It is this quantity which allows the determination of the "macroscopic" geometry of the interferometer, i.e. the baseline-source geometry that relates the location of the radio telescopes on the revolving Earth to the "infinitely" distant compact

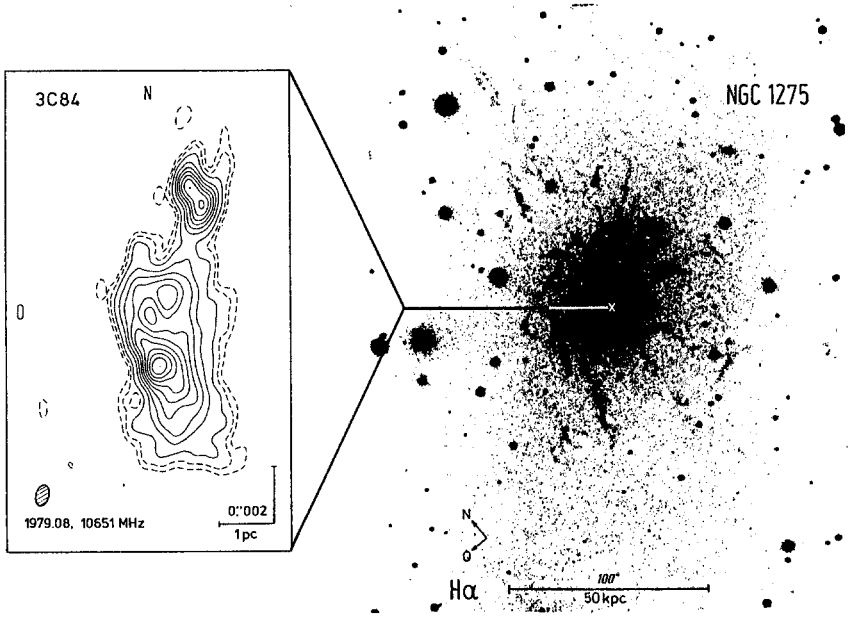

Fig. 6. 2.8 cm VLBI map of the compact source 3C84 obtained with a 7-station transatlantic array (Alef 1982). The mother galaxy NGC 1275 is shown here in the light of H_α. Note the scale difference of more than 4 orders of magnitude

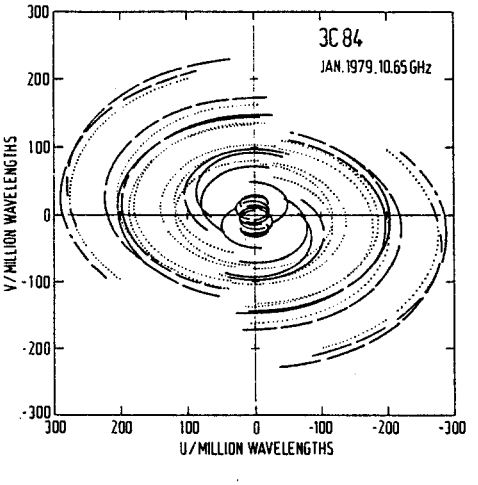

Fig. 7. Coverage of the aperture plane achieved in the 13 h observation which provided the data for the VLBI map shown in Fig. 6. The 21 elliptical traces in the (u,v) plane describe the changes of each projected baseline due to the Earth rotation as seen by the observed source.

radio sources. These pointlike emitters without proper motion are ideally suited to serve as fixed beacons in the heavens, allowing us to monitor even the smallest departures from the computed motions of the receiving stations.

The research fields that profit most from the geometric potential of VLBI are those dealing with the motions of the celestial bodies, in particular the Earth-Moon sytem, and the orientation and the size of the Earth itself: astrometry and geodesy. These fields are usually meant to imply a much broader area, namely fundamental astronomy and geo-sciences, such as geodesy, geophysics, oceanography etc.. The topic of gravitational light deflection is intimately related to all of these fields, because it forms part of the fundamental model describing the physical reality of VLBI.

The VLBI model is developed using the knowledge presently available to mathematically recreate, as closely as possible, the situation at the time of observation. This model then sets a standard against which a least squares parameter estimation algorithm is posed to determine the best values of the quantities to be solved for. Before this process starts, the raw observations have to be purged of several systematic effects, which in fact limit the final accuracy of the results. The flow diagram of a typical geodetic VLBI data analysis software package is shown in Fig. 10.

The systematic instrumental effects include clock instabilities, electronic delays in cables and circuitry and deformations of the telescope structure. The instrumental delay changes are monitored by the phase and delay calibration which is part of the MkIII system. In the telescope the distance between the feed horn and the axis intersection which constitutes the baseline reference point (Fig. 9), is assumed to be constant at the mm-level; in this case it becomes part of the clock offset parameter. Large telescopes such as the Effelsberg 100m antenna exhibit changes which can, however, be modeled to a level of a few millimeters (Rius et al. 1987).

The effect of the atmosphere on VLBI observations is considered to be the most serious problem, because at widely separated stations the elevation angles of the source during a scan differ greatly as well as the meteorological conditions themselves. The ionosphere, which is a highly dispersive medium in the lower radio frequency bands, can be dealt with to first order by using two different observing frequencies. In geodetic VLBI the NASA frequency pair of 2.3GHz (S-band) and 8.4GHz (X-band) is used throughout.

The influence of the neutral atmosphere, essentially the troposphere, on radio signals adds up to an extra zenithal path of 1.8 to 2.5 meters. The contribution of the dry part is rather stable, although care has to be taken to choose a proper model function for the lower elevation angles (Davis et al. 1985). The wet component, although the smaller part of the total tropospheric effect, changes rapidly and has to be monitored by some means. The most promising - albeit costly - method appears to be the radiometer technique, which consists of measuring the microwave thermal emission from water vapour near 22GHz in the line-of-sight (Elgered et al. 1982).

Now let us turn to the model side of the geodetic analysis.

The fundamental observation equation relating the group delay to the baseline and source vectors may be written in its simplest form:

$$\tau_0 = -\frac{\boldsymbol{b} \cdot \boldsymbol{k}}{c} \tag{17}$$

The baseline vector components b_x, b_y, b_z are referred to the instantaneous Earth rotation axis. The unit vector of the source $\boldsymbol{k}$ points to the apparent position at the time of observation[3]. The negative sign accounts for the fact that the motion of the incoming wave front is opposite to the direction of the unit vector k and the time delay is defined $\tau = t_2 - t_1$.

At this stage there are 3 + 2n fundamental parameters to be determined in a least squares fit: the three baseline components b_x, b_y, b_z and the coordinates α, δ of n observed sources. Due to the fact that the offset between the clocks at both stations is not known to better than around 100ns, clock parameters (usually offset and rate) have to be added to the solution. Therefore, minimal solutions are possible only with observations on at least three epochs and to at least two different sources. Usually a set of 12 to 18 sources spread over the sky as evenly as possible is used in a schedule of 24 hours during which these sources are observed in an interleaved mode in order to optimise the geometric condition of the solution.

In view of the extremely high precision inherent to VLBI the modeling accuracy has to be brought down to a level of better than one centimeter on the global scale. Great efforts have been made to develop comprehensive geodetic VLBI data analysis software systems which include all aspects of the multi-faceted reality of VLBI. Here we shall take a summary look at the most important model components with more emphasis only on the relativistic part of the model.

The fundamental geometric model of the time delay τ_g forms the heart of the system. This model has evolved from its basic form in a geocentric system to the fairly complex relativistic formulation in the solar system barycenter (SSB) (Finkelstein et al. 1983, Hellings 1986, Soffel et al. 1991):

$$\begin{aligned}\tau_g = {} & \tau_0 \left[1 - (\dot{\boldsymbol{R}} + \dot{\boldsymbol{r}}_2) \cdot \boldsymbol{k}\right] / c + \tau_0 \left[(\dot{\boldsymbol{R}} \cdot \boldsymbol{k})^2 + 2(\dot{\boldsymbol{R}} \cdot \boldsymbol{k})(\dot{\boldsymbol{r}}_2 \cdot \boldsymbol{k})\right] / c^2 \\ & + (\boldsymbol{b} \cdot \dot{\boldsymbol{R}}) \left[(\dot{\boldsymbol{R}} \cdot \boldsymbol{k})/2 + (\dot{\boldsymbol{r}}_2 \cdot \boldsymbol{k})\right] / c^3 - \tau_0 (U + \dot{R}^2/2 + \dot{\boldsymbol{R}} \cdot \dot{\boldsymbol{r}}_2)/c^2 \\ & - (\boldsymbol{b} \cdot \dot{\boldsymbol{R}})/c^2 + \tau_{grav}\end{aligned} \tag{18}$$

$$\tau^{\odot}_{grav} = \frac{(1+\gamma) r^{\odot}}{c} \ln \left[\frac{R_1 + \boldsymbol{R}_1 \cdot \boldsymbol{k}}{R_2 + \boldsymbol{R}_2 \cdot \boldsymbol{k}}\right] \tag{19}$$

$$\tau^{\oplus}_{grav} = \frac{(1+\gamma) r^{\oplus}}{c} \ln \left[\frac{r_1 + \boldsymbol{r}_1 \cdot \boldsymbol{k}}{r_2 + \boldsymbol{R}_2 \cdot \boldsymbol{k}}\right] \tag{20}$$

where $r^{\odot}$ and $r^{\oplus}$ are the Schwarzschild radii of the sun and the Earth. For the other bodies in the solar system (at present only Jupiter is considered when it is closer than 10° to any of the observed sources) the corresponding Schwarzschild

[3] Note: in this chapter and in Fig. 8 the symbols $\boldsymbol{b}$ and $\boldsymbol{k}$ are employed (instead of $\boldsymbol{d}$ and $\boldsymbol{s}$, Eq. (6)) in order to conform to standard usage in this field.

radii and the vectors from these bodies to the VLBI antennas have to be used in the above formula (19).

The complete relativistic formulation above (as implemented in the CALC, MkIII Data Analysis System, Ma 1978) includes both the effects of special relativity (SRT) and of general relativity (GRT), but for practical reasons these are treated separately and added together on the level of the time delay.

The effects of special relativity arise from the fact that quantities defined in coordinate frames moving relative to each other have to be related by transformations of the Lorentz type with v/c as the characteristic quantity in the time delay correction terms. The choice of two particular coordinate systems (the SSB and the geocentric system) used to describe the VLBI model arises from practical considerations: the motions of bodies in the solar system and the positions of the radio sources are most readily defined in the SSB, while the actual baselines between the telescopes are usually required in a geocentric system (Fig. 8). The vector $\dot{\boldsymbol{R}}$ describes the velocity of the geocenter with respect to the SSB ($\approx 30\,\mathrm{km/s}$) and $\dot{\boldsymbol{r}}_2$ is the velocity of the station 2 with respect to the geocenter ($\leq 0.46\,\mathrm{km/s}$). In the geometric VLBI model these velocities have to be computed with an accuracy of 10^{-6} in order to guarantee picosecond delay accuracy.

The last two terms in the full equation (18), which is a polynomial representation for picosecond accuracy, account for the difference in SSB-coordinate time and geocentric proper time as well as for the fact that the station clocks are located at fixed points on the Earth's crust. Here, U is the magnitude of the gravitational potential of the solar system at the geocenter.

The effect of gravity on the propagation of electromagnetic waves (GRT) is no less important, as was pointed out by Shapiro in 1979 (NASA VLBI conf.). Even at an angle of 180° away from the sun the differential delay effect for a 6000 km baseline is still 0.4 ns (Tab. 2).

According to GRT, space-time is deformed by the presence of masses. The most massive object in our vicinity is of course the sun, which accounts for more than 99% of the total effect. An impression of the enormous impact of $\tau^{\odot}_{\mathrm{grav}}$ on the observed group delays of a typical geodetic VLBI experiment is shown in Fig. 12. The residual spread is greater by a factor of ≈ 5 if τ_{grav} is neglected (Schuh 1987). However, at the present accuracy level of VLBI the major planets also contribute a bending effect which cannot be entirely neglected. If Jupiter arrives within a few degrees of an observed source, its influence on ray bending has to be taken into account (Tab. 2). Another small but significant contribution ($\leq 20\,\mathrm{ps}$) comes from the gravity field of the Earth itself.

The formulation for τ_{grav} shown here in (19) and (20) is implemented in the current version 7.0 of the CALC MkIII VLBI software (Eubanks (ed.) 1991).

Since the early 80's VLBI observations have been used extensively to test Einstein's theory in its Parameterized Post Newtonian formulation (PPN). Two approaches have been used, one designing special experiments to observe sources such as 3C279 and 3C273 during their close approch to the sun (Fomalont and Sramek 1977, Shapiro 1967) and the other using all available data from routine geophysical experiments to achieve the accuracy by the sheer number of the ob-

servations (Carter et al. 1985). In the case of specially designed experiments with observations close to the sun, one has to cope with the effects of the sun's radio noise (loss in SNR) and the coronal path delays. While the latter can be removed to a large extent by dual frequency observations (e.g. in the S/X bands), the former requires strong sources and a trade-off consideration between distance from the sun and strength of the gravitational bending effect.

The γ-factor, which in the Einstein theory should be equal to unity, has been found to show no significant departure from this value to the level of 0.1%. Recently this accuracy level has been further improved to 0.02% (Robertson et al. 1991). Attempts have also been made to verify the gravitational bending near Jupiter, but the effect is only marginally significant ($\lesssim 100$ ps) at close ($\lesssim$ several arcmin) encounters (Campbell 1989, Treuhaft and Lowe 1991).

The description of the Earth's orientation with repect to the celestial system (precession, nutation) and the motion of the Earth's axis with respect to the crust (polar motion) has to reach the same level of accuracy as all the other model components, which means roughly $0''.001$. The same holds true for the rotational speed of the Earth about its axis. To compute the phase angle of the Earth's rotation to 1 milliarcsec the UT1 variations have to be known to better than 0.1 msec. For VLBI these requirements cannot be met without parametrisation. Therefore, with longer series of VLBI experiments, the nutation parameters in longitude and obliquity, and the components of polar motion x_p, y_p plus UT1 are included as parameters in the least squares solution. Also, precession can be solved for if longer time spans of data are analysed.

Periodic and aperiodic deformations of the Earth's crust have to be taken into account as well. Solid Earth tides show diurnal and semidiurnal oscillations with vertical amplitudes of about 40 cm and horizontal displacements of about 10% of the vertical effect. Although good models are available, the relevant parameters (the Love numbers) can be estimated from larger sets of data (Herring et al. 1983). More difficult to model are the tidal loading effects of the oceans, which amount to as much as a decimeter on some coastal or island sites (Schuh and Möhlmann 1989). The loading effects of the atmosphere also reach the level of significance in VLBI modeling.

Solutions with large data sets are stable enough to produce precise source positions simultaneously with the baseline components and other parameters. The positions of some fifty compact radio sources, distributed over the whole northern sky, are by now known at an accuracy level of 1 milliarcsec (Ma and Shaffer 1988). Over short ($\lesssim 0.^\circ 5$) arclengths even much higher positional accuracies ($\gtrsim$10 microarcsec) have been achieved by means of differential methods (Shapiro et al. 1979, Marcaide and Shapiro 1983). Using the southern stations at Hartebeesthoek/South Africa and Hobart/Tasmania the source list is being extended southwards to achieve a uniform coverage of the entire celestial sphere (Carter et al. 1988).

A major problem is constituted by the fact that most of the observed compact sources tend to show structure at the level of a few milliarcsec. These effects, in particular the changes in the structure, pose a limit on the accuracy of the radio

reference system. Continuous monitoring of the structure, which is also accomplished by analysing VLBI data can be done in parallel to the geodetic analysis, thus providing a means of correcting for the structure effects (Schalinski et al.1988, Campbell et al. 1988).

The system shown in Fig. 10 can be seen to have two main streams, one containing the actual observations which undergo the successive instrumental and environmental corrections, and the other to produce the so-called theoreticals, beginning with the "a priories", a set of starting values for the parameters to be estimated. Both streams converge at the entrance to the least squares algorithm, where the "observed minus computed" are formed.

In geodetic VLBI data processing there are two levels of least squares solutions, one in which only the "local" unknowns are estimated (such as clocks and atmospheric parameters) thus creating a first data base version of each particular experiment, and another which collects all available experiments for a combined solution including the "global" unknowns such as station and source positions, Earth rotation parameters, etc. and – last but not least – the γ_{PPN} parameter.

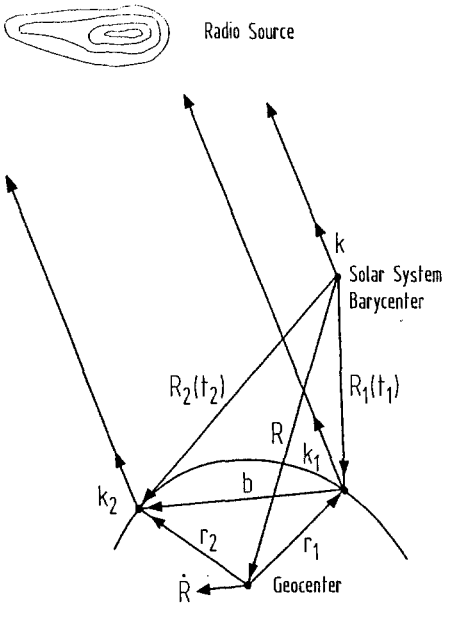

Fig. 8. VLBI geometry referred to solar system barycenter (Schuh 1987)

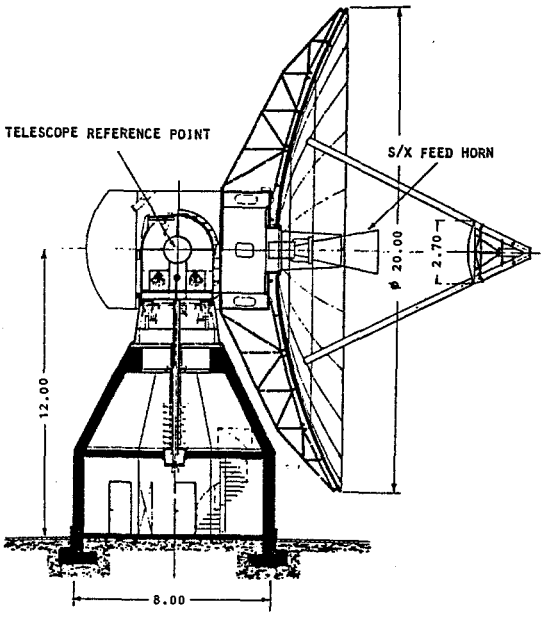

Fig. 9. The 20 m radiotelescope of the geodetic fundamental station Wettzell, Bavaria, Germany

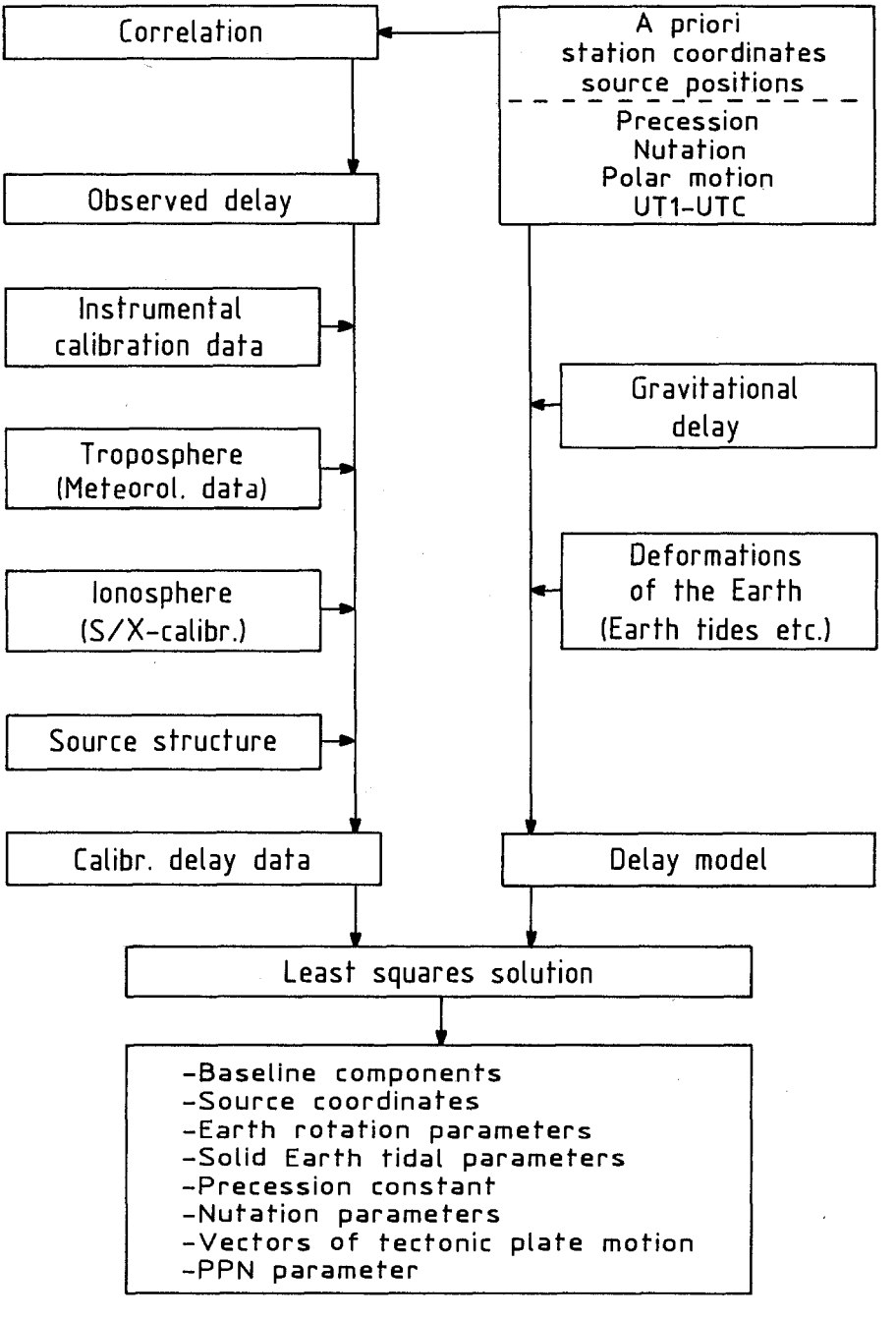

Fig. 10. Flow diagram of a geodetic VLBI software system

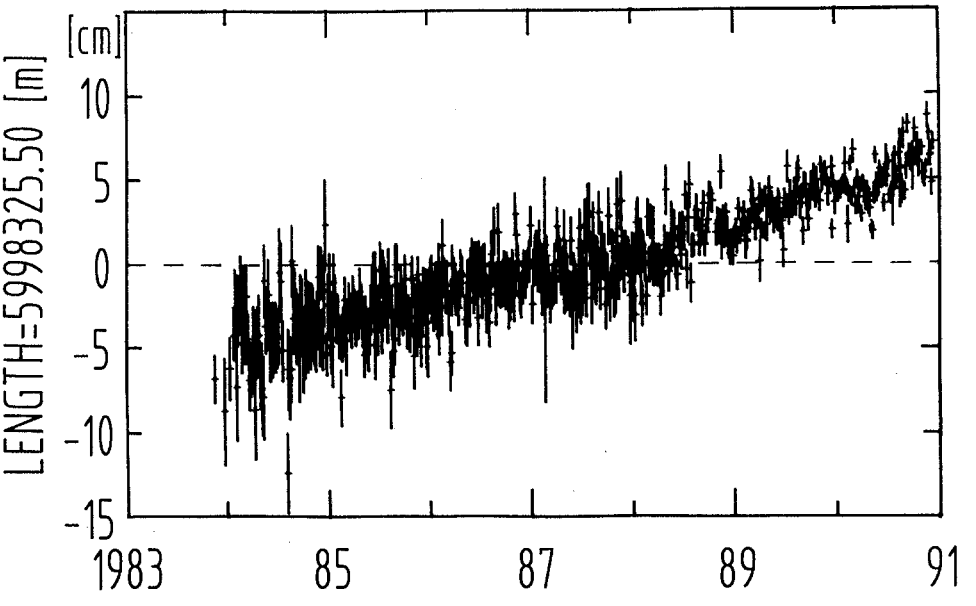

Fig. 11. Evolution of baseline length [cm] between Wettzell, Germany, and Westford, Mass., USA, 1984–1991

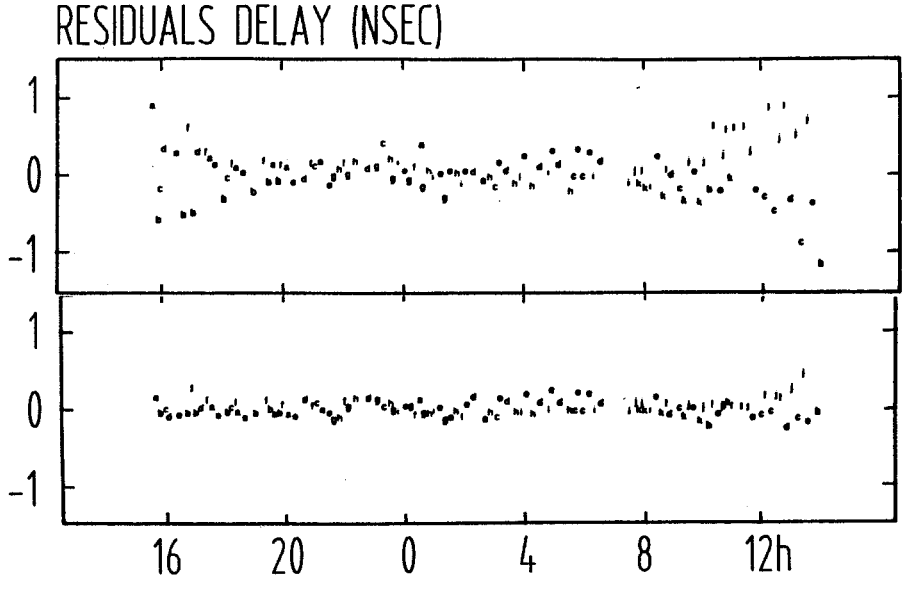

Fig. 12. Delay residuals [ns] from fit for a Bonn– Haystack observation of 12 radio sources, May 1983, without (top) and with (bottom) corrections for the solar gravitational delay (Schuh 1987)

Table 1. Instrumental Characteristics of VLBI (1991)

Angular resolution $(0.5 \cdot \lambda/d_\perp)$	$0''.001 \cdot (\lambda/\mathrm{cm}) / (d_\perp/1000\,\mathrm{km})$
rms noise $\sigma(S_{corr})$	$\gtrsim 1\,\mathrm{mJy} = 1 \cdot 10^{-29}\mathrm{Wm}^{-2}\mathrm{Hz}^{-1}$
Point source detection limit	$\approx 7 \cdot \sigma(S_{corr})$
Imaging capability	
Minimum object flux S_ν(min)	$\gtrsim 80\mathrm{mJy}$ $(B = 56\,\mathrm{MHz})$
Observing time required	≈ 12 hours
Dynamic range	$\lesssim 100:1$ (few cases: O(1000:1))
Polarisation mapping	first steps at $0''.001$ resolution
Measurement accuracies	
Group delay τ	$\gtrsim 1.5 \times 10^{-11}$ s (B_s=360 MHz)
Delay rate $\dot{\tau}$	$\gtrsim 1.0 \times 10^{-13}$ s/s
Baseline length d	$\lesssim 1.5$ cm for d$\lesssim 6000$ km
Polar motion	$\lesssim 0''.001$
UT1 variation	$\lesssim 0.1$ ms
Time resolution	$\lesssim 1$ day $(\approx 2$h for UT1)
Source position	$\lesssim 0''.001$ (whole sky)
Relative source position	$\gtrsim (10 \times 10^{-6})''$ (arc length $\lesssim 0.^\circ 5$)
Technical characteristics	
Standard observing wavelengths λ	92, 18, 13, 6, 3.6, 2, 1.3 cm
Standard frequencies ν	.33, 1.7, 2.3, 5.0, 8.4, 15, 22 GHz
Experimental wavelengths	7 mm, 3.5 mm, 1.3 mm
Experimental frequencies	43 GHz, 100 GHz, 230 GHz
Antenna diameter D	10 to 100 m
Baseline length d	$\lesssim 10{,}000$ km
Relative frequency stability $\delta\nu/\nu$	$\gtrsim 10^{-15}$ (H-maser)
Clock synchronisation accuracy	$\approx 1\,\mu$s a priori
Interferometer coherence time t_{coh}	$\gtrsim 10$ min $(\lambda \gtrsim$ cm)
Coherent integration time t_{int}	$\leqq t_{coh}$
System noise temperature T_{sys}	$\lesssim 100$ K $(\lambda \gtrsim$ cm)
Sampling rate	$\leqq 224$ Mbit/s (future: 2 Gbit/s)
Recording bandwidth B	$\leqq 112$ MHz (future: 500 MHz)
Spanned bandwidth B_s	$\leqq 360$ MHz (future: 720 MHz)
Group delay τ_g	$\approx 3.3\,\mathrm{ms} \cdot d_\parallel[1000\,\mathrm{km}]$
Natural fringe frequency $\nu_f = \nu\dot{\tau} \approx$	$7.3\,\mathrm{kHz}\,\cos\delta\,d_{EW}[1000\,\mathrm{km}] / \lambda[\mathrm{cm}]$
Crosscorrelation coefficient ρ_{12}	$\approx 10^{-5}$ to 10^{-3}
Accumulation time t_{acc}	≈ 1 s def.: $\hat{\rho}_{12} \propto \langle v_1 v_2^* \rangle_{t(acc)}$
Residual fringe frequency ν_f^{res}	1 to 10 mHz
Duration of observing run t_{obs}	≈ 12 h (geodesy: ≈ 24 h)

Table 2. Gravitational path delay as a function of spherical distance Θ from the Sun and Jupiter. Given are maximum values for a 6000 km baseline (Schuh 1987)

Θ(Sun)[°]	τ^{S}_{grav}[ns]	Θ(Jupiter)[°]	τ^{J}_{grav}[ns]
0.267	169.52	Rim	1.582
1	45.30	0.017 ($\hat{=}$ 1′)	0.605
5	9.06	0.167 ($\hat{=}$ 10′)	0.062
10	4.54	0.5	0.021
30	1.53	1	0.010
60	0.79	5	0.002
90	0.56	10	0.001
120	0.46		
150	0.41		
180	0.40		

5 VLBI Today and Tomorrow

It is clear from what we have said before that VLBI observations yield detailed information on all astrophysical processes which produce compact radio sources (brightness temperature $T_b \gtrsim 10^6$K), and on all geophysical processes which change the vector spacings between radio telescopes fixed to the Earth's surface or which cause irregularities in the Earth's rotation (geodynamical processes). VLBI also enables us to test and discriminate between gravitational theories by measuring the light deflection or path delay caused by the sun and its planets.

The increasing angular resolution and precision of VLBI allow the investigation of processes causing short term variability of the astro- and geophysically relevant observables. Both the structure of radio sources on milliarcsec scales and the geophysical quantitities have been found to vary on time scales ranging from years to weeks or even days. It is the signature of the temporal behaviour of the observed phenomena which carries key information on the underlying physics. This is why monitoring programs have become more and more important for both astro- and geophysics.

The apparently brightest and therefore most easily observable compact sources ($\lesssim$ 0.″01) at dm/cm wavelengths are the nonthermal continuum sources found in the nuclei of quasars and radio galaxies, and the line-emitting H_2O (λ 1.35 cm) and OH (λ 18 cm) maser sources found near newly formed or evolved stars. There are thousands of compact extragalactic and hundreds of maser sources in the sky. Most of the current astronomical VLBI observations are devoted to these objects. Higher sensitivity is required for the observation of radio nuclei in nearby (mildly) active galaxies, of nonthermal continuum sources associated with nearby flare stars or X-ray binary stars, and for the observation of molecular masers in external galaxies.

With the milliarcsec resolution of intercontinental VLBI networks one typically achieves spatial resolutions on scales of about 1 to 10 pc for radio galaxies and quasars, of about 0.5 to 5 AU for interstellar molecular masers within our galaxy, and of about 0.05 AU for nearby stars (Fig. 13).

The exact location of very compact objects relative to a well-defined mass centre (star, galaxy) is mostly uncertain. They are, as far as we know, often associated with, or embedded in, zones of inflow on to or outflow from their parent objects covering wide ranges of size and power. Physical processes which can generate compact emission are: accretion on to collapsed massive objects in the centres of galaxies (AGNs) or on to collapsed stars (X-ray binaries); collimated outflow (jets) from AGNs or collapsed stars; outflow in star forming regions (H_2O, OH masers), "superwinds" from late-type stars (H_2O, OH masers), and outbursts on stellar surfaces.

Compact radio sources have a multiple function: they are fascinating objects in their own right, but at the same time they are diagnostic tools for probing the underlying larger scale processes such as collimated outflow from AGN or newly formed stars. Most remarkable is the use of molecular masers (expanding point source clusters) as distance indicators, finally the use of extragalactic sources as reference points for geodetic and geodynamical measurements.

Some important areas of impact of VLBI on astrophysics are summarized below; the order is by classes of compact sources, i.e. the immediate targets of observation.

- Extragalactic radio sources (quasars, radio galaxies, BL Lac type objects, Seyfert galaxies, mildly active galaxies):
  - physics of the "central engine" in AGNs,
  - physics of collimated outflow, the role of relativistic bulk motion,
  - relation between radio emission and emission in other spectral domains,
  - reference sources for astrometry and geodesy,
  - gravitational lenses,
  - scintillating background sources (interstellar medium).
- H_2O and OH masers in our galaxy:
  - physics of masers and their environment,
  - kinematics of outflows in proto-stellar nebulae,
  - mass-loss from late-type stars, giants, supergiants,
  - distances within the Galaxy via the "statistical parallax" method,
  - scattering properties of the interstellar medium.
- Extragalactic masers:
  - physics of "megamasers",
  - distance to nearby galaxies, Hubble constant (space VLBI).
- Continuum sources associated with stars:
  - nearby flaring stars,
  - X-ray binary stars,
  - pulsars (astrometry).

Astronomical VLBI observations are nowadays normally done within the framework of networks which dispose of block time committed to VLBI by member observatories. About 30 radio telescopes located in about 20 countries take part in VLBI activities, and more than fifty are expected to do so in the near future (Fig. 1). The largest global VLBI experiments have used 18 stations.

While early VLBI observations were organised as ad hoc collaborations of observatories, VLBI networks in the USA and Europe, starting in the late 1970's, have turned into facilities run like single observatories and open to any competent user from the international community. The European and US VLBI networks typically run 4 observing sessions per year. The 100 m telescope near Bonn, for example, is committed to 90 days of VLBI observations per year, and it actually spends about 30% of its local observing time on VLBI. The networks have program committees for assessment of observing proposals, central scheduling of VLBI block time, "absentee observing", VLBI processing centres, and software packages for signal analysis and image reconstruction.

The US Network is about to be replaced by the VLBA (Very-Long-Baseline Array) (Romney 1988). The VLBA will be the world's first dedicated multipurpose VLBI network. The array will be available for astronomical and astrometric/geodetic observations. It will consist of ten 25 m antennae, in a configuration optimized to provide high resolution and high image quality in a large field of view over a wide range of declinations. 9 frequency bands from 330 MHz to 43 GHz are planned. The operations centre will be in Socorro, New Mexico. By combining the VLBA with large telescopes, it will be possible to form a very sensitive global ($\approx$ 20 station) array.

Table 1 summarises the current instrumental performance of VLBI. Processing centres running broad band ($\gtrsim$50 MHz) correlators, currently in operation or in progress, are in (type, number of station inputs, comment):

- Haystack, Ma., USA (MkIII A, 6, astronomy, geodesy),
- Washington, D.C., USA (MkIII A, 5, geodesy),
- Bonn, Germany (MkIII A, 6, astronomy, geodesy),
- Kashima, Japan (K3, 2, geodesy; K4, $\gtrsim$10, space VLBI, planned)
- Socorro, NM, USA (VLBA, 20, astronomy, geodesy, under construction)

Some technical developments will have a considerable impact on future VLBI; these include improved VLBI systems, mm-VLBI, and space VLBI.

Important figures of merit describing the practical use of VLBI systems are: the maximal recording rate (= 2 × B for 1-bit sampling), the recording density on given storage media (magnetic tapes), and the number of baselines which can be correlated simultaneously at a given bandwidth. Corresponding numbers for the widely used Mark IIIA system are: 224 Mbit/s (recording rate), 3 hours of 56 MHz data on one tape, correlation of 12 baselines at 28 MHz simultaneously (Bonn Mk IIIA correlator, 1992). The Mk III system and its more advanced versions (Mk IIIA, Mk IV) have been developed at Haystack Observatory, USA (Rogers et al. 1983, Whitney 1988, Webber and Hinteregger 1988). The Mk IV system will be operational in the near future. Maximum recording rates of 1 Gbit/s and

recording at 128 Mbit/s for 24 hours on one tape will be possible. Other VLBI systems which have been developed in recent years, and which are about to come into operation, are: the VLBA system (Romney 1988), the Canadian S2 system (Wietfeldt et al. 1992), and the Japanese K4 system (Kawaguchi 1992, Chikada et al. 1992). Some of these systems are (within limits) compatible. In principle all of them can be made compatible via appropriate interfaces.

Millimeter-VLBI has made considerable progress in recent years; see Krichbaum and Witzel (1992) for observations at 43 GHz and Baath (1992) for 100 GHz VLBI. At 3 mm an angular resolution of 50 microarcseconds has been achieved. After improvements in receiver techniques the sensitivity of mm-VLBI is at the moment mainly limited by the short instrumental coherence time ($\approx$30 s at 7 mm, $\approx$10 s at 3 mm) and the limited number of the required large telescopes (collecting area). The future broad band VLBI systems will be most helpful in further enhancing the sensitivity of mm-VLBI.

The technical feasibility of space VLBI has been demonstrated by two experiments (Levy et al. 1986) involving a 4.9 m antenna on a TDRSS (Tracking and Data Relay Satellite System) satellite in conjunction with two ground based 64 m radio telescopes in Australia and Japan. The longest baseline used in these experiments was 2.2 Earth diameters. There are two approved dedicated space VLBI missions, both planned to be launched around 1995: the Japanese mission VSOP (9 m antenna, 20,000 km orbit) and the Russian mission RADIOASTRON (10 m antenna, 80,000 km orbit). Observing wavelengths will be 1.35 cm, 6 cm, 18 cm, and (Radioastron) 92 cm.

From the very beginning the scientific motivation for the development of VLBI came from both astrophysics ***and*** geophysics. Shapiro gave a detailed description of the expected geophysical applications of VLBI as early as in 1969 at a conference held in London, Canada, on "Earth-quake displacement fields and the rotation of the Earth" (Shapiro and Knight 1970). In the following two decades virtually all of the goals mentioned there and even more could be achieved:

- the creation of a quasi-inertial extragalactic reference system as a basis for astrometry to study galactic rotation and to improve the distance scale of the universe. Recent results of stellar astrometry with HIPPARCOS are being tied to the extragalactic reference system.
- the realisation of a global terrestrial reference system in order to satisfy the needs of global geodetic and navigational systems (including spacecraft navigation),
- monitoring the Earth rotation parameters (polar motion and UT1 variations) with the highest possible resolution for a better understanding of the kinematics and dynamics of the "System Earth", i.e. the Earth in space, the effects of the atmosphere and the oceans, and the processes in the Earth's interior,
- the determination of improved coefficients for precession and nutation and the estimation of the Earth's elasticity parameters, thereby also contributing to a more comprehensive understanding of the System Earth (Herring et al. 1986),

- the determination of regional and global crustal motions to verify the plate tectonical models and to study the processes at the plate boundaries, with the aim to contribute to earthquake prediction research,
- the determination of the γ_{PPN} parameter in General Relativity.

Today the accuracies required to attain these goals have been demonstrated by hundreds of VLBI experiments on baselines connecting almost all major continents of the globe. In order to combine the efforts in different countries around the world in realising these goals, several programs of international cooperation have been launched, among which the following are the most important:

- The NASA Crustal Dynamics Project (CDP)
  This project is part of a US Federal program involving several government agencies for the application of space technology to crustal dynamics and earthquake research. Cooperative arrangements have been made with European and other countries extending the project to a global research program (NASA 1988). The VLBI part of the CDP comprises regular experiments (10-20 each year) of one to three days duration between the major geodetic VLBI facilities in the US, Europe and in and around the Pacific Ocean. In addition so-called bursts of observations are carried out each year using the mobile VLBI units to monitor tectonically interesting sites in California and Alaska (NASA 1988, Ma et al. 1989).
- Project IRIS (International Radio Interferometric Surveying)
  The aim of the IRIS program is to conduct VLBI observations at regular intervals to monitor polar motion and UT1. This observational program began in 1980 under the acronym of POLARIS and has been extended in several steps. It now comprises three networks, the original IRIS-A (Atlantic) network with three stations in the USA (Westford, Richmond and Mojave) and two in Europe (Wettzell and Onsala), the IRIS-S (South) network with Hartebeesthoek added to IRIS-A and the IRIS-P (Pacific) network combining four stations around the Pacific (Kashima/Japan, Hobart/Tasmania, Fairbanks/Alaska, and Mojave/California) (Carter et al. 1985, Carter et al. 1988).

Both programs have profited significantly from the broad international cooperation and have produced impressive results. The determination of Earth rotation parameters by the IRIS network has superseded the classical methods with respect to accuracy and time resolution. The data have begun to show hitherto unseen phenomena such as the influence of the zonal winds of the atmosphere and departures from the normal state, e.g. the El Niño events in the southern Pacific (Chao 1989). Even the short period tidal influence of the oceans could be detected using large numbers of data sets (Brosche et al. 1991).

Both the IRIS project and NASA's Crustal Dynamics Program (CDP) could in recent years, for the first time see relative motions of tectonic plates; this is a most remarkable moment in the development of plate tectonics which started with Wegener's continental drift hypothesis in 1912. The baseline length changes detected by VLBI can be seen to confirm the plate models to a surprisingly good level (Ma et al. 1989). A prominent example is the 6,000 km Westford to Wettzell

baseline which now has a record of nearly 8 years of uninterrupted observations and displays a very significant trend (1.8 ± 0.1 cm/y) which is in good agreement with the predicted relative tectonic motion of the American and Eurasian plates (Fig. 11). Problems occur of course at stations near the plate boundaries, where the influence of the processes associated with the formation and subduction of the crust can be clearly seen (Heki et al. 1990).

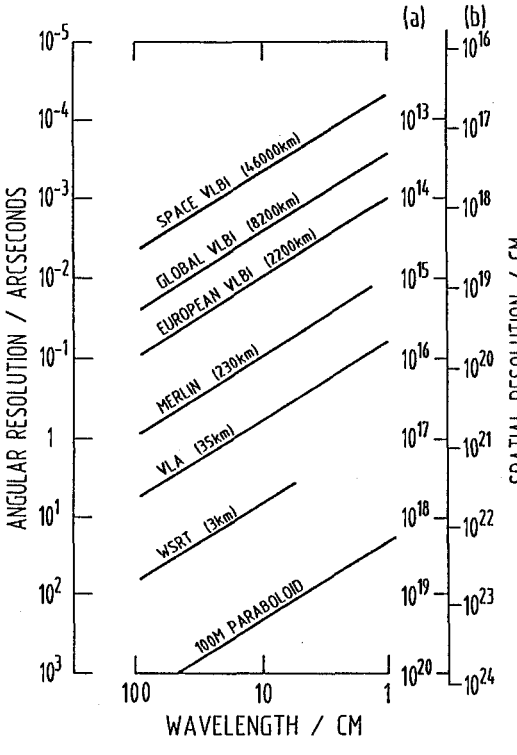

Fig. 13. Angular resolution (λ/d(max)) for VLBI and some local instruments in the wavelength range 1–100 cm. The spatial resolution is given for a) the Galactic Center (distance 7.1 kpc), and b) 3C84 = NGC 1275 (distance 55 Mpc)

Fig. 14. The 100m radiotelescope, Effelsberg, Germany, the biggest VLB interferometer element routinely used for VLBI

Acknowledgment.

The authors thank David Graham, Ivan Pauliny-Toth, Geoffrey Ediss, and Harald Schuh for critically reading the manuscript and Edward Fomalont for comments regarding light deflection measurements.

References

Alef, W. (1982): Diplomarbeit, Max-Planck-Institut für Radioastronomie, Bonn

Alef, W. (1989): *Introduction to phase-reference mapping*, in (eds.) Felli and Spencer (1989), pp. 261–274

Baath, L.B. (1992): *Results from 100 GHz VLBI*, in (eds.) Hirabayashi et al. (1992)

Brosche, P., Wünsch, J., Campbell, J., Schuh, H. (1991): *Ocean tide effects in Universal Time detected by VLBI*, Astron. Astrophys. **245**, pp. 676–682

Campbell, J. (1989): *Mesure de l'effect relativiste de Jupiter par VLBI*, in *Systemes de reference spatio-temporels*, N. Capitaine (ed.), (Observatoire de Paris), pp. 55–63

Campbell, J., Schuh, H., Zeppenfeld, G. (1988): *On the Computation of Group Delay Corrections Caused by Radio Source Structure*, in (eds.) Reid and Moran (1988), pp. 427–428

Carter, W.E., Robertson, D.S., MacKay, J.R. (1985): *Geodetic Radiointerferometric Surveying: Applications and Results*, J. Geophysical Research **90**, pp. 4577–4587

Carter, W.E. et al. (1988): *IRIS-S: Extending Geodetic Very Long Baseline Interferometry Observations to the Southern Hemisphere*, J. Geophys. Research **93**, pp. 14947–14953

Carter, W.E. (Convenor) (1991): Proceedings of the AGU Chapman Conference on *Geodetic VLBI: Monitoring Global Change* held at Washington, D.C., Apr.1991. NOAA Technical Report NOS 137 NGS 49, US Dept. of Commerce, NOAA/NOS, Rockville, MD, USA

Chao, B.F. (1989): *Length-of-day variations caused by El Niño – Southern Oscillation and quasi-biennial oscillation*, Science **243**, pp. 923–925

Chikada, Y. (1992): *The VSOP Correlator*, in (eds.) Hirabayashi et al. (1992)

Clark, T.A. et al. (1985): *Precision Geodesy using the Mk III Very-Long-Baseline Interferometer System*, IEEE Transactions on Geoscience and Remote Sensing GE **23** pp. 438–449

Cornwell, T.J. (1987): *Imaging at Radio Wavelengths*, Proc. of ESA Workshop on *Optical Interferometry in Space*, Granada, Spain, June 1987, ESA SP–273, pp. 31–36

Cornwell, T.J., Perley, R.A. (eds.) (1991): *Radio Interferometry: Theory, Techniques, and Apllications*, (Astronomical Society of the Pacific, San Francisco)

Davis, J.L. et al. (1985): *Geodesy by Radio Interferometry: Effects of Atmospheric Modeling Errors on Estimates of Baseline Length*, Radio Science **20** pp. 1593–1607

Elgered, G., Rönnäng, B., Askne, J. (1982): *Measurements of Atmospheric Water Vapour with Microwave Radiometry*, Radio Science pp. **17**, pp. 1258–1264

Eubanks, T.M. (ed.) (1991): Proceedings of the U.S. Naval Observatory Workshop on *Relativistic Models for Use in Space Geodesy*, June 1991, (U.S. Naval Observatory, Washington D.C., USA)

Felli, M., Spencer, R.E. (eds.) (1989): *Very Long Baseline Interferometry: Techniques and Applications*, (Kluwer Academic Publishers, Dordrecht)

Finkelstein, A.M., Kreinovich. V.J., Pandey, S.N. (1983): *Relativistic Reductions for Radiointerferometric Observables*, Astrophys. and Space Science **94**, pp. 233–247

Fomalont, E.B., Sramek, R.A. (1977): *The Deflection of Radio Waves by the Sun*, Comments on Astrophysics **7**, pp. 19–33

Heki, K., Takahashi, Y., Kondo, T. (1990): *Contraction of northeastern Japan*, Tectonophysics **181**, pp. 113–122

Hellings, R.W. (1986): *Relativistic Effects in Astronomical Timing Measurements*, Astron. J. **91**, pp. 650–659

Herring, T.A. et al. (1983): *Determination of Tidal Parameters from VLBI Observations*, Proc. Ninth Int. Sympos. on *Earth Tides*, ed. J. Kuo, (Schweizerbart'sche Verlagsbuchhandlung, Stuttgart), pp. 205–240

Herring, T.A., Gwinn, C.R., Shapiro, I.I. (1986): *Geodesy by Radio Interferometry: Studies of the Forced Nutations of the Earth, I. Data Analysis*, J. Geophysical Research **91**, pp. 4745-4754

Hirabayashi, H., Inoue, M., Kobayashi, H. (eds.) (1992): *Frontiers of VLBI*, (Universal Academy Press, Tokyo)

Högbom, J.A. (1974): Astron. Astrophys. Suppl. **15**, pp. 417–26

Kawaguchi, J. (1992): *VLBI Recording System in Japan*, in (eds.) Hirabayashi et al. (1992)

Krichbaum, T.P., Witzel, A. (1992): *Astronomical Results from Recent 7mm-VLBI Campaigns*, in (eds.) Hirabayashi et al. (1992)

Levy, G.S. et al. (1986): *Very Long Baseline Interferometric Observations made with an Orbiting Radio Telescope*, Science **234**, pp. 187–189

Ma, C. (1978): PhD Thesis, NASA/GSFC, Greenbelt, MD, USA, Techn. Memo No. 79582

Ma, C., Shaffer, D.B. (1988), in (eds.) Reid and Moran (1988), pp. 325–326

Ma, C., Ryan, J.W., Caprette, D. (1989): *Crustal Dynamics Project Data Analysis - 1988, VLBI Geodetic Results 1979-87*, NASA Technical Memorandum 100723, Greenbelt, Md., Feb. 1989

Mandel, L., Wolf, E. (1965): *Coherence Properties of Optical Fields*, Rev. Mod. Phys. **37**, pp. 231–287

Marcaide, J.M., Shapiro, I.I. (1983): *High Precision Astrometry via Very-Long-Baseline Radio Interferometry: Estimate of the Angular Separation between the Quasars 1038+528A and B*, Astron. J. **88**, pp. 1133–1137

Meeks (ed.) (1976): *Methods of Experimental Physics*, Part C: Radio Observations, (Academic Press, New York)

Moran, J.M. (1976): *Very Long Baseline Interferometric Observations and Data Reduction*, in *Methods of Instrumental Physics*, Vol. 12, Part C, *Radio Observations*, edited by M.L. Meeks, (Academic Press, New York)

Narayan, R., Nityananda, R. (1986): *Maximum Entropy Image Restoration in Astronomy*, Ann. Rev. Astron. Astrophys. **24**, pp. 127–170

NASA (1988): *NASA Geodynamics Program, Summary Report: 1979–1987, Progress and Future Outlook*, NASA Technical Memorandum 4065

Pearson, T.J., Readhead, A.C.S. (1984): *Image Formation by Self-Calibration in Radio Astronomy*, Ann. Rev. Astron. Astrophys. **22**, pp. 97–130

Perley, R.A., Schwab, F.R., Bridle, A.H. (eds.) (1989): *Synthesis Imaging in Radio Astronomy*, A Collection of Lectures from the Third NRAO Synthesis Imaging Summer School, (Astronomical Society of the Pacific, San Francisco)

Readhead, A.C.S., Wilkinson, P.N. (1978): *The Mapping of Compact Radio Sources from VLBI Data*, Astrophys. J. **223**, pp. 25–36

Reid, M.J., Moran, J.M. (eds.) (1988): *The Impact of VLBI on Astrophysics and Geophysics*, Proc. IAU Symposium No. 129, (Kluwer Academic Publishers, Dordrecht)

Rius, A., Rodriguez, J., Campbell, J. (1987): *Geodetic VLBI with Large Antennas*, in (eds.) J. Campbell and H. Schuh: Mitt. Geod. Inst., Univ. Bonn, No. 72, pp. 59–67

Roberts, D.H., Wardle, J.F.C., Brown, L.F., Gabuzda, D.C., Cawthorne, T.V. (1990): *Parsec-Scale Linear-Polarization Properties of Quasars, Galaxies, and BL Lacertae Objects*, in (eds.) Zensus and Pearson (1990), pp. 110–116

Robertson, D.S., Carter, W.E., Campbell, J., Schuh, H. (1985): *Daily Earth Rotation Determinations from IRIS Very Long Baseline Interferometry*, Nature **316**, pp. 424–427

Robertson, D.S., Carter, W.E., Dillinger, W.H. (1991): *New measurement of solar gravitational deflection of radio signals using VLBI*, Nature **349**, pp. 768–770

Robertson, D.S. (1991): *Geophysical Applications of Very-Long-Baseline Interferometry*, Rev. Mod. Phys. **63**, 899–918

Rogers, A.E.E. (1970): *Very Long Baseline Interferometry with Large Effective Bandwidth for Phase Delay Measurements*, Radio Science **5**, pp. 1239–1248

Rogers, A.E.E. et al. (1983): *Very-Long-Baseline Interferometry: The Mark III System for Geodesy, Astrometry, and Aperture Synthesis*, Science **219**, pp. 51–54

Romney, J.D. (1988): *The Very Long Baseline Array*, in (eds.) Reid and Moran (1988), pp. 461–468

Schalinski, C.J. et al. (1988): *First Results of the VLBI Investigation of Sources from Geodetic IRIS-Experiments*, in (eds.) Reid and Moran (1988), pp. 359–360

Shapiro, I.I. (1967): *New Method for the Detection of Light Deflection by Solar Gravity*, Science **157**, 806–808

Shapiro, I.I., Knight, C.A. (1970): *Geophysical applications of long baseline radio interferometry*, in *Earthquake Displacement Fields and the Rotation of the Earth*, eds. L. Mansinha, D.E. Smylie, and A.E. Beck, (Springer, New York), pp. 284–301

Shapiro, I.I. et al. (1979): *Submilliarcsecond Astrometry via VLBI: Relative Position of the Radio Sources 3C345 and NRAO 512*, Astron. J. **84**, pp. 1459–1469

Schuh, H. (1987): *Die Radiointerferometrie auf langen Basen zur Bestimmung von Punktverschiebungen und Erdrotationsparametern*, PhD Thesis, Universität Bonn, 1986, publ. in: Deutsche Geodätische Kommission, Reihe C, Heft 328, München

Schuh, H., Möhlmann, (1989): *Ocean loading Displacements by VLBI*, Geophysical Res. Letters **16**, pp. 1105–1108

Soffel, M.H.(1989): *Relativity in Astrometry, Celestial Mechanics and Geodesy*, (Springer-Verlag, Berlin)

Soffel, M.H., Müller, J., Wu, X., Xu, C. (1991): *Consistent Relativistic VLBI Theory with Picosecond Accuracy*, Astron. J. **101**, pp. 2306–10

Thomas, J.B. (1972): *An Analysis of Long Baseline Radio Interferometry*, DSN Report, Technical Report 32-1526, Vols. VII, VIII, XVI, Jet Propulsion Laboratory, Pasadena

Thompson, A.R., Moran, J.M., Swenson, G.W., Jr. (1986): *Interferometry and Synthesis in Radio Astronomy*, (J. Wiley & Sons, New York)

Treuhaft, R.N., Lowe, S.T. (1991): *A Measurement of Planetary Relativistic Deflection*, Astron. J. **102**, pp. 1879–88

Webber, J.C., Hinteregger, H.F. (1988): *Mark IIIA and VLBA High-Density Recording*, in (eds.) Reid and Moran (1988), pp. 501–502

Whitney, A.R. et al. (1976): *A Very-Long-Baseline Interferometer for Geodetic Applications*, Radio Science **11**, pp. 421–432

Whitney, A.R. (1988): *The Mark IIIA Correlator System*, in (eds.) Reid and Moran (1988), pp. 503–504

Wietfeldt, R.D. et al. (1992), *The Canadian S2 Recorder for Radioastron*, in (eds.) Hirabayashi et al. (1992)

Wohlleben, R., Mattes, H., Krichbaum, T. (1991): *Interferometry in Radioastronomy and Radar Techniques*, (Kluwer Academic Publishers, Dordrecht)

Zensus, J.A., Pearson, T.J. (eds.) (1989): *Parsec-Scale Radio Jets*, (Cambridge University Press, Cambridge)

A GRADIOMETER EXPERIMENT TO DETECT THE GRAVITOMAGNETIC FIELD OF THE EARTH

Dietmar S. Theiss

Institut für Theoretische Physik
Universität zu Köln
D-5000 Köln 41
Federal Republic of Germany

Abstract

We describe a space-borne experiment to detect the Lense-Thirring field produced by the proper rotation (mass-current) of the Earth. This gravitomagnetic field will generate an increasing signal in a gravity gradiometer orbiting the Earth in local inertial (gyroscope) orientation. For a polar orbit of $600\,km$ altitude, the signal will grow with a (constant) rate of about $2 \times 10^{-4} E$ per month ($1E = 1$ Eötvös $= 10^{-9} sec^{-2}$). In view of instrumental accuracies achieved in the last years, this effect could, in principle, be detected *at present* by Paik's high-sensitive superconducting gravity gradiometer in combination with precise gyroscopes placed in a drag-free Earth's satellite. A preliminary error analysis for the experiment indicates that the effect could already be measured after ≈ 1 *month* with sufficient accuracy (relative error of $\approx 1\%$). To achieve this, precise gyroscopes would be necessary, which, however, were allowed to be less precise than the *present* Stanford gyroscopes by a factor of $\simeq 20$. In addition, we present a method for *isolating* the gravitomagnetic signal from the dominant Newtonian background.

1. Introduction

It belongs to one of the fundamental consequences of Einstein's general relativity that the proper rotation of a mass will generate a gravitational "magnetic" field, which, for masses such as the Earth or the Sun, is similar to the magnetic dipole field of a rotating electric charge (Thirring 1918, Lense and Thirring 1918).[1] For slowly rotating masses (Earth, Sun), the Lense-Thirring field is expected to be extremely weak. As yet, it has *not* been detected.

The proper rotation of a mass gives rise to a dragging of local inertial frames. An experiment to measure this (mass-current) effect was initiated by Schiff, Fairbank, and Everitt[2,3] in the 1960's, the Stanford Gyroscope Experiment. This space experiment, now also known as Gravity Probe B (GPB), is expected to be performed in the next few years. The possibility of measuring the Lense-Thirring drag by analyzing the orbital motion of Earth's satellites has also been investigated. The most recent proposal in this direction is due to Ciufolini.[4]

The present paper is concerned with a further consequence of the proper rotation of a mass. It is the influence of the gravitomagnetic field of the mass on the relative (tidal) acceleration of neighboring test particles. Within the weak-field

approximation, this has first been discussed by Braginsky and Polnarev,[5] where the gravitomagnetic "forces" acting on two test masses at the ends of a spring (in an Earth's orbit) have been calculated with respect to the post-Newtonian coordinate frame. Using the invariant concept of Fermi frames, the gravitomagnetic contribution to the relative acceleration of test particles orbiting a rotating spherical body has been investigated by Mashhoon and Theiss.[6–9] (For more recent work on this subject see also References 10–15.) It was shown that this contribution will increase with time, and, for a circular inclined Earth's orbit of low altitude ($\simeq 600\,km$), it would become comparatively large already after about 1 *month*.

This secular (cumulative) relativistic effect could open a new possibility for detecting the Lense-Thirring field of the Earth by using already available technologies.

2. The Gravitomagnetic Effect in the Gravity Gradient and Its Origin

Consider a set of three spinning (orthogonal) test gyroscopes falling freely along an inclined circular geodesic orbit with (constant) radius r about a slowly rotating central body[16] (such as the Earth or the Sun) of mass M and proper angular momentum of magnitude J ($J \in \mathbf{R}$). The motion of the spin axes of these (torque-free) gyroscopes, which constitute a local inertial frame along the geodesic orbit, is governed by the equations of parallel transport (to a good approximation). By solving these equations[6–9] using the post-Schwarzschild approximation,[17] it can be shown that the motion of the gyroscope axes with respect to an effective Newtonian frame (sidereal frame)[18] consists of precessional motions superposed by a specific nodding up-and-down movement. The latter motion, a new relativistic effect of purely gravitomagnetic origin, has been referred to as *relativistic nutation*; see also References 13–15 and 19. The other, precessional, parts of the motion are dominated by the Fokker effect (geodetic precession). This (Fokker) effect has been recently detected in the motion of the Moon.[20]

Denote the proper time of the geodesic orbit by τ and choose one axis of the sidereal frame of reference so that, at the beginning of measurement ($\tau = 0$), it coinsides approximately with a vector normal to the orbital plane. Relativistic nutation is a periodic variation of the angle between this axis of the sidereal frame and a gyroscope axis. The leading contribution Θ_n of relativistic nutation to this angle can be written as

$$\Theta_n \simeq \xi \sin(\alpha)\,[\sin(\omega_F \tau + \eta_0) - \sin\eta_0]\ , \tag{1}$$

where $\xi = J/M\omega r^2$ and the constant η_0 denotes the azimuthal position of a test particle (satellite) in the plane of the geodesic orbit at $\tau = 0$ measured from the line of the ascending nodes. Here $\omega = \sqrt{GM/r^3}$ approximately describes the orbital frequency in the absence of rotation ($J = 0$), and α denotes the inclination of the orbit to the equatorial plane of the central body; the gravitational constant is denoted by G. The frequency of this nutational oscillation is equal to the Fokker

frequency

$$\omega_F = \left[\left(1 - 3\frac{GM}{c^2 r}\right)^{-\frac{1}{2}} - 1\right]\omega \simeq \frac{3}{2}\frac{GM}{c^2 r}\omega\,. \tag{2}$$

The nutation amplitude $\xi \sin\alpha$, remarkably, does not depend on the velocity of light c. This fact can be traced back to the occurrence of a small divisor[6–9,19] involving the Fokker frequency ω_F. In the post-Newtonian limit of the post-Schwarzschild approximation,[17] Eq. (1) reduces to

$$\Theta_n \simeq \omega_n \tau\,, \tag{3}$$

which represents a *precessional* motion with frequency

$$\omega_n = \frac{3}{2}\frac{GJ}{c^2 r^3}\sin\alpha\cos\eta_0 \simeq \xi\sin(\alpha)\omega_F\cos\eta_0\,. \tag{4}$$

(Eq. (3) simply follows from the expansion of Eq. (1) to second order in $1/c$.) The comparison of Eq. (1) with Eq. (3) shows that the post-Newtonian approximation *breaks down* over time scales of the order of the Fokker period $\tau_F = 2\pi/\omega_F$,[9,19] since, e.g., after $\tau = \left(\frac{\pi}{2} - \eta_0\right)/\omega_F$, the angle Θ_n of relativistic nutation decreases according to the (more precise) expression (1), whereas the corresponding post-Newtonian result, Eq. (3), shows a further *increase* of Θ_n; see also References 14 and 15. However, along the orbit of a satellite about the Earth, the Fokker period amounts to $\approx 10^5$ years; for the orbit of the Earth about the Sun it amounts to $\simeq 67$ million years. The post-Newtonian approximation is, therefore, sufficient to describe the observations in the solar system. From Equations (3) and (4) it follows that, in the post-Newtonian limit, relativistic nutation shows up as a specific (non-vanishing) part of Schiff precession,[21] where the magnitude of this nutation part is given by $\omega_n\tau$.

Let us now consider the influence of the Lense-Thirring field of the central body on the relative (tidal) acceleration of two nearby test particles T_1 and T_2 and the origin of this relativistic effect. Assume that T_1 moves along the circular geodesic orbit (reference orbit). With respect to a local inertial (orthonormal) frame $\{\lambda_i\} = \{(\lambda_i{}^\mu)\}$,[22] represented by the spin axes of the three orthogonal test gyroscopes, the tidal acceleration, i.e., the acceleration of T_2 as measured from T_1, is given by

$$\frac{d^2 x_i}{d\tau^2} = -K_{ij}x^j\,, \tag{5}$$

where the K_{ij} and the x_i denote the elements of the tidal matrix and the relative (Fermi) coordinates of T_2, respectively. The elements K_{ij} are defined by $R_{\mu\nu\varrho\sigma}\lambda^\mu\lambda_i{}^\nu\lambda^\varrho\lambda_j{}^\sigma$ with the (covariant) components $R_{\mu\nu\varrho\sigma}$ of the spacetime curvature tensor and the components λ^μ of the tangent vector of the geodesic orbit. Along the orbit, the gravity gradient Γ of the gravitational field of the central body in the direction of particle T_2 can be described by the tidal acceleration per

separation $x = \sqrt{x_i x^i}$ of T_1 and T_2 projected to the unit three-vector $\mathbf{n} = \mathbf{x}/x$ ($\mathbf{x} = \left(x^1, x^2, x^3\right)^{\mathrm{T}}$), i.e.,[8,7]

$$\Gamma = -K_{ij} n^i n^j \,, \tag{6}$$

where the $n^i = x^i/x$ are the direction cosines of T_2 measured with respect to the gyroscope axes ($\{\lambda_i\}$). In the following, we fix a specific *initial* orientation of the λ_i relative to the circular geodesic orbit (of T_1): At the beginning of measurement ($\tau = 0$) let λ_1 point approximately in radial direction away from the central body, λ_2 in a direction opposite to the orbital angular momentum vector of particle T_1 about the central mass, and λ_3 in the direction of motion of T_1 along the geodesic orbit. By such an orientation of the λ_i one can achieve that the angle Θ_n of relativistic nutation essentially *vanishes* for each gyroscope axis at $\tau = 0$. This is of importance in connection with the measurement of the Lense-Thirring contribution to the gravity gradient, since the leading terms of this contribution (perturbation) then vanish initially (see below); for details cf. References 7, 9, 14, and 19. In what follows, we always refer to this specific orientation of the local inertial frame of reference. According to the post-Schwarzschild approximation,[17] the contribution of the Lense-Thirring field of the central body to the gravity gradient Γ contains the expression

$$\left| 6\omega^2 \xi \sin\alpha \sin\left(\frac{1}{2}\omega_F \tau\right) \right| \tag{7}$$

as the leading amplitude. It results from the *off-diagonal*, purely gravitomagnetic elements K_{12} (radial-normal at $\tau = 0$) and K_{23} (normal-tangential at $\tau = 0$) of the tidal matrix; see References 6–9, 14, 15, and 19. This amplitude (7) is directly proportional to the amplitude of relativistic nutation ($\xi \sin\alpha$) and, therefore, shows a maximum (at $\tau = \tau_F/2$) which is independent of the speed of light c. In the post-Newtonian limit of the post-Schwarzschild approximation [17] ($\omega_F \tau \ll 1$), the amplitude (7) becomes *secular* and directly proportional to τ; one has

$$3\omega^2 \xi \sin(\alpha) \omega_F \tau \,. \tag{8}$$

Here $\omega_F = 3GM\omega/2c^2 r$ (to second order in $1/c$); cf Eq. (2). (The result (8) is in agreement with the post-Newtonian calculations performed more recently in References 11 and 12.) From the comparison of Equations (1)–(4) with (7) and (8), it becomes obvious that this secular amplitude originates from a coupling of the *nutation part* of Schiff precession with the amplitude ($\simeq 3\omega^2$) of the Newtonian contribution to the gravity gradient; see especially References 9, 14, and 15. In this connection it is also important to note that ω_n and the amplitude (8) vanish for an equatorial orbit ($\alpha = 0$) in contrast to the frequency of the *full* Schiff precession. As was mentioned above, for time scales much shorter than the Fokker period, the post-Newtonian approximation is completely adequate to describe relativistic effects of gravitational sources such as the Earth or the Sun. In the following, we, therefore, restrict our considerations, concerning the detection of the Lense-Thirring field of the Earth, to the post-Newtonian limit ($\omega_F \tau \ll 1$) of the post-Schwarzschild approximation.

3. The Experiment

Let us now examine the possibility of detecting the Lense-Thirring field of the Earth by measuring the gravity gradient along the orbit of an artificial satellite. For this purpose, we proceed with a more specific description of the relativistic effect in question.

It follows from the calculation of the (symmetric) tidal matrix $K = (K_{ij})$ (for the linearized Kerr metric[16]) that the leading contribution Γ^{LT} of the Lense-Thirring field of the central body to the gravity gradient Γ along the circular geodesic orbit has the form

$$\Gamma^{\mathrm{LT}} = -2\left(K_{12}n^1n^2 + K_{23}n^2n^3\right) , \tag{9}$$

and can be written explicitely[7,8] as

$$\Gamma^{\mathrm{LT}} = \mathcal{A}(\tau)\left[a_0 + a_1 \sin 2\omega\tau + a_2 \cos 2\omega\tau\right] , \tag{10}$$

where the (increasing) amplitude $\mathcal{A}(\tau)$ is given by[23]

$$\mathcal{A}(\tau) = 3\omega^2\xi \sin(\alpha)\omega_F\tau \equiv \frac{9}{2}\sin(\alpha)\frac{G^2JM}{c^2r^6}\tau , \tag{11}$$

cf. Eq. (8). The (constant) coefficients a_0, a_1, and a_2 are of order unity and depend on the angle η_0 and on the direction cosines n^i (cf. Eq. (6)); see References 7 and 8. An instrument that allows the detection of very small (relative) accelerations is beeing developed at the University of Maryland by H.J. Paik. It is designed for a space mission to measure the gradient of the gravitational field of the Earth along a satellite orbit of low altitude with extremely high accuracy. These measurements would enable, e.g., a very precise determination of the multipole moments of the Earth and would thereby improve our knowledge of the structure of the Earth. Paik's superconducting gradiometer allows the measurement of gravity gradients in three orthogonal directions, simultaneously. Each direction is realized by a linear channel (arm) along which a proof mass (particle T_2; cf. Eq. (6)) is constrained to move. The proof masses are attached to mechanical springs, where the tidal forces acting on the masses in the direction of the corresponding gradiometer arms are directly measured through suitable compensating forces. The gravity gradient in the direction of each gradiometer axis can then be determined by the compensating forces and the equilibrium separation of the proof masses from the origin (particle T_1) of the orthogonal frame formed by the three gradiometer arms; for details see, e.g., Reference 24 and also 10. The intrinsic noise level for Paik's superconducting gravity gradiometer has now been reduced to $\approx 10^{-3}EHz^{-1/2}$, corresponding to $\approx 10^{-6}E$ for an integration time of 1 month.[25] ($1E = 1$ Eötvös $= 10^{-9}sec^{-2}$.) It follows from Eq. (11) that, for a circular polar orbit ($\alpha = \pi/2$) of 600 km altitude, the amplitude $\mathcal{A}(\tau)$ of the gravitomagnetic contribution Γ^{LT} of the Earth to the gravity gradient will increase to $\simeq 2 \times 10^{-4}E$ after 1 month;[7,8] cf. also References 6, 9, and 11–13. Thus, this relativistic effect could be detected, in principle, by

means of Paik's *present* gradiometer[25] already after $\approx 1\,month$ with a signal-to-noise ration of *100*.

The experiment to detect the gravitomagnetic tidal effect, essentially requires a high-sensitive gravity gradiometer in combination with precise gyroscopes both placed in a drag-free satellite orbiting the Earth. A drag-free satellite will be necessary to guarantee that, at low altitude above the Earth, the system is falling (almost) freely along the orbit (geodesic motion). To measure the secular effect described by Equations (10) and (11), the orientation of the gradiometer axes has to be kept *constant* relative to the spin axes of the gyroscopes (local inertial frame) during the experiment. In other words, the gradiometer arms will be guided *inertially* along the orbit.

A modified experiment, in which the gyroscopes are replaced by telescopes, has also been discussed.[10] In that experiment the orientation of the gradiometer axes would be kept constant relative to the *sidereal* frame.[18] The results in Ref. 10 show that, along a polar circular orbit of low altitude, the contribution of the Lense-Thirring field of the Earth to the gravity gradient (measured in sidereal orientation) will be *periodic* with frequency 2ω and will show a (constant) amplitude of $\approx 10^{-7}E$. In the present paper, we will not consider this version of the experiment in further detail. We, instead, refer the reader to the analyses provided in Ref. 10.

The entire gradient of the gravitational field of the Earth is greatly dominated by the Newtonian background, which possesses an amplitude of order ω^2 corresponding to $\approx 10^3 E$, for an orbit of low altitude. It would therefore be desirable to find a method to *isolate* the gravitomagnetic contribution, described by Eq. (10), from these dominant parts. For this purpose, let us consider a specific (constant) orientation of the orthogonal gradiometer axes relative to the (parallel transported) spin axes of the gyroscopes, representing the co-moving, locally inertial frame of reference, $\{\lambda_i\}$. Imagine, e.g., two gradiometer axes, A_1 and A_2, lying in the plane spanned by the orthogonal unit vectors λ_2 and λ_3. (The initial orientation of the spacelike orthonormal frame $\{\lambda_i\}$ relative to the geodesic orbit was described above.) Furthermore, assume that each of these gradiometer arms, A_1 and A_2, forms an angle of 45° with the λ_2-axis; i.e., one arm, say A_1, points in the direction of $(\lambda_2 + \lambda_3)$, the other, A_2, in the direction of $(\lambda_2 - \lambda_3)$. (The remaining arm, of course, then points in the direction of λ_1, or $-\lambda_1$.) Let $\mathbf{n} = (n^i)$ and $\tilde{\mathbf{n}} = (\tilde{n}^i)$ be (orthogonal) unit 3-vectors pointing in the $(\lambda_2 + \lambda_3)$-direction (A_1) and in the $(\lambda_2 - \lambda_3)$-direction (A_2), respectively. The (constant) components[22] n^i and $\tilde{n}^i$ are the direction cosines of $\mathbf{n}$ and $\tilde{\mathbf{n}}$ measured with respect to the basis $\{\lambda_i\}$, cf. Eq. (6), and can be written as

$$n^1 = 0\,, \qquad n^2 = \frac{1}{\sqrt{2}}\,, \qquad n^3 = n^2\,, \tag{12}$$

$$\tilde{n}^1 = 0\,, \qquad \tilde{n}^2 = n^2\,, \qquad \tilde{n}^3 = -n^2\,. \tag{13}$$

From the general expression, Eq. (6), it now follows that the gravity gradients Γ — in the direction of $\mathbf{n}$ (axis A_1) — and $\tilde{\Gamma}$ — in the direction of $\tilde{\mathbf{n}}$ (axis A_2)— take

the simple form

$$\Gamma = -\frac{1}{2}(K_{22} + K_{33} + 2K_{23}) , \tag{14}$$

$$\tilde{\Gamma} = -\frac{1}{2}(K_{22} + K_{33} - 2K_{23}) . \tag{15}$$

By *subtracting* these two gradiometer "signals" one gets[26]

$$\hat{\Gamma} \equiv \tilde{\Gamma} - \Gamma = 2K_{23} . \tag{16}$$

(Quite similar results are obtained when rotating the arms A_1 and A_2 about the λ_1-axis by 90°, 180°, or 270°.) If we place the gradiometer arms A_1 and A_2 (with 45°-orientation) in the plane spanned by λ_1 and λ_2, the corresponding signal difference $\hat{\Gamma}$ will be given by $2K_{12}$, and in the (λ_1, λ_3)-plane by $2K_{13}$. Thus, a direct measurement of the *off-diagonal* elements of the tidal matrix K will be possible by applying such a signal-subtraction method. As was mentioned above, the matrix elements K_{12} and K_{23} are directly proportional to the proper angular momentum J of the central body (Earth) and are, thereby, of purely *gravitomagnetic* origin; cf. Eq. (9). The explicit evaluation of Eq. (16) yields

$$\hat{\Gamma} \simeq \mathcal{A}(\tau)\,[\sin\eta_0 - \sin(2\omega\tau + \eta_0)] , \tag{17}$$

where the increasing amplitude $\mathcal{A}(\tau)$ is given by Eq. (11). (The right-hand side of Eq. (17) is equal to the leading contribution to $\hat{\Gamma}$, which is secular.) When placing the gradiometer axes A_1 and A_2 correspondingly in the (λ_1, λ_2)-plane, the isolated gravitomagnetic signal $\hat{\Gamma}$ is given by $2K_{12}$ ($\simeq 2K_{23}\cot\omega\tau$).

Since the Earth is not completely spherically symmetric, the quadrupole moment (oblateness) and higher moments will contribute to the gravity gradient. For instance, the quadrupole moment will also affect the off-diagonal elements of the tidal matrix; see, e.g., Ref. 14. Preliminary investigations indicate that these off-diagonal contributions of the quadrupole moment will vanish along a *polar* Earth's orbit ($\alpha = \frac{\pi}{2}$), for which the gravitomagnetic effect just reaches its *maximum*; cf. Equations (11) and (17). More detailed studies of the influences of the central body's multipole moments on the gravity gradient will be left to future publications.

It follows from our preliminary error analysis for the gradiometer experiment in question that there is one major source of error. Namely, the deviation of the motion of the *gyroscopes'* spin axes from ideal parallel-transport along the orbit. The deviations are caused by small perturbing torques. Let us examine the influence of this *drift* of gyroscopes on the isolated gravitomagnetic signal, $\hat{\Gamma}$, more closely. For this purpose, we can restrict our considerations, without loss of generality, to the specific case of orientation of the gradiometer axes treated above (Equations (12)–(17)). A small drift of gyroscopes can be described by a time-dependent infinitesimal rotation of the local inertial frame of reference $\{\lambda_i\}$. The resulting slowly *rotating* frame $\{\lambda_i'\}$ is then given by[22]

$$\lambda_i' = D_i{}^j(\tau)\lambda_j , \tag{18}$$

where the elements D_{ij} $(= D_i{}^j)$ of the infinitesimal rotation matrix D can be expressed as $D_{ij} = \delta_{ij} + I_{ij}$ with $I_{ij} = -I_{ji}$ and $|I_{ij}| \ll 1$. In this rotating frame, the isolated gravitomagnetic signal, $\hat{\Gamma}'$, takes the form

$$\hat{\Gamma}' = 2\left[DKD^{\mathrm{T}}\right]_{23} = \hat{\Gamma} + \delta\hat{\Gamma}\,, \tag{19}$$

where the perturbation $\delta\hat{\Gamma}$ of $\hat{\Gamma}$, which is due to the drift of the gyroscopes, can be written as

$$\delta\hat{\Gamma} = -2\left[I_{12}K^N_{13} + I_{23}\left(K^N_{22} - K^N_{33}\right)\right]\,. \tag{20}$$

(Here it is sufficient to consider only that part of $\delta\hat{\Gamma}$, which results from the leading, Newtonian, contribution[27] K^{N} to the tidal matrix K.) If we write the relative error of $\hat{\Gamma}$ as

$$\frac{\delta\hat{\Gamma}}{\hat{\Gamma}} = \epsilon \tag{21}$$

($|\epsilon| \ll 1$), we can determine the (infinitesimal) quantities I_{12} and I_{23} from Eq. (20) by means of Equations (11) and (17). One has

$$I_{12} = -\frac{3}{2}\epsilon\frac{GJ}{c^2r^3}\tau\sin\alpha\cos\eta_0\,, \tag{22}$$

$$I_{23} = \epsilon I_{12}\tan\eta_0\,. \tag{23}$$

Hence, the permissible *drift-rate* of gyroscopes, at a given value of ϵ, can be described by the (constant) frequency [28]

$$\delta\Omega = \sqrt{\left(\frac{dI_{12}}{d\tau}\right)^2 + \left(\frac{dI_{23}}{d\tau}\right)^2} = \frac{3}{2}|\epsilon|\frac{G|J|}{c^2r^3}\sin\alpha\,. \tag{24}$$

Thus, it follows that, if we tolorate an error ($|\delta\hat{\Gamma}/\hat{\Gamma}|$) of 1%, i.e., $|\epsilon| = 1/100$, a drift-rate $\delta\Omega$ of $\simeq 1.3$ *milliarcsec/yr* would be allowed, for a polar circular Earth's orbit of 600 *km* altitude; see also Ref. 13. (Such a circular orbit will be chosen for the planned Stanford Gyroscope Experiment (GPB).) It should be noted that the drift-rate of the *present* Stanford gyroscopes amounts to $\simeq 0.06$ *milliarcsec/yr* (for a $10^{-11}g$ drag-free satellite)[29] and is, thereby, already *smaller* than the tolerable value for $\delta\Omega$ (at $|\delta\hat{\Gamma}/\hat{\Gamma}| = 1\%$) by a factor of $\simeq$ *20.*

The treatment of additional errors, e.g., the influence of small deviations of the satellite from ideal geodesic motion (free fall) on the gravitomagnetic signal $\hat{\Gamma}$, will be the subject of further publications.[30]

In conclusion, the gradiometer experiment described in the present paper could open the possibility of detecting the Lense-Thirring field of the Earth with sufficient accuracy after only a *few weeks* by already *existing* instruments.

Acknowledgments

I would like to thank Prof. F.W. Hehl for his support.

References and Notes

[1] H. Thirring, Phys. Z. **19**, 33 (1918); J. Lense and H. Thirring, Phys. Z. **19**, 156 (1918). See also B. Mashhoon, F.W. Hehl and D.S. Theiss, Gen. Rel. Grav. **16**, 711 (1984).

[2] L.I. Schiff, Phys. Rev. Lett. **4**, 215 (1960); Proc. Nat. Acad. Sci. U.S. **46**, 871 (1960).

[3] C.W.F. Everitt *et al.*, in *Near Zero: Festschrift for William M. Fairbank*, edited by C.W.F. Everitt (Freeman, San Francisco, 1986).

[4] I.Ciufolini, Phys. Rev. Lett. **56**, 278 (1986).

[5] V.B. Braginsky and A.G. Polnarev, Pis'ma Zh. Eksp. Teor. Fiz. **31**, 444 (1980) [JETP Lett. **31**, 415 (1980)].

[6] B. Mashhoon and D.S. Theiss, Phys. Rev. Lett. **49**, 1542 (1982).

[7] D.S. Theiss, Ph.D thesis, University of Cologne (Köln, 1984).

[8] D.S. Theiss, Phys. Lett. A **109**, 19 (1985).

[9] B. Mashhoon, Found. Phys. **15** (Bergmann Festschrift), 497 (1985).

[10] H.J. Paik, B. Mashhoon, and C.M. Will, in *Experimental Gravitational Physics*, edited by P.F. Michelson, Hu En-ke, and G. Pizella (World Scientific, Singapore, 1988), p.229; B. Mashhoon, H.J. Paik, and C.M. Will, Phys. Rev. D **39**, 2825 (1989). See also H.J. Paik, Adv. Space Res. **9**, 41 (1989).

[11] E. Gill, J. Schastok, M.H. Soffel, and H. Ruder, Phys. Rev. D **39**, 2441 (1989).

[12] C.A. Blockley and G.E. Stedman, Phys. Lett. A **147**, 161 (1990).

[13] D.S. Theiss, in *Proc. First William Fairbank Meeting on Relativistic Gravitational Experiments in Space* (Rome, 1990), to appear in the *Advanced Series in Astrophysics and Cosmology*, edited by L.Z. Fang and R. Ruffini (World Scientific, Singapore).

[14] B. Mashhoon and D.S. Theiss, Nuovo Cimento B **106**, 545 (1991).

[15] D.S. Theiss and B. Mashhoon, *Comment on "On the Mashhoon-Theiss 'Anomaly' "*, preprint.

[16] The exterior gravitational field of a slowly rotating spherical body can be described by the Kerr metric linearized in the angular momentum parameter $a = J/M$; see References 6–9, 14, and 19.

[17] In the post-Schwarzschild approximation, deviations of the gravitational field from spherical symmetry — caused, e.g., by the proper rotation (cf. Ref. 16)) or the oblateness of the body — are considered to first order, whereas the *mass* of the central body is taken into account to *all orders*. To the appropriate order in $1/c$ (c velocity of light) this scheme reduces to the standard post-Newtonian approximation.

[18] This (sidereal) frame refers to distant stars.

[19] B. Mashhoon and D.S. Theiss, Phys. Lett. A **115**, 333 (1986).

[20] I.I. Shapiro, R.D. Reasenberg, J.F. Chandler, and R.W. Babcock, Phys. Rev. Lett. **61**, 2643 (1988). See also B. Bertotti, I. Ciufolini, and P.L. Bender, Phys. Rev. Lett. **58**, 1062 (1987).

[21] Schiff precession is a gravitomagnetic effect, caused by the proper rotation of a central mass and describes the dragging of local inertial frames (relative to the sidereal frame) in the post-Newtonian (weak-field) approximation; cf. Ref. 2.

[22] Greek indices run from 0 to 3; Latin indices run from 1 to 3. Summation convention is used throughout.

[23] In the post-Schwarzschild approximation, Γ^{LT} is of the same form as Eq. (10), where $\mathcal{A}(\tau)$ is given by the expression (7); cf. References 7 and 8.

[24] M.V. Moody, H.A. Chan, and H.J. Paik, J. Appl. Phys. **60**, 4308 (1986); H.A. Chan and H.J. Paik, Phys. Rev. D **35**, 3551 (1987); H.A. Chan, M.V. Moody, and H.J. Paik, Phys. Rev. D **35**, 3572 (1987).

[25] See, e.g., M.V. Moody and H.J. Paik, in *Relativistic Gravitational Experiments in Space*, proceedings of a workshop sponsored by the National Aeronautics and Space Administration, edited by R.W. Hellings (Scientific and Technical Information Division, NASA Conference Publication 3046, Washington, D.C., 1989), p. 211.

[26] This signal subtraction method is similar to that described in Ref. 10; see also Ref. 13.

[27] This Newtonian contribution is given by $K^N_{11} = \omega^2(1 - 3\cos^2\omega\tau)$, $K^N_{22} = \omega^2$, $K^N_{33} = -\left(K^N_{11} + K^N_{22}\right)$, $K^N_{13} = -\frac{3}{2}\omega^2 \sin 2\omega\tau$, $K^N_{12} = K^N_{23} = 0$.

[28] The components[22] $\delta\Omega^i$ of the frequency vector describing the gyroscopes' drift are given by the equations $-\left[(dD^{\mathrm{T}}/d\tau)D\right]_{ij} \equiv dI_{ij}/d\tau = \epsilon_{ijk}\delta\Omega^k$; $\delta\Omega \equiv \sqrt{\delta\Omega_i \delta\Omega^i}$.

[29] See, e.g., C.W.F. Everitt, B.W. Parkinson, and J.P Turneau, in *Relativistic Gravitational Experiments in Space*, proceedings of a workshop sponsored by the National Aeronautics and Space Administration, edited by R.W. Hellings (Scientific and Technical Information Division, NASA Conference Publication 3046, Washington, D.C., 1989), p. 118.

[30] There are, however, several errors that can be treated similarly as in Ref. 10 (sidereal orientation of gradiometer axes).

The International Atomic Time and the PTB's Clocks

Dr. K. Dorenwendt, Physikalisch-Technische Bundesanstalt, Braunschweig

1. Time Scales

A time scale serves the purpose of giving a date to events by associating numerical values to them. It is characterized by an arbitrarily fixed origin and a scale unit which - in the cases of interest to us - is the second of the International System (SI) of Units.

To realize such scales, time depending physical phenomena are used whose evolution in time can be followed on the basis of measurements and whose laws are known. Archaeologists, for example, establish dates with reference to the radioactive decay, whereas periodical processes are preferred in the field of technology. These were formerly the processes of celestial mechanics which were considered particularly regular; we know today that atomic processes are superior in this respect.

1.1 Astronomical Time Scales

1.1.1 Universal Time UT and Zonal Time

The time in normal life is determined by the revolution of moon and earth round the sun. Days, months and years follow one another to form a natural time scale. From ancient times the astronomers have therefore been responsible for the determination of the time. However, if one looks more closely at the time intervals obtained by observing the stars, it can be seen that their length varies and that they are therefore only conditionally suitable for the establishment of a time scale.

If, for example, the time of 12 o'clock is associated to the moment when the sun is at its zenith as it is observed every day from a fixed point on earth, so-called true solar days are obtained whose lengths deviate from the mean length of a day by up to $\pm$ 30 s in the course of a year. In a time scale based on the true sun, these differences in the length of individual days add up to seasonal variations of $\pm$ 15 min. These irregularities are due to the earth's elliptic orbit and to its axis of rotation being inclined to the plane of its ecliptic.

If the variations are averaged with the aid of the well-known relations, a mean solar day and a mean solar time is obtained. Until 1956, the second was defined as the 86 400th part of the mean solar day.

The mean solar time referred to the meridian of Greenwich is called Universal Time, UT, previously also Greenwich Mean Time. The so-called zonal times result from the Universal Time in that full hours are added or deducted. The zonal times are therefore identical with the mean solar times of certain selected meridians. The relevant Act of the German Reich of 1893, which was applicable until the 1978 Time Act entered into force, therefore stipulated the following: "The legal time is the mean solar time of the fifteenth meridian east of Greenwich."

1.1.2 Corrected Universal Time UT1 and UT2

More exact observations of the sky with its fixed stars show that the earth's axis of rotation varies inside the planet by some arc seconds. This so-called polar motion has a period of about 14 months, subject to irregular seasonal changes. Depending on where he is standing, the observing astronomer is given the deceptive impression of a relative change of the earth's rotational frequency ($\pm$ 10^{-8} in these latitudes). This influence on the Universal Time was taken into account in 1956, leading to the introduction of the corrected Universal Time, UT1. It is proportional to the earth's angle of rotation and is to date being used for navigation purposes.

In the years 1934/35, physicists of the Physikalisch-Technische Reichsanstalt succeeded in demonstrating seasonal variations in the earth's rotation with the aid of quartz clocks [1] (Fig. 1). This phenomenon is caused by periodical changes of the earth's moment of inertia (heating of the air above the continents). When these variations are allowed for by suitable corrections, a still better Universal Time, UT2, is obtained which is, however, hardly used any longer. It is the most uniform time scale which can be derived from the earth's rotation. The difference between UT1 and UT2 amounts to up to $\pm$ 30 ms in the course of a year.

In addition to the irregularities already referred to, there are other influences which have a disturbing effect on the earth's movement. Asymmetrical forces from the moon and the sun act on our oblate planet. The earth then behaves like a gyroscope. Its axis of rotation slowly rotates about the perpendicular to the

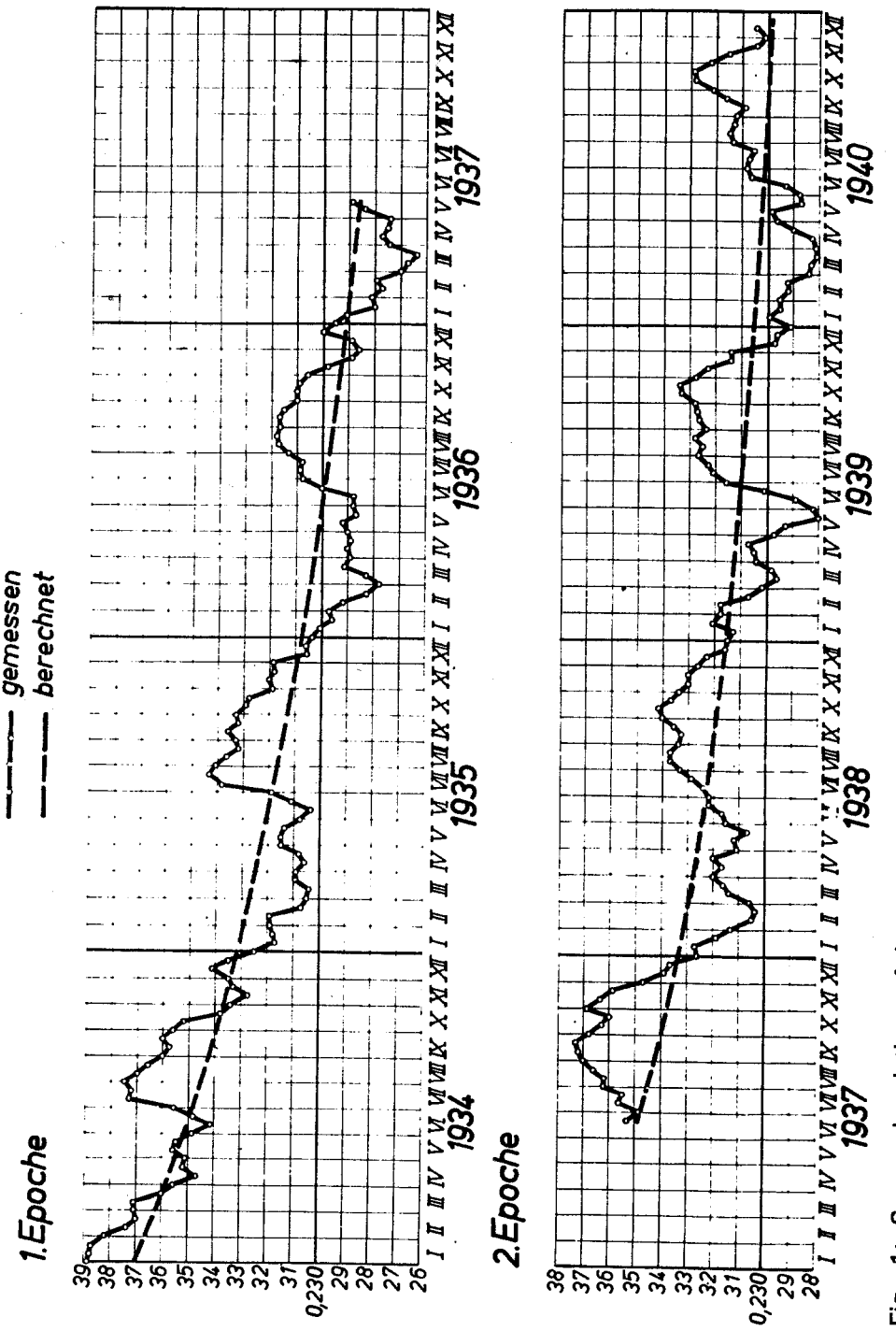

Fig. 1: Seasonal variations of the mean solar day in seconds per day, historical measurements with quartz clocks.

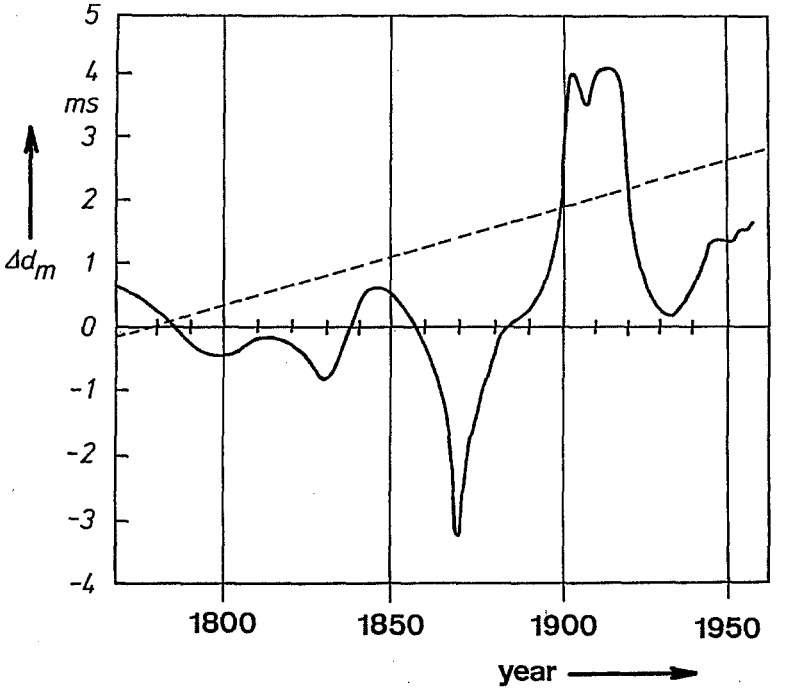

Fig. 2: Long term variations Δdm of the mean solar day determined with the aid of astronomical observations, the dashed line indicates the influence of tidal friction.

ecliptic. A full period of this so-called precession takes 26 000 years. The earth is also subject to smaller forces from the other planets, which change constantly. Combined with the effects of slow changes in the moment of inertia of our planet, long-term variations of the earth's rotational frequency result. The astronomers have observed this phenomenon for centuries by comparing the length of the mean solar days with other periodical processes in celestial mechanics (Fig. 2). Passages of planets and the revolution of our moon or of the Jovian moons were used for this purpose [2].

The large variations shown in Fig. 2 hide a slow but constant decrease in the earth's rotational frequency due to the friction of the tides. Coral fossils from Devon, on which daily and annual rings can be counted, have shown, however, that 400 million years ago the year had 400 days [3]. All this demonstrates that the Universal Time derived from the earth's rotational movement is not a uniform

time scale. Alone as a result of the tidal friction, the Universal Time and an ideal, uniform time scale drift apart by one hour in 1 000 years.

To become independent of the irregularities of the earth's rate of rotation, the Ephemeris Time, ET, with the ephemeris second as the fundamental unit was introduced in 1956. Here the determination of the time was deduced from the revolution of the earth round the sun and no longer from its daily rotation on its own axis. The ephemeris second was defined as a certain fraction of the tropical year 1900. In practice the ephemeris second proved to be as unsuitable as was the second derived from the mean solar day.

1.1.3 Pulsars

The question of whether pulsars can be used as time standards has recently been brought up for discussion [4]. In general, pulsars are isolated neutron stars which remain after the gravitational collapse of a star. Their periods of rotation range between 10^3 s and 10^{-3} s.

Radio radiation is released above the magnetic poles, by a mechanism which is still not fully understood. Pulses are observed because the magnetic dipole axis and the axis of rotation include a certain angle. The pulsar thus acts like the beacon light of a lighthouse, whose radiation sweeps us in the rhythm of its period of rotation (Fig. 3). However, a time scale derived from these pulses has the following characteristic features:

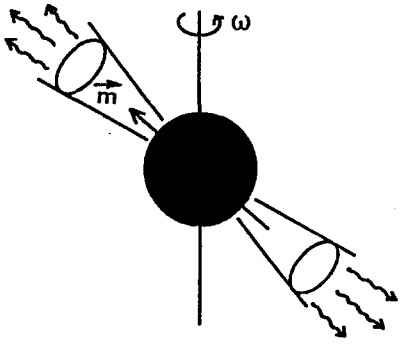

Fig. 3: The radiation of a pulsar is aligned with its magnetic dipole axis.

- The pulses are affected by noise. Even the ms-pulses of the neutron stars rotating at highest speed can be picked up only with an uncertainty of some 100 ns. This leads to very long averaging times when a clock on the earth is to be synchronized with their aid.

- The arrival times of the pulses forming the time markers of such a scale depend on the earth's position during its revolution round the sun. To reach an uncertainty of 100 ns, the earth's orbit must be known correct to 30 m. It has, however, been pointed out at the beginning that the earth's path is subject to complex influences from the planetary constellation.

- The pulse frequencies of individual neutron stars show individual drifts. This phenomenon has to do with the radiated power which, in the last analysis, is fed by a decrease of the rotational energy. In a time scale established on such a basis, the drift of the pulse frequency must be calibrated by means of a superior standard, i.e. an atomic clock.

1.2 Atomic Time Scales

Atomic time scales differ from astronomical time scales in an essential aspect. The latter are derived from a single time standard, an astronomical object whose signals are accessible to everybody. In contrast to this, atomic time scales are obtained by forming the average of a great number of clocks. It is true that here, too, a particular, exceptionally good clock might serve as a maserclock; however, this will probably never happen, be it only for political reasons.

1.2.1 Time Comparisons

International time comparisons in the ns range are a prerequisite for the calculation of a mean time scale on the basis of the readings of many clocks installed all over the world. Three procedures are essentially followed for this pupose:

- **Transportable atomic clock:** The stationary clocks of two institutes are successively placed side by side with a travelling clock and the intervals between the second pulses are measured by means of a time interval counter. Various stations are usually included in such comparisons. The uncertainty attained depends on the stability of the travelling clock and on the duration of

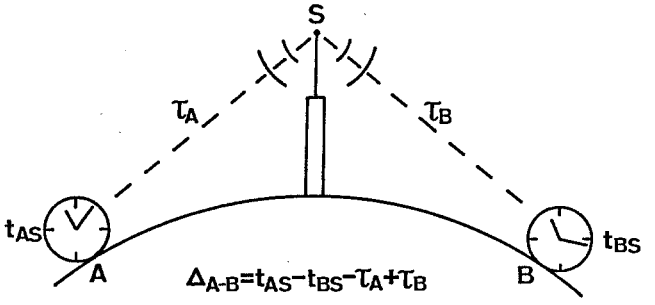

Fig. 4a: One-way time comparison: Δ_{A-B} time difference between the clock at station A and B; t_{AS}, t_{BS} arrival times of a selected signal of transmitter S; τ_A, τ_B propagation times.

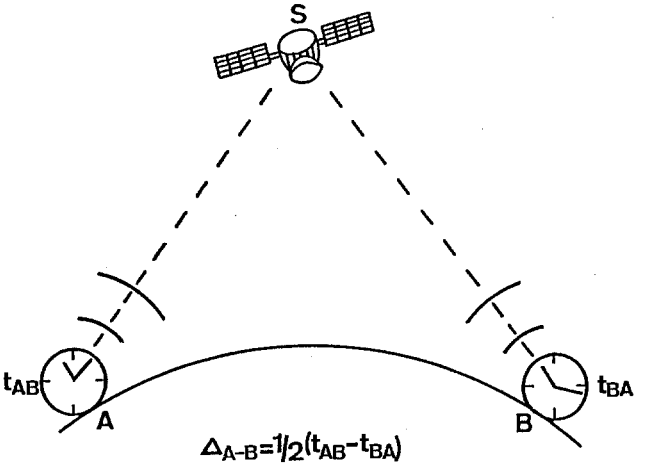

Fig. 4b: Two-way time comparison: t_{AB}, t_{BA} time intervals between transmitted and received second pulses.

the journey. It varies between 10 ns and some 100 ns. The relativistic time shifts of the travelling clock must be corrected; in the case of air journeys they may amount to some 100 ns.

- **One-way time comparisons:** The procedure is based on the reception of radio signals (Fig. 4a). Prerequisite is the exact knowledge of the distances, as the propagation times of the signals, τ_A and τ_B, enter into the result. The time difference of the clocks can be calculated from the arrival times t_{AS} and t_{BS} of a selected signal read on both clocks.

In the North American and European region, preference was given in the past to signals of the LORAN C navigation system installed on the earth. Uncertainties of about 100 ns were attained in the time comparisons, which were essentially due to disturbing interferences between the ground wave travelling on the earth's surface and the sky wave reflected by the ionosphere.
Today better results are obtained using signals from satellites in the case of which such interferences cannot occur. As the locations of satellites can be determined with sufficient accuracy, the propagation times can be calculated as well. Uncertainties of about 20 ns are at present being attained with the aid of the GPS satellites [5]. This progress is also due to an improved signalling technique which makes use of pseudorandom sequences and autocorrelation techniques.

- **Two-way time comparisons:** Via a satellite which acts like a mirror, each of the stations transmits its second pulses to the other station (Fig. 4b). On each side, the time interval between the second transmitted by the respective station and the second pulse received from the other station is measured. When the time difference between the clocks is calculated, the propagation times, which are the same in both directions, cancel out. Strictly speaking, this is, however, only true of geometrical paths. In the electronics of the stations involved different delay times of the signals transmitted and received may occur, which must be determined separately. In spite of this, uncertainties of 1 ns are achieved by this method.

1.2.2 The International Atomic Time, TAI

The International Atomic Time is defined as follows: Its origin corresponds with January 1, 1958; 0.00 Universal Time. The SI second realized at sea level has been fixed as the scale unit. TAI therefore is a coordinate time scale on our geoid in rotation.

Approximately 200 atomic clocks in about 20 time institutes contribute at present to the formation of TAI. At ten-day intervals, the time differences of the clocks are measured using the procedures referred to before, and the results communicated to the Bureau International des Poids et Mesures (BIPM), Paris. At the BIPM, TAI is calculated by a method which has already been changed several times. Approximately two months later each participant in these clock comparisons is informed

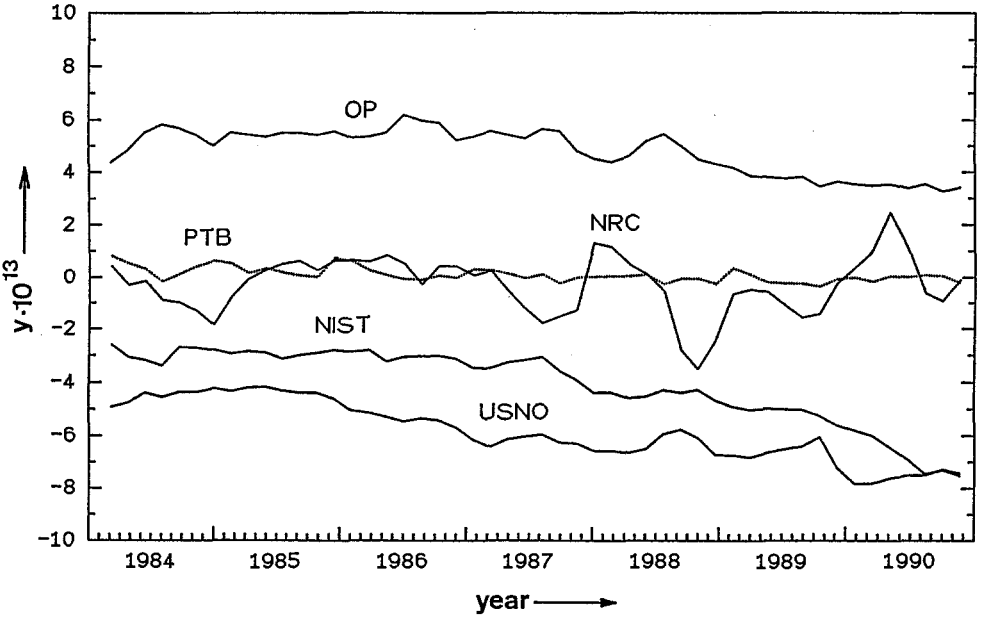

Fig. 5: Relative frequency difference y between TAI and independent atomic time scales: OP (Observatoire de Paris, France); PTB (Phys.-Techn. Bundesanstalt, Germany); NRC (National Research Council, Canada); NIST (National Institute of Standards and Technology, USA); USNO (United States Naval Observatory, USA).

to which extent his clocks differ from TAI. He can then subsequently convert dates established in the time scale of one of his clocks to the more exact TAI.

For the calculation of TAI, different weight is given to the clocks involved. This weight is independent of the clock's frequency difference with respect to TAI; it solely depends on the clock's stability observed in the past year in comparison with TAI. The time scale calculated in this way therefore excels by a particular stability. However, it has not yet the correct scale unit, the SI second, as it is mainly based on commercial Cs clocks which are affected by systematic errors. Comparisons with the primary clocks (see below) of great metrological institutes show that the second of the so-called free atomic time scale obtained from the averaging process is too great by more than 10^{-12}. Only after a suitable steering is TAI obtained from the free atomic time scale.

In this context the PTB's clocks are of special importance. In these primary clocks, all frequency-shifting effects are taken into account by appropriate corrections so that the second, which is finally realized, has the defined length up to a very small limit of uncertainty [6]. Details will be given in the next section. There are only a few such primary clocks in the world and not all of them are operated continually. For approximately four years, the BIPM has even been solely dependent on the results furnished by the PTB's clocks, CS1 and CS2.

Fig. 5 is intended to give an impression of the long term stability of atomic time scales. The differences between TAI and the independent time scales of some institutes reflect, that usually commercial clocks with uncontrolled systematic frequency departures are involved. The good agreement of the PTB time with TAI demonstrates the great influence of the PTB clocks in the steering of TAI.

2. Primary Atomic Clocks

Since 1967 the second of the International System of Units has been defined as follows:

> The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom.

The definition does not refer to a certain velocity or a certain gravitational potential.As a consequence, depending on the conditions under which it is realized, the SI second appears to differ in length. This was taken into consideration when the International Atomic Time, TAI, was introduced which has been defined as a coordinate time scale on our geoid.

The definition of the second is put into practice by means of so-called Cs atomic clocks. As these clocks are subject to a number of influences which affect their rate or frequency, a distinction is made between the commercial standards where the amount of these influences is not known, and the primary clocks where the frequency errors are determined by measurement of parameters and then corrected, and where the limit of residual errors is given by an uncertainty estimation. Such primary clocks - in particular those of the PTB - will be dealt with in the following.

2.1 Principle of a Cs Clock

In order to show the limits of a time standard, the principle of a Cs clock (Fig. 6) shall be explained first.

Cesium is heated in an oven O which the atoms leave as divergent beam through a small opening. The solid dots and open dots mark atoms at the two energy levels referred to in the definition of the second. As the levels are close together, they are equally populated.

The atoms first reach the state selecting magnet M1 whose inhomogeneous field influences their path. Atoms at the upper energy level have a negative magnetic moment. They are deflected towards the axis and continue on their path as parallel beam when the velocity is appropriate. Atoms at the lower energy level with positive magnetic moment are expelled from the beam.

Inside the cavity our atoms cross an electromagnetic field whose frequency can be tuned with the help of the voltage controlled oscillator VCO. If this frequency corresponds to the transition frequency f_0 of the cesium, the atoms are stimulated

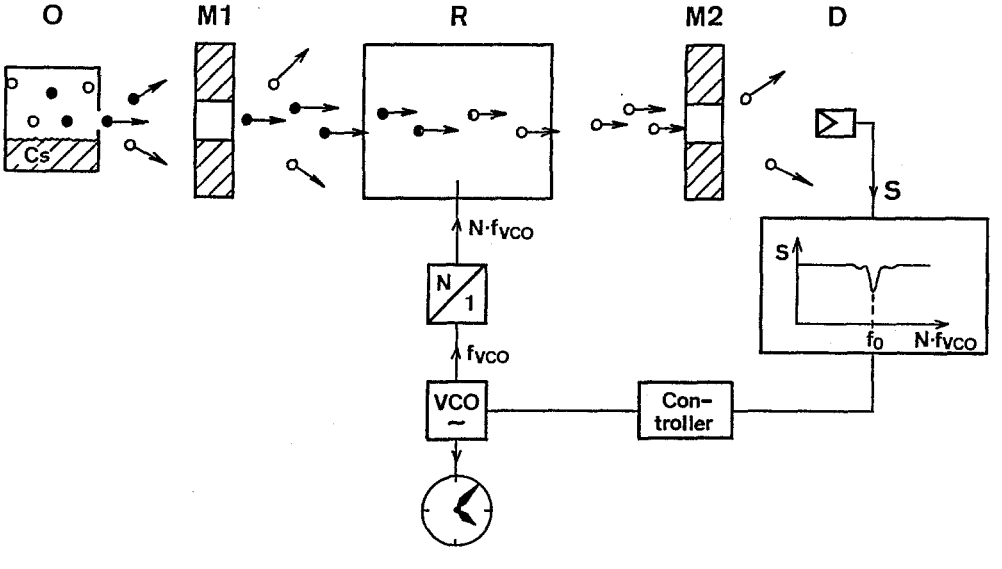

Fig. 6: Principle of a Cs clock: O oven; M1, M2 state selecting magnets; R microwave cavity; D detector; VCO voltage controlled oscillator.

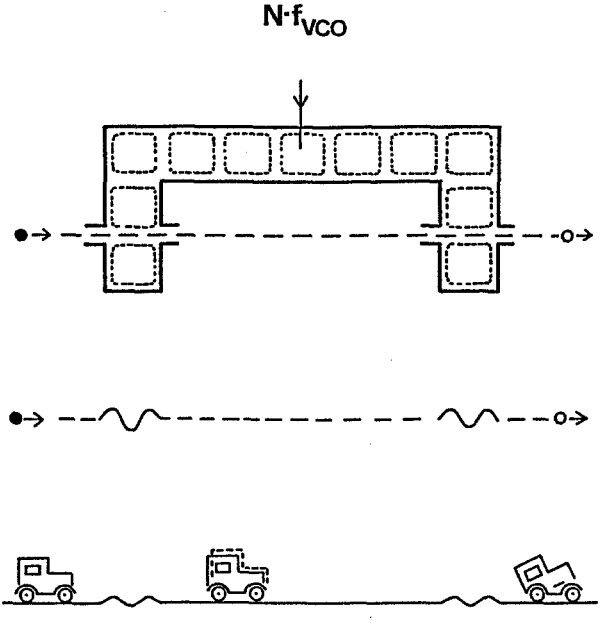

Fig. 7: Ramsey cavity or separated field exposure: the dashed line represent magnetic field lines of the standing wave pattern.

to change their energy level. This is indicated by the variation from solid to open dots inside the cavity. The atoms then reach the state selecting magnet M2. Having made the transition, they are defocussed, thus making the detector signal S a minimum. S can therefore be used to tie the frequency fed into the cavity to the atomic resonance f_o.

For clarity, closer details of the microwave cavity must be given. It is in fact possible to expose the atoms to a single extended radiation field according to Fig. 6; however, different oscillation modes or phase curvatures as they occur in an extended cavity would disturb the clock's function. A more ideal field distribution can be realized inside a waveguide. Norman F. Ramsey developped the method of the "separated field exposure" (Nobel Prize for physics in 1989) which makes use of this advantage. Fig. 7 shows the Ramsey resonator, a U-shaped waveguide, with its standing wave field in the interior. The atoms which traverse the walls of the waveguide through bores some millimeters in diameter are exposed twice to the field.

The effect of this separated exposure can be demonstrated by a picture which Ramsey himself used. Instead of an atom excited by a microwave, we consider a car on a corrugated road. The first disturbance shakes the car which will continue to oscillate at its eigenfrequency on the even road section. The second disturbance will further intensify this oscillation if its periodic excitation occurs in phase with the car's oscillations. In this case the car can be seriously damaged or - when we return to the picture of the atoms - they can change their energy level under stimulated emission or absorption. Obviously the transition probability will reach a maximum if the microwave frequency agrees with the atoms' eigenfrequency and if both excitations are in phase. Additional secondary maxima occur if the frequency of the cavity field does not correspond to the atoms' eigenfrequency but is just so high that between the two exposures, the atoms can perform an integer number of eigenoscillations more or less. The central maximum to which the microwave frequency must be tuned can, however, be easily identified so that no ambiguities occur in the clock's operation.

Serious consequences for the clock's operation arise when the fields in the two exposure regions are slightly out-of-phase. This can be caused, for example, by non-symmetrical feeding of the cavity or by faults in the waveguide geometry. In the ideal resonator with no phase difference the flying atoms always stay in phase with the field. No Doppler shifts will happen as everything is equivalent to an atom which moves parallel the wave fronts of a travelling wave. A phase difference between the two interaction regions, however, can be regarded as an inclination of the wavefronts with respect to the atomic path which results in a Doppler shift of the resonance frequency.

All commercial Cs clocks are affectd by this error which may assume relative values of some 10^{-12}. In primary clocks, its amount is determined experimentally by reversal of the atomic beam. As the error then changes in sign, it can be evaluated from the frequency comparisons with a stable reference clock. Despite the correction, which is applied in the PTB's primary clocks, the contribution of this error predominates in the uncertainty estimate.

2.2 Properties of Cs Clocks

2.2.1 Uncertainty

The most important property of a primary clock is its uncertainty. It describes the possible deviation of the clock frequency from the true value given by the definition of the second. In other words, the uncertainty is a measure of how well one has succeeded in stabilizing the tunable oscillator to the centre of the atomic resonance and how well all line shifting effects have been corrected. As this can, however, be done only with finite accuracy, a residual uncertainty remains. Table 1 shows the various error sources and their contributions to the uncertainty of the PTB's primary clocks, CS1 and CS2.

Table 1: Contributions to the relative uncertainty of clocks CS1 and CS2 (in 10^{-14})

	CS1	CS2
phase difference of resonator	3	1
C field	0,3	0,5
spectral purity	0,4	0,4
microwave power	0,3	0,3
servo system	<0,2	<0,2
relativistic Doppler effect	0,1	0,1

The predominant influence of the phase difference of the two interaction fields can be seen. The origin and the implication of the other error sources cannot be discussed in detail here. Worth mentioning is that the relativistic Doppler effect prevails in the primary clocks of most other institutes as faster atoms are made use of there. In general, it can be said that it will hardly be possible to reduce the uncertainty of the primary Cs clocks by an additional order of magnitude. After the predominant contribution to the uncertainty has been eliminated, one is usually confronted with a number of effects of approximately the same size which would then have to be fought against at the same time.

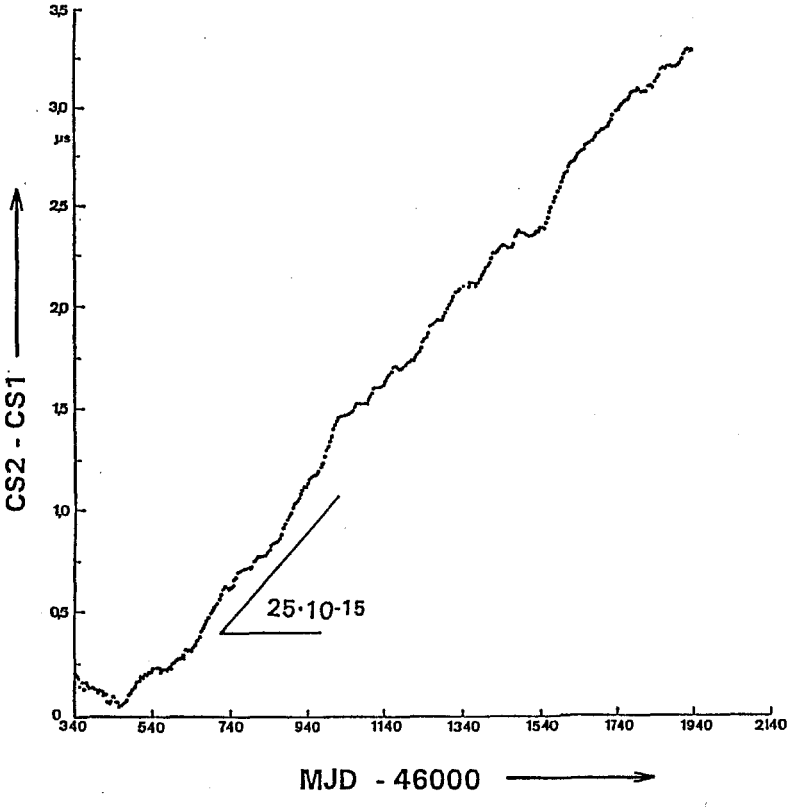

Fig. 8: Accumulated time difference of the two primary clocks CS1 and CS2 at PTB: MJD (Modified Julian Date) is a decimal day count; MJD = 46 340 corresponds to Oct. 2nd, 1985.

Fig. 8 shows the time difference between CS1 and CS2 adding up in the course of the years. The clocks' frequency difference is reflected in the curve's slope. It can be seen that the clocks deviate on an average by $2.5 \cdot 10^{-14}$. This frequency difference is compatible with the uncertainty estimates of Table 1.

2.2.2 Instability

Apart from the uncertainty, the instability is the next important characteristic of an atomic clock. It describes the frequency variations of the standard and determines the averaging time τ required to obtain reproducible frequency measurement values. The instability is usually stated in the form of the two-sample standard deviation $\sigma(\tau)$, often also known by the name of Allan variance [7].

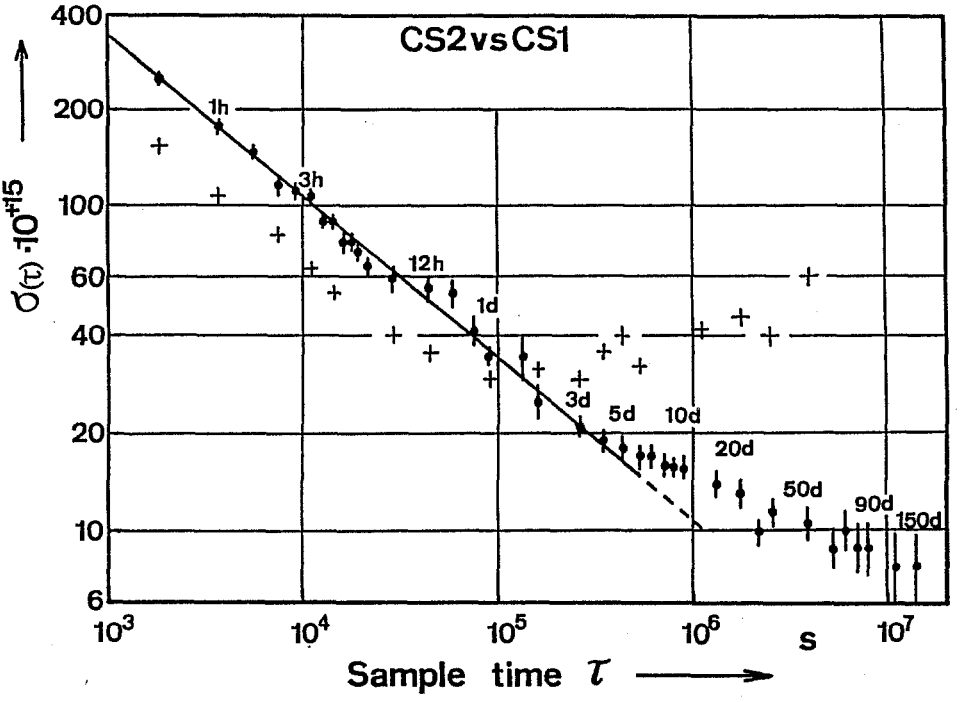

Fig. 9: Two-sample standard deviation $\sigma(\tau)$ of the primary clocks CS1 and CS2: crosses show the measurement of a commercial high-performance Cs clock versus CS2.

Fig. 9 shows a $\sigma(\tau)$ plot of the PTB's primary clocks, CS1 and CS2. It can be seen that the instability still decreases up to the averaging times of 150 days reached now. The behaviour of a commercial Cs clock, a so-called high-performance Cs clock, has been plotted for comparison. An increase of $\sigma(\tau)$ for averaging times which exceed a few days is evident. This points to a frequency drift, which can be caused by a change in one of the many parameters, which affect the clock operation (see table 1).

As long as any systematic change of this kind can be neglected the instability is only limited by the shot noise of the atomic beam and the linewidth of the transition. For this case the dependence of $\sigma(\tau)$ from the different physical quantities is given in Fig. 10.

Whereas $(N \cdot \tau)^{-1/2}$ reflects the well known influence of the shot noise, the linewidth linearly depends upon the time of flight T of the atoms through the cavity including the area between the two interaction zones.

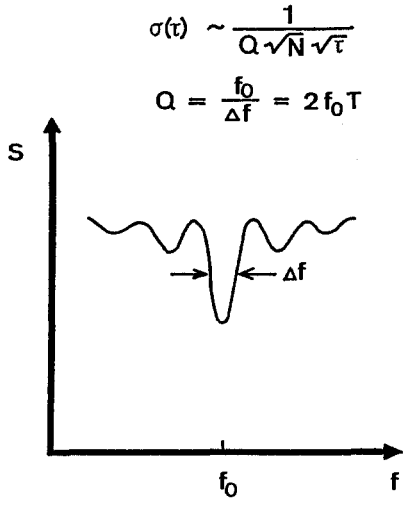

Fig. 10: Dependence of the two-sample standard deviation upon shot noise and linewidth: S detector signal; f frequency of cavity field; τ averaging time; N number of atoms per second, T interaction time.

If a Cs clock of especially small instability at finite averaging times is to be developed, either the number N of the particles in the beam must be increased (oven temperature) or the time of flight T extended (longer resonator, selection of slower atoms). For technical reasons, unfortunately neither of these means can be used to achieve greater progress for classic Cs clocks [8].

3. New Approaches

When 25 years ago the second was redefined on the basis of an atomic transition, there were good reasons to select the cesium.

Table 2: Characteristics of Cs and consequences

heavy atoms and low melting point	===>	beam of slow atoms
large differences in the magnetic moments	===>	magnetic state selection
low ionization energy	===>	efficient hot wire detector

While the two last-named advantages are rather of a technical nature, the first one is of the fundamental type. Both the cavity phase difference and the relativistic Doppler effect cause frequency errors and thus contributions to the uncertainty, which decrease with decreasing velocity. In addition, the instability is reduced by the associated prolonged time of flight.

All these advantages must, however, today be seen in relative terms. New technologies like ion trapping, laser cooling, optical pumping and optical detection suggest both the use of other atoms or molecules (Hg, Be, Mg, Yt, O_SO_4) and other clock designs. The two concepts at present considered most promising will be briefly described in the following.

3.1 Ion Traps

Small ion clouds, and even individual ions, can be enclosed in a confined space in electromagnetic cages (Fig. 11). Such traps consist of a ring-shaped electrode and two caps to which a suitable voltage is applied. A distinction is made between radio frequency or Paul traps where this voltage consists of d.c. and a.c. components, and Penning traps which require a static magnetic field iln addition to a d.c. voltage.

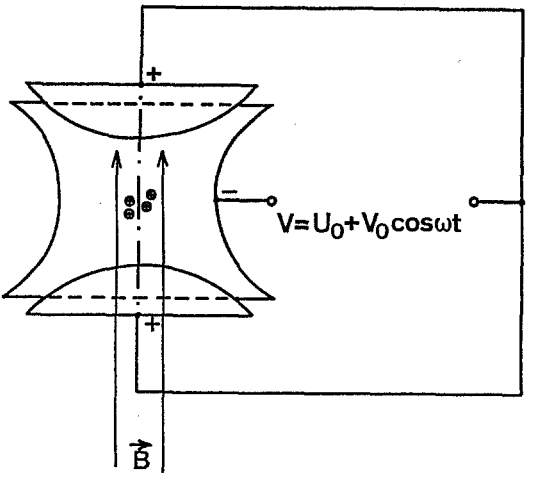

Fig. 11: Jon traps are possible to work in one of two ways: $V_o \neq O$, $B = O$ (Paul trap), $B \neq O$, $V_o = O$ (Penning trap).

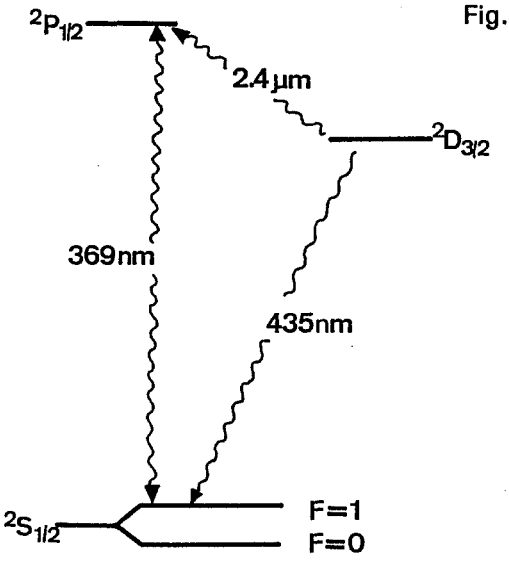

Fig. 12: Partial term scheme of $^{171}Yb^+$.

The residual movement of the ions around the trap centre is reduced with the aid of so-called laser cooling. For this purpose, the trap centre is exposed to a laser radiation whose frequency is red-shifted in relation to a transition in the ion. Due to the Doppler effect, only ions flying towards the beam are able to absorb photons. Each time a momentum is transmitted in the direction of the beam, while emission takes place statistically in any direction so that a slowing-down of the ions follows as the net effect.

Very slow ions are then obtained on which long times of interaction with fields can be realized [9]. As an example, the clock operation of an ion trap with ytterbium will be explained.

Fig. 12 shows the term scheme of $^{171}Yb^+$. As in the case of cesium, the hyperfine splitting of the ground state, which is here 12.6 GHz, is used as clock transition. Tuning of a 12.6 GHz microwave to this transition is achieved in intervals. To begin with, the ions are cooled using radiation in the blue spectral region, which belongs to a strong transition between the S and the P state. Then the state selection is achieved by means of the same radiation which, for this purpose, is tuned to a transition between a level of the P state and the upper hyperfine level of the ground state (F = 1). As a result, atoms are continually pumped to the P level from which they fall back again into the two hyperfine

levels of the ground state. After a short time the upper level is empty, laser light is no longer absorbed and the fluorescence radiation at 369 nm extinguishes. Now the microwave is switched on which again populates the upper hyperfine level of the ground state if its frequency is correctly tuned. As a result absorption takes place again and the fluorescence radiation reappears. Its intensity can therefore be made use of to tune the microwave to the clock transition.

The decay of the P state, however, does not always lead to the ground state. With a branching ratio of 1:300 the ions can escape to the metastable D state, where they are lost for the clock operation due to the long livetime of this state. Therefore, an additional laser with a wavelength of 2.4 μm is required for pumping the ions back to the clock experiment.

Higher frequencies - such as the 435 nm transition - cannot yet be made use of for a clock. It is not yet possible to count the cycles of an optical transition. Nevertheless, research is going on in this field mainly for two reasons: the instability is inversely proportional to the frequency (see Fig. 10) and a greater distance of the energy levels in the term scheme makes the associated frequency less sensitive to shifts of these levels due to electromagnetic fields (Zeeman or Stark effect).

3.2 Cs Fountain Clock

The new techniques can also be made use of to improve the classic Cs clocks. Work is at present in progress on the so-called fountain clocks (Fig. 13).
In this case a cloud of neutral Cs atoms is captured in a kind of molasses by exposure to laser light from six directions. The laser frequencies are slightly red-shifted with respect to an atomic resonance at 850 nm. As a result, atoms trying to escape from the molasses get into resonance with one of the beams on account of the Doppler effect, and they are pushed back by absorbed photons.

Then, by properly changing the frequencies of the vertical beams, a moving molasses is created that drags the atoms upwards. When they have reached a velocity of a few m/s all molasses lasers are switched off and the state selection is achieved by a properly tuned laser pulse of a few ms. On their way up and down the launched atoms interact twice with the field of a microwave cavity, exactly as they do in a normal Cs clock. The microwave transition or the correct frequency of

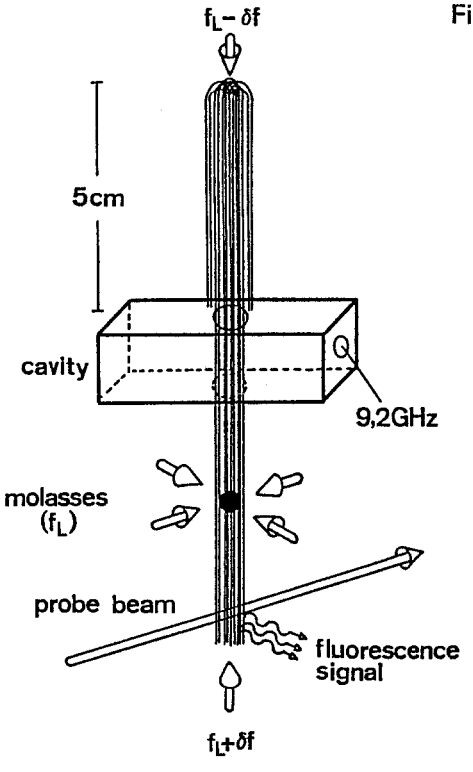

Fig. 13: Cs fountain clock.

the cavity field is checked by measuring the induced fluorescence of the atoms when they cross a probe beam on their way down.

Although such a fountain clock seems to be very complicated, its feasibility has already been demonstraded [10] and it may be expected that its realization will improve the accuracy and the instability of Cs clocks by another order of magnitude. Problems related to the cavity phase difference of the classic Ramsey resonator are avoided as the atoms interact twice with the same field and the generation of slow atoms with long interaction times guarantees a narrow linewidth of the microwave transition.

References

1. Scheibe, A.; Adelsberger, U.: Phys. Zeitschr., 37(1936), p. 185
2. Danjon, A.: Astronomie générale, Paris (1959)
3. Wells, J.W.: Nature, 197(1963), p. 948

4. Taylor Jr., J.H.: Proc. IEEE, 79(1991), p. 1054
5. Allan, D.W. et al.: IEEE Trans., IM-34(1985), p. 118
6. Bauch, A. et al.: IEEE Trans., IM-36(1987), p. 613
7. Allan, D.W.: Proc. IEEE, 54(1966), p. 221
8. Dorenwendt, K. et al.: Physica Scripta, 41(1990), p. 712
9. Wineland, D.J.: Science, 226(1984), p. 395, TN-141
10. Clairon, A. et al.: Europhys. Lett., 16(1991), p. 165

Gravity-Wave Astrophysics

Gerhard Schäfer

Max-Planck-Institut für Astrophysik, W-8046 Garching bei München, and MPG-Arbeitsgruppe "Gravitationstheorie", Universität Jena, O–6900 Jena

Abstract: An overview is given of the gravity-wave emission from astrophysical sources. The detectability of gravity-wave signals on Earth and in space is discussed.

1 Introduction

Albert Einstein predicted gravitational waves shortly after his final formulation of general relativity: In 1916, within linearized approximation, Einstein discussed the generation and propagation of gravitational waves, and in 1918 he derived his famous quadrupole formula for the gravitational energy emission from non-self-gravitating systems (only in 1941, L. Landau and E. Lifshitz generalized this formula to weakly self-gravitating systems). It took then about 40 years before in particular through the work by Hermann Bondi the physical reality of gravitational waves became evident: In 1957 Bondi devised a primitive detector for gravitational waves. Joseph Weber then started the later world-wide activity in the search for cosmic gravitational waves (Weber 1960).

The first who gave mathematical expressions for cosmic gravitational waves from realistic sources were Peters and Mathews (1963). They worked out the gravity-wave emission from Newtonian binary star systems in bounded Keplerian motion. Gravitational waveforms emitted by test bodies falling radially into a Schwarzschild black hole were given for the first time by Davis, Ruffini, and Tiomno (1972). And the first gravity-wave signals from quasi-realistic models of collapsing stars (collapse of homogeneous ellipsoids) were obtained by Saenz and Shapiro (1978).

Our most advanced knowledge of gravity-wave signals from astrophysical sources will be the subject of the present article. For a rather exhaustive

account of the gravity-wave research, theoretically and experimentally, up to the year 1986, the reader is referred to the review by Kip Thorne (1987).

By the year 1992, the existence of gravitational waves has been revealed with high precision. Arrival-time measurements of the radio signals from the binary pulsar PSR 1913+16, running since 1974, show up an orbital-motion decay consistent with the gravity-wave emission according to general relativity with an accuracy better than 0.5% (Taylor et al. 1992).

2 Gravitational Waves

General relativity tells us that gravitational waves are ripples in the geometry of spacetime which propagate with the speed of light. They have two independent degrees of freedom which live in the plane orthogonal to the propagation direction. For the description of the waves on Earth where the gravitational field is weak, it is sufficient to treat the waves as ripples in Minkowski space.

Let $\eta_{\mu\nu} = \text{diagonal}(-,+,+,+)$, with $\mu,\nu = 0,1,2,3$, be the Minkowski metric and let h_{ij}^{TT}, with $i,j = 1,2,3$, denote the gravity-wave field (h_{ij}^{TT} is a transversal and traceless symmetric 3-tensor which, for each component, fulfils the usual wave equation in Minkowski space), then the metric $g_{\mu\nu}$ of spacetime, on Earth, takes the form $g_{00} = -1 +$ small non-propagating terms, $g_{0i} =$ small non-propagating terms, $g_{ij} = \delta_{ij} + h_{ij}^{TT} +$ small non-propagating terms. The two "linear" polarization states of a gravitational wave are usually denoted by h_+ ("plus"-polarization) and $h_\times$ ("cross"-polarization). Their orientations enclose an angle by 45°. In spherical polar coordinates (θ and ϕ), centered at the source of the wave, $h_+ = h_{\theta\theta} = -h_{\phi\phi}$ and $h_\times = h_{\theta\phi} = h_{\phi\theta}$ hold.

The influence gravitational waves have on gravity-wave detectors results from the solutions of the general covariant detector equations of motion. For laserinterferometric gravity-wave detectors, the covariant mirror and the covariant light-ray equations of motion have to be solved. These problems are easiest treated in the coordinate system x^i to which the metric coefficients, given above, belong. In this coordinate system, the mirrors, treated as free test masses, remain at rest when a gravitational wave passes through, and the influence of the gravitational wave onto the laser light results from the solution of the light-ray equations i.e. the equation $0 = g_{\mu\nu}dx^\mu dx^\nu$, essentially, with the simple boundary condition of fixed 3-space coordinates. In the case of bar detectors the influence of the wave is best described in terms of the tidal force the bar experiences during the passage of the wave. Under the assumption that the reduced wavelength of the gravitational wave is large compared to the extension of the bar, the potential of this force reads $(1/2)c^2 R_{0i0j}X^iX^j$ where the curvature tensor (Riemann tensor) $R_{\mu\nu\rho\sigma}$ is related with the second time derivative of the wave field

through $R_{0i0j} = -(1/2)c^{-2}h^{TT}_{ij,tt}$. The coordinates X^i are different from the coordinates x^i used above. They refer to an orthonormal basis of vectors in the mass-center of the bar, i.e. they are geometrical objects. In this coordinate system, free test masses do not remain at rest when a gravitational wave passes through and, for laserinterferometric gravity-wave detectors, the equations of motion for the light rays, now in Minkowski space in leading approximation, have to be solved under time-dependent boundary conditions. The relation between the two coordinate systems in question, (t, x^i) and (T, X^i), reads, $x^i = X^i - (1/2)h^{TT}_{ij}X^j$ and $t = T - c^{-2}(1/4)h^{TT}_{ij,t}X^iX^j$. Hereof the relative elongation of two free test masses at the positions X^i and $X^i + L^i$ is easily deduced. It holds, $dL^i = (1/2)h^{TT}_{ij}L^j$. In Fig.1 of the contribution by Danzmann et al. (these proceedings), the action of the plus–polarization is shown. For further details see, e.g. Thorne (1983).

The gravity-wave field, h^{TT}_{ij}, can be characterized by mass- and current-(or spin-) radiative multipole moments, respectively time derivatives of them (l-th time derivative for the 2^l-pole moments), starting with the quadrupole moments (e.g. see Thorne 1980). Usually, the quadrupole contribution is the most important one. In a slow-motion framework, the radiative quadrupole moments are known as functionals of the source variables up to the first post-Newtonian approximation (Blanchet and Damour 1989; Blanchet et al. 1990; Damour and Iyer 1991). To leading order, the radiative moments are identical with the corresponding source moments in the Newtonian theory. For test bodies falling into non-rotating or rotating black holes, simple uncoupled linear differential equations (Regge-Wheeler equation, Zerilli equation, Teukolsky equation) relate the source multipole moments with the radiative multipole moments. With the exception of black-hole collapse and black-hole collision, all simulations referred to in the present article are based on the concepts and structures just mentioned.

3 Burst Sources

The most important burst sources for gravitational waves are (1) the collapse of iron stellar cores to neutron stars, (2) the collapse of compact stars to black holes, (3) the coalescence of neutron stars and/or black holes, and (4) the fall of stars or small black holes into supermassive black holes.

3.1 Collapse to Neutron Stars

The collapse of iron stellar cores to neutron stars is very likely the trigger mechanism of Type II supernova explosions. This mechanism is one of the most ambitiously investigated processes in astrophysics (Woosley and Weaver 1986). However, in spite of big efforts, the underlying dynamics is still not fully understood. Also collapse simulations with rotation were not able to solve the problem (Müller and Hillebrandt 1981, Mönchmeyer and Müller 1989). The shock wave which propagates through the outer core still turns into an accretion shock without being able to heat and to emit the envelope. That the model calculations nevertheless show some realistic features stems from the Kamiokande and Irvine-Michigan-Brookhaven neutrino events of the supernova 1987A which are in accord with the theoretical predictions. Thus one may expect that also the gravitational waves, emitted in case of collapse simulations, are not totally unrealistic, and this so much the more, as the gravitational waves are mainly generated through the same process which also triggers the shock wave, namely the rather well understood bounce of the inner core. The question if the shock wave stalls into an accretion shock in the outer core or not depends very much on processes in the outer core, e.g. dissociation of the iron nuclei in the outer core or reviving of the shock wave through neutrino-antineutrino annihilations - and these processes have no influence on the gravity-wave signal.

Stars with initial mass in the range of about $10 - 20 M_\odot$ end up with low entropy, typically $1.4 M_\odot$-stellar iron cores. The cores are supported primarily by relativistic electron degeneracy pressure and are thus only marginally stable against collapse. Electron capture onto free protons lowers the electron concentration and the pressure support, and thus triggers the gravitational collapse. For stars in the mass range of about $20 - 40 M_\odot$ mainly photodesintegration of iron-peak elements to α-particles triggers the instability. Those stars develop more massive, higher entropy iron cores before collapse.

The collapse of rotationally deformed stellar cores produces gravitational waves. The transition from spherically symmetric, non-rotating stellar evolution calculations, the only detailed microscopic calculations performed so far, to rotationally deformed collapse simulations has found several treatments: (i) perturbations superimposed on "realistic" spherically symmetric core collapse simulations (Turner and Wagoner 1979; Seidel and Moore 1987; Seidel et al. 1988), (ii) aspherical Newtonian collapse dynamics of homogeneous and inhomogeneous, uniformly rotating ellipsoids with a relatively crude treatment of the microphysics (Novikov 1975; Shapiro 1977; Saenz and Shapiro 1978, 1979, 1981; Moncrief 1979; Ipser and Managan 1984), (iii) axisymmetric Newtonian collapse dynamics from initially rigidly rotating, hydrostatic equilibrium configurations, with polytropic equation of state with density dependent adiabatic index (Finn and Evans 1990), (iv) axisymmetric Newtonian collapse simulations with differentially rotating,

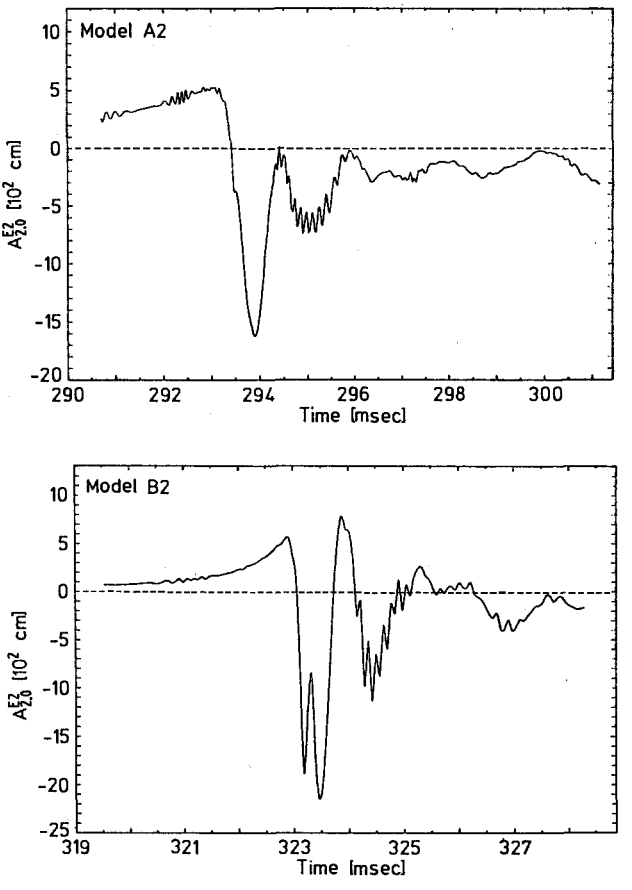

Fig. 1. Quadrupole waveforms for the Models A2 and B2.

spherically symmetric initial conditions and explicit treatment of the microphysics (Müller and Hillebrandt 1981; Mönchmeyer and Müller 1989). The models (i), (iii), and (iv) allow in a natural manner for shocks. For numerical reasons the shocked matter was treated by the addition of artificial viscosity.

For several reasons, the gravitational waves of the models (iii) and (iv) seem to be the more realistic ones. Concerning models (i), the applied perturbation technics do not allow for bulk motion, they are only applicable to slowly rotating bodies, and one has to impose rather arbitrary initial conditions; and the models (ii) suffer somewhat from their rigidity assumptions. In contrast herewith, the models (iii) and (iv) allow for differential rotation and free volume deformations, respecting the assumed axisymmetry.

The Fig.1 presents waveforms obtained by Müller (1982). Shown is the plus-polarization state, h_+; the cross-polarization state, $h_\times$, is zero in the actual approximation. The rotation axes of the stellar cores are assumed to lie along the z-axis; therefore the axisymmetric core dynamics does not depend on the azimuthal angle ϕ. The amplitude $A^{E2}_{2.0}$ is defined by $rh_+ = (15/64\pi)^{1/2} A^{E2}_{2.0} \sin^2\theta$.

Initially, the models are spherically symmetric with differential rotation $\Omega(r) = \text{const.}$ for $r < 10^7$cm ($r < 10^8$cm) and proportional to $r^{-1}(r^{-2})$ for $r > 10^7$cm ($r > 10^8$ cm), and with rotational energy of 1.6×10^{49} erg (6.2×10^{49} erg) for the innermost $1.4M_\odot$, respectively for the Models B2 (A2). The collapse was initiated by reducing the hydrostatic equilibrium entropy by 5% (Müller and Hillebrandt 1981).

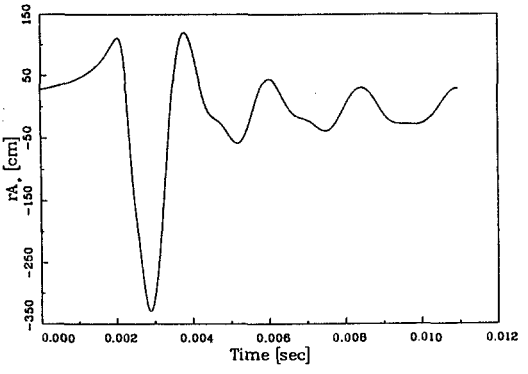

Fig. 2. Quadrupole waveform.

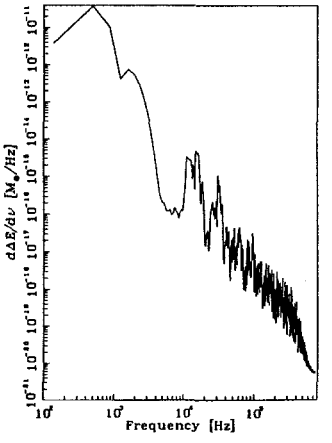

Fig. 3. Energy spectrum corresponding to Fig.2.

The Fig.2 gives the waveform obtained by Finn and Evans (1990). A_+ is identical with h_+. In Fig.3 the energy spectrum is shown (the velocity of light is put equal to one in the unit for the energy).

Initially, the angular velocity is uniform, 0.16 rad sec^{-1}, and the matter is in hydrostatic equilibrium. The angular momentum of the $1.4M_\odot$-stellar core is 2×10^{49} g cm^2 sec^{-1} and the "radius" amounts to 9×10^8cm. The collapse was initiated by reducing the hydrostatic equilibrium internal energy by 1%.

The presently most detailed collapse simulations of rotating stellar cores are the Models A, B, C, and D introduced and investigated by Mönchmeyer and Müller (1989). Not in hydrostatic equilibrium, initially, these simulations are more complete than the previous adiabatic (with the exception of the shock generation) simulations by Müller and Hillebrandt (1981). These simulations have taken into account, equation of state data based on detailed nuclear-statistical equilibrium and Hartree-Fock calculations (Hillebrandt and Wolff 1985), electron and neutrino captures on free protons and neutrons, respectively, neutrino trapping at densities $\rho > 3 \times 10^{11}$ g cm^{-3},

and local angular momentum conservation. In particular, in contrast to the previous simulations, the neutrino pressure from the kinetic equilibrium of the weak interaction was taken into account. Inside the trapping region the neutrinos were assumed to reach thermal equilibrium instantaneously. The hydrodynamical equations were solved by the aid of a conservative, explicit numerical code of second order accurate differencing (Mönchmeyer and Müller 1989). The initial collapse configurations are summarized in Table 1.

Table 1. The quantities X, Z, R, Ω_0, $E_{\rm rot}^{1.4}$, $E_{\rm rot}^{1.0}$, and β are the distance from the core center in the equatorial plane, the distance from the equatorial plane, the corresponding radial distance, the central angular velocity, the rotational energy for the inner $1.4M_\odot$ and $1.0M_\odot$, and the ratio of rotational to potential energy for the innermost $1.4M_\odot$, respectively. The rotational energy is growing along the sequence A-C-D-B.

Model		A	B	C	D
Ω/Ω_0		$\frac{R^2}{(r^2+R^2)}$	$\frac{R^2}{(r^2+R^2)}$	$\frac{R}{(r+R)}$	$\frac{X^2}{(x^2+X^2)}\,\frac{Z^4}{(z^4+Z^4)}$
R	[10^8 cm]	1.0	1.0	0.1	–
X	[10^8 cm]	–	–	–	1.0
Z	[10^8 cm]	–	–	–	1.0
Ω_0	[s^{-1}]	4.0	8.0	22.0	5.5
$E_{rot}^{1.4}$	[10^{49} erg]	2.2	8.7	2.9	4.3
$E_{rot}^{1.0}$	[10^{49} erg]	1.2	4.7	1.9	2.3
β		0.004	0.018	0.007	0.010

The four waveforms derived hereof by Mönchmeyer et al. (1991) are shown in Fig.4. The amplitude $d^2M_{20}^{E2}/dt^2$ is identical with the amplitude of Fig.1, i.e. with $A_{2,0}^{E2}$. The Fig.5 presents the energy spectra of the waveforms of Fig.4.

The curves in Fig.1 are much more noisy compared to the curves in the Figs.2 and 4. The main reason for this is the use of a wave extraction formula in case of Fig.1 which involves two numerical time differentiations. The Fig.2 was obtained by the performance of only one time differentiation, and in Fig.4 even a wave extraction formula without numerical time differentiations (Epstein 1978; Nakamura and Oohara 1989; Blanchet et al. 1990) was applied. Like the other two extraction formulae, the latter formula has compactly supported integrands. At high frequencies the energy spectra of the waves of Fig.1 are strongly dominated by noise. They have been omitted on this reason.

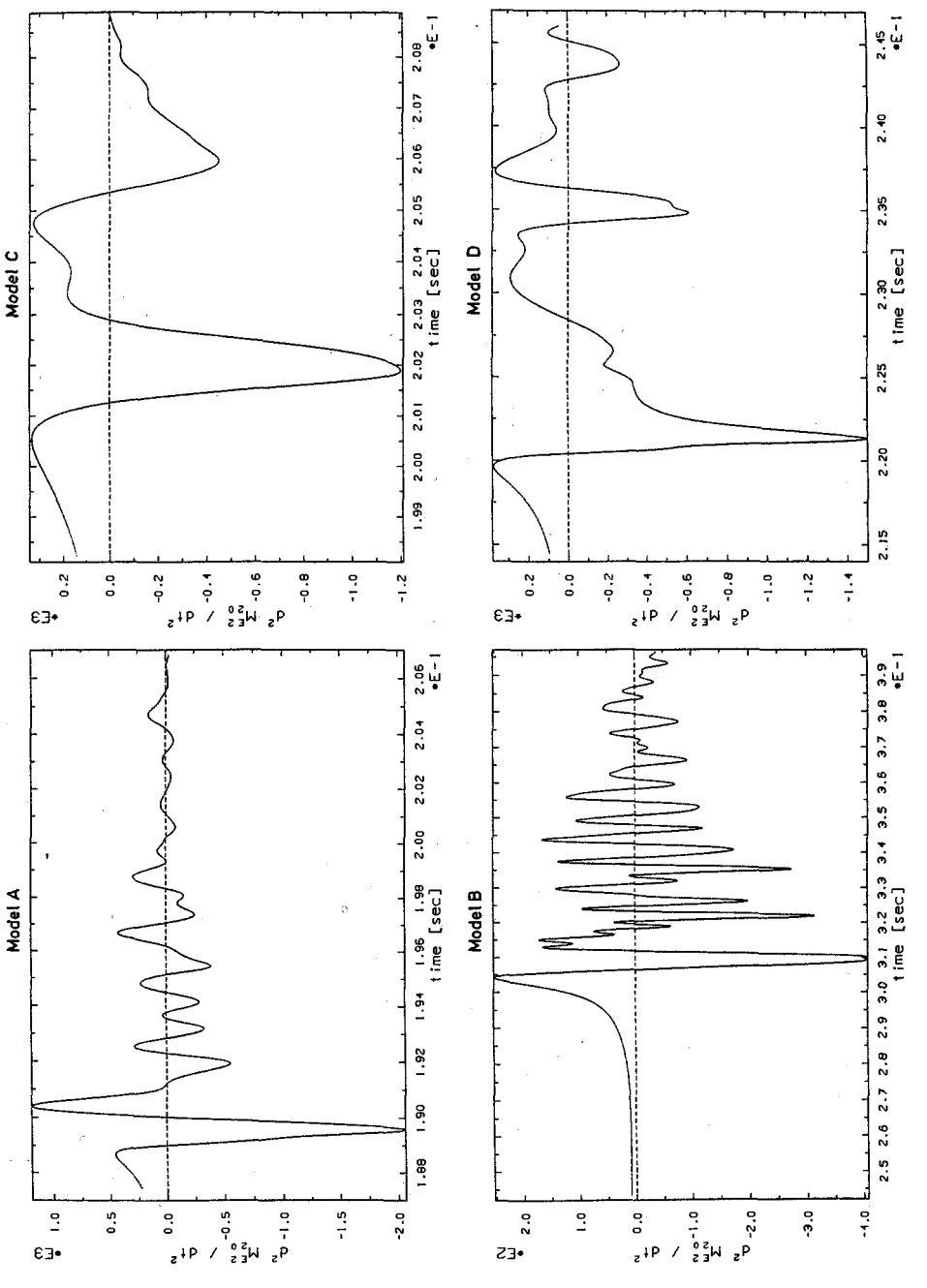

Fig. 4. Quadrupole waveforms for the Models A, B, C, and D.

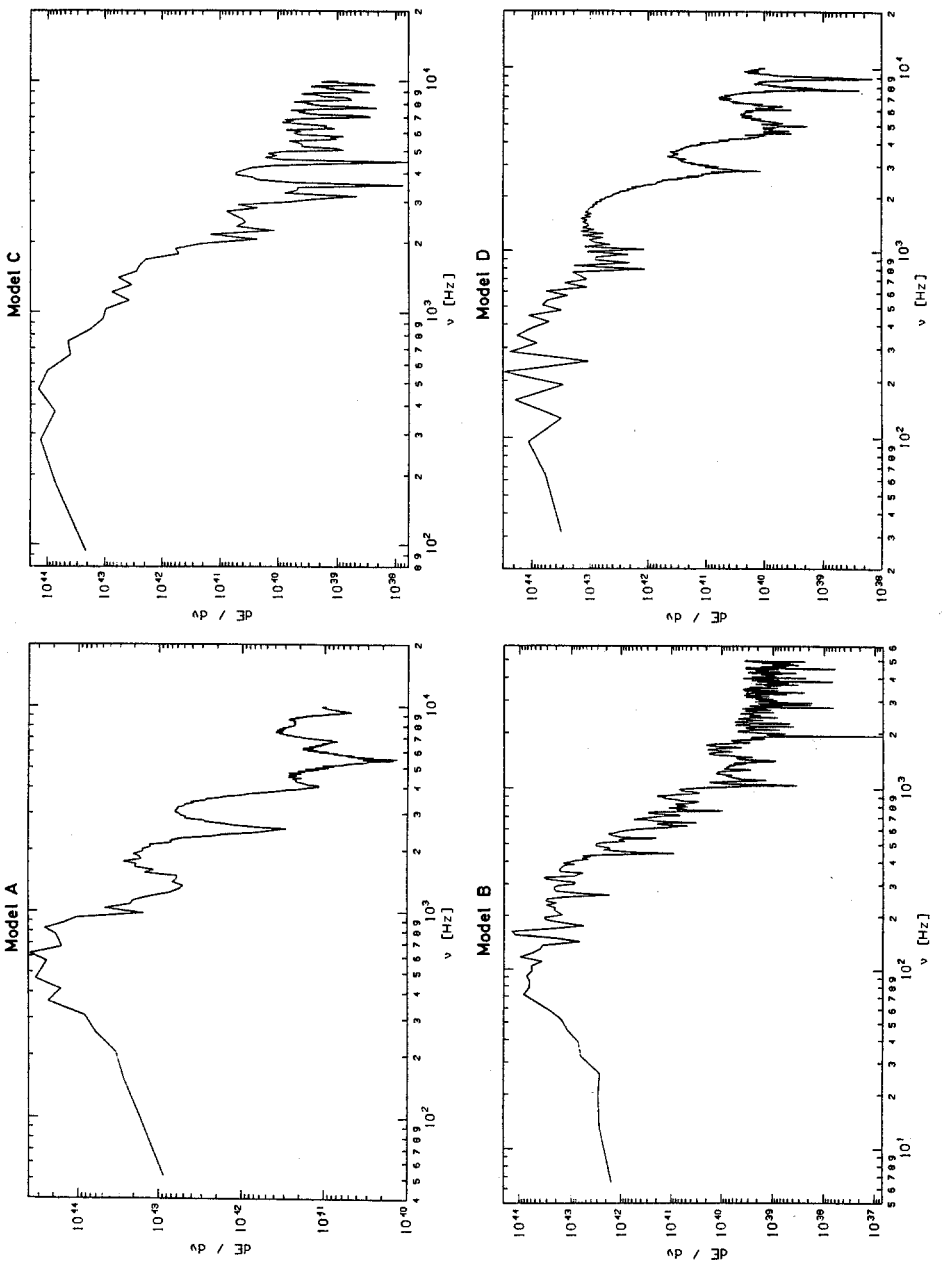

Fig. 5. Energy spectra corresponding to Fig.4.

The rotational energies of the Models A, B, C, and D are growing along the sequence A-C-D-B, see Table 1. The rotational energy of Model A2, Fig.1, lies in-between the rotational energies of the Models D and B, the rotational energy of Model B2 is slightly smaller than the one of Model A. The collapse of Model A is stopped by nuclear forces, whereas for the Models D and B centrifugal forces are responsable for the bounce (at bounce, Model D is just approaching nuclear density). The bounce of Model C has its origin in both forces. The Model D is exceptional because it shows two bounces, separated by a large-scale oscillation (notice also the special differential rotation pattern, initially, see Table 1). Correspondingly, the gravitational wave signal of Model D is different from the wave signals of all the other Models A, B , C, A2, B2, and Fig.2. The more complicated structure of the waveform of Model B, compared to e.g., Model A, results from the superposition of several 2d-oscillation modes.

The maximum wave amplitudes of the Models A, C, D, A2, B2, and Fig.2 are practically the same, 10^{-23} for a source at a distance of 15 Mpc (approximate distance to the Virgo cluster). The energy in the waves, in units of $10^{-8}M_{\odot}c^2$, amounts to 7.6, 1.2, 1.1, 0.2, and 2, for the Models A, C, D, B, and Fig.2, respectively. We do not quote the corresponding results for the Models A2 and B2 because of the mentioned noise problem.

3.2 Collapse to Black Holes

The collapse of compact stars to black holes, predicted by theory, besides the big bang, is the most interesting and spectacular process in gravity theory. Its verification in Nature will be of utmost importance. Gravitational waves can help to reveal non-spherically symmetric collapse to black holes and consequently the existence of black holes themselves.

The gravitational radiation emission from axisymmetric black-hole collapse has found an excellent treatment by Stark and Piran (1985), Piran and Stark (1986) (also see Nakamura et al. 1987). The starting point there was a spherically symmetric, rigidly rotating compact star of radius $6GM/c^2$ with pressure deficit of 1% compared to a corresponding non-rotating star in hydrostatic equilibrium (M denotes the mass of the star). The assumed polytropic equation of state with adiabatic index two resulted in an initial central density of $1.9 \times 10^{15}(M/M_{\odot})^{-2}$ g/cm^3. In units of cm, the initial radius was $8.8 \times 10^5 M/M_{\odot}$. The total angular momentum J of the star was measured by the dimensionless parameter $a = Jc/GM^2$.

For $a < 1$ the collapse proceeds to black-hole configuration. The gravitational waveforms are shown in Fig.6 for the two polarization states, h_+ and $h_\times$. The gravitational waves were taken at $r = 50GM/c^2$. The energy spectra for the $a = 0.79$ case are given in Fig.7.

The maximum h_+-amplitude is closely approximated by (rc^2/MG) $|h_+|_{\max} = \min(0.1a^2, 0.06)$, cf. Fig.6, the energy spectra peak at frequencies

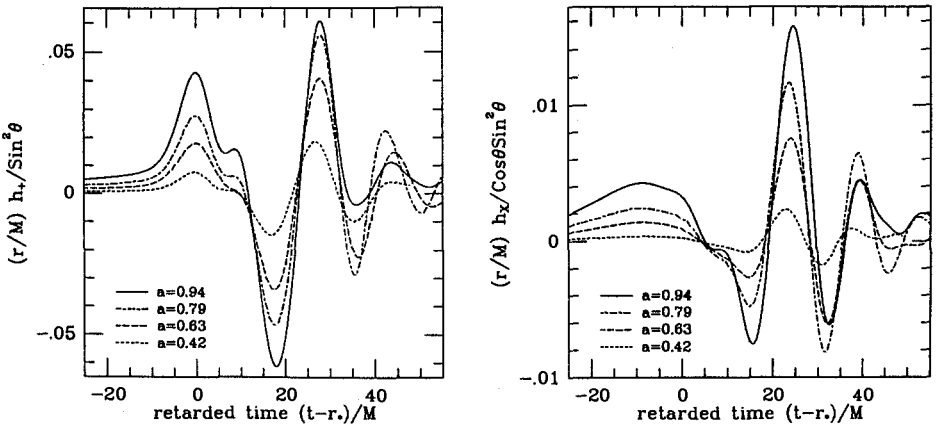

Fig. 6. Waveforms vs. retarded time for various black hole collapse processes, $G = c = 1$.

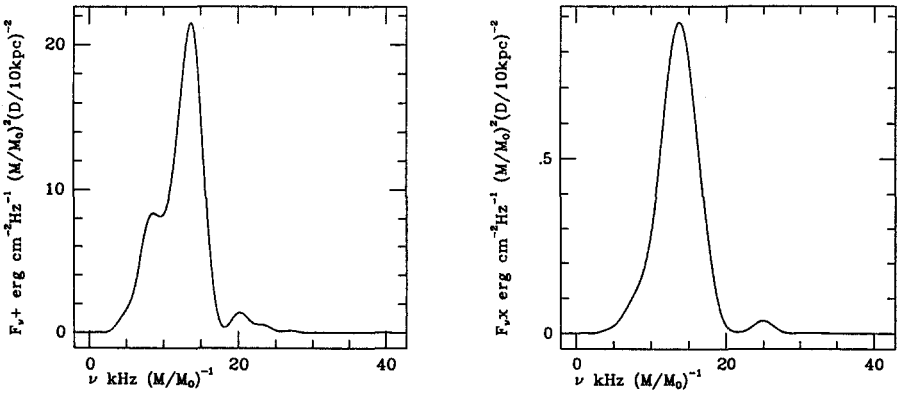

Fig. 7. Energy spectra for the collapse with a = 0.79, derived from the energy flux and normalized to a star of $1M_\odot$ at a distance of 10kpc.

of $7 < \nu < 16\, M_\odot/M$ kHz, cf. Fig.7, and the radiated energies ΔE behave as $\Delta E/Mc^2 = 10^{-3}$ min$(1.4a^4, 0.6)$. The cross- and plus- polarizations are

related by $|h_+| > 5|h_\times|$ and $(\Delta E/Mc^2)_+ > 10\,(\Delta E/Mc^2)_\times$. The similarity of the waveforms for different values of "a" is quite remarkable. It can be understood in terms of black-hole quasi-normal modes: The gravitational waves are excited strongest when the size of the star, in Schwarzschild coordinates, is about $3GM/c^2$ (marginally stable circular orbit for light rays), i.e. when the exterior metric of the star is already very nearly a black-hole metric.

Several collapse calculations with dust have found very similar wave forms (Cunningham et al. 1978, 1979, 1980; Petrich et al. 1985).

For a $20M_\odot$ collapse to a black hole at a distance of 15 Mpc the gravitational wave amplitudes could reach values of up to 4×10^{-21} and frequencies of the maxima of the energy spectra of up to 800 Hz.

3.3 Coalescence of Compact Objects

The coalescence of neutron stars and/or black holes is a very strong source for gravitational waves. In Nature, the usually realized coalescing situation is from quasi-circular orbits because gravitational radiation reaction circularizes inspiraling orbits rather quickly.

The presently most complete picture of gravitational–wave signals from coalescing binaries stems from Newtonian hydrodynamical simulations with gravitational radiation damping (Oohara and Nakamura 1989, 1990; Nakamura and Oohara 1989, 1991). These simulations give some insight into the to be expected gravitational–wave signals from coalescing neutron stars. Of course, as long as tidal effects are negligible, the waveforms of coalescing neutron stars and/or black holes are identical (Lincoln and Will 1990; Junker and Schäfer 1992). For the head-on collision of equal-mass black holes the reader is referred to the investigations by Smarr (1979).

As representative example we pick up a simulation performed by Oohara and Nakamura (1990). The neutron stars, each with a mass of $1.49M_\odot$, were assumed to be in rigidly rotating equilibrium with center of mass distance of 15 km just when they touched each other (starting point of the hydrodynamical simulation,T = 0; for the earlier evolution, $T < 0$, the neutron stars were treated as point-like bodies moving in gravitationally damped, quasi–circular orbits). The neutron-star matter fulfilled a polytropic equation of state with adiabatic index two. The initial central mass densities of the neutron stars and the angular frequency of the binary system amounted to $4 \times 10^{15}\mathrm{g/cm^3}$ and $6.6 \times 10^3\ \mathrm{sec}^{-1}$, respectively. The coalescence was further driven by gravitational quadrupole radiation damping.

For $T > 0$, the waveforms for the plus- and cross-polarizations are shown in Fig.8; for the whole process, the plus-polarization is given in Fig.9.

The wave amplitude grows up to about 3×10^{-21} for a source at a distance of 15 Mpc, and the efficiency of the gravitational energy emission amounts to about 4%. The frequency of the coalescing part of the wave is

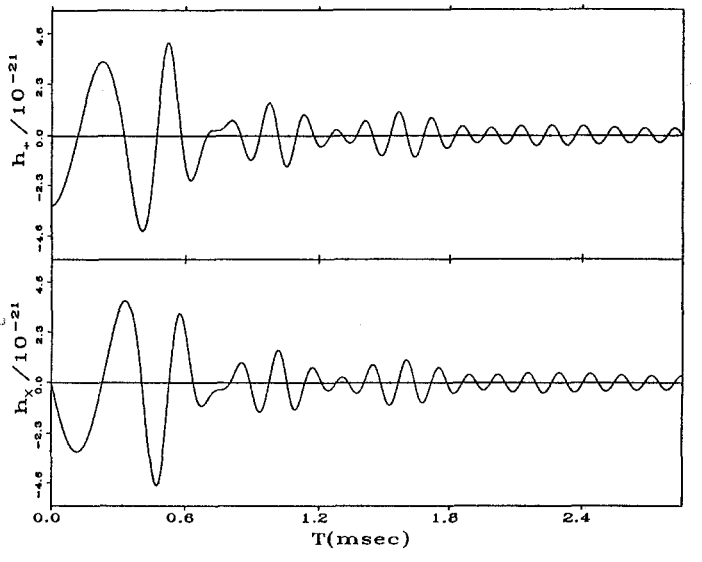

Fig. 8. Waveforms on the rotation axis at 10Mpc.

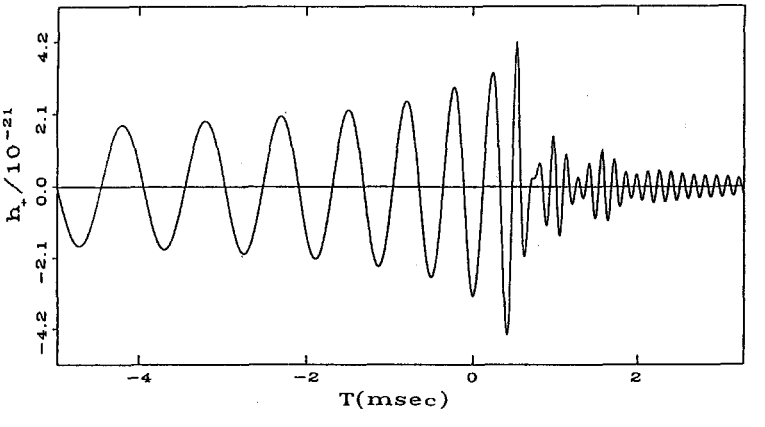

Fig. 9. Waveform on the rotation axis at 10Mpc.

about 7 kHz. The measurements of the frequency and of the amplitude at some instant of time before coalescence, and of the time interval from that time to the time of maximum amplitude allow the direct determination of the distance of the coalescing binary system (Schutz 1986).

The Fig.10 shows the time variation of the two-arm-phase-difference, $\Delta\varphi_{GR}$ of the laser light for a delay line and for a Fabry-Perot type laser interferometric detector as caused by the wave of Fig.9, see Oohara and Nakamura (1990). These detectors, with effective armlength of 75 km and cavity length of 3 km with appropriate reflection and transmission coefficients, are mainly sensitive to gravitational waves with frequencies at about 1 kHz.

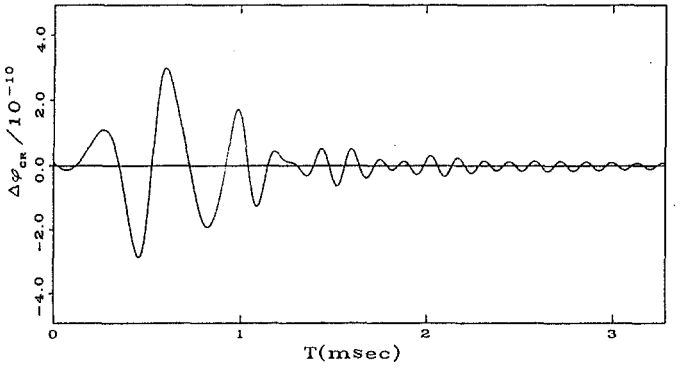

Fig. 10a. Time variation of phase difference for a delay line type detector with 75km effective armlength.

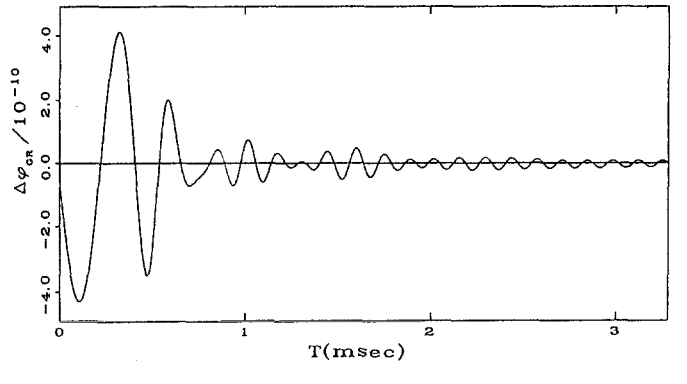

Fig. 10b. Time variation of phase difference for a Fabry-Perot type detector with 3km cavity length.

3.4 Fall of Stars or Small Black Holes into Supermassive Black Holes

The gravity-wave-burst problem of the radial fall of stars or small black holes (test bodies) into non-rotating supermassive black holes has found its representative treatment by Ferrari and Ruffini (1981). The Fig.11 shows the radial function of the quadrupole gravitational radiation perturbation (in units of $\frac{mG}{c^2}\sqrt{\frac{2}{\pi}}$) for the cases $\gamma_0 = 1$ and 1.5, where by definition $\gamma_0 = (1 - v_\infty^2/c^2)^{-1/2}$ holds. v_∞ is the speed of the test body (mass m) at infinity.

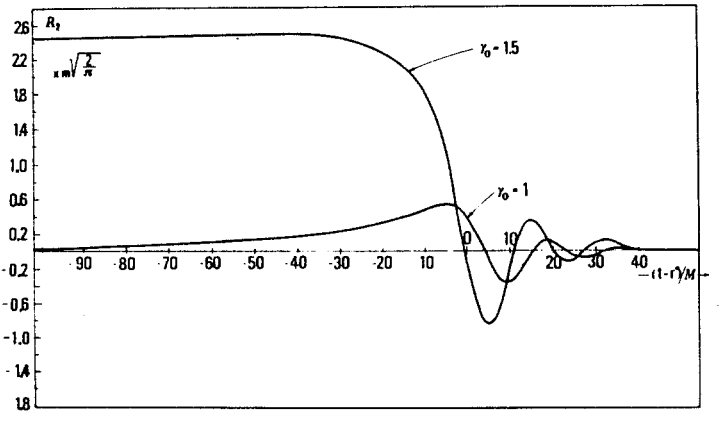

Fig. 11. The radial function of the gravitational quadrupole perturbation as function of the retarded time $(t - r_*/c)c^3/MG$, M denotes the mass of the black hole.

In Fig.11, it is interesting to note the "memory effect" for $\gamma_0 > 1$, i.e. the difference in the initial and final values of the perturbation (see Sec.6). This difference has been pointed out already by Ferrari and Ruffini (1981). As in the case of the black-hole collapse of Sec.3.2, the waveforms can be understood as superpositions of black-hole quasi-normal modes (also see Detweiler 1979).

The step from the head-on infall from infinity with zero initial velocity to the infall with non-vanishing orbital angular momentum has been made by Detweiler and Szedenits (1979); for a more recent treatment, see Oohara (1986) where one can find also the waveform from a ring plunging into a Schwarzschild black hole. Detweiler and Szedenits found an enhancement in the total emitted energy by a factor of 50 as the ratio $J/\mu M$ (J is the angular momentum, μ and M are the masses of the test body and the

supermassive black hole, respectively) increases up to 3.9. There, the total emitted energy, ΔE, takes the value $\Delta E = 0.5(\mu/M)^2\, Mc^2$ and the angular frequency, σ, peaks at about $\sigma = 0.32c^3/MG$, see Fig.12.

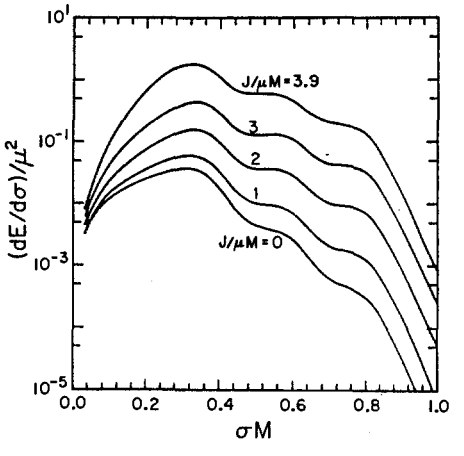

Fig. 12. The energy spectrum of the gravitational radiation emitted by a body falling into a Schwarzschild black hole with angular momentum J.

The special structure of the waveforms, "precursor - main-burst - ringing tail", first noticed by Davis et al. (1972) and later conjectured by Ruffini (1978) to be valid for all collapse processes to black holes, can indeed be seen in all Figs.6 and 11, and even in the Figs.1, 2, 9, and in parts in Fig.4.

The gravitational radiation from bodies moving in the gravitational field of rotating black holes has been calculated by Kojima (1986). Only for very fast-rotating black holes the results differ significantly from the non-rotating case. For a more detailed overview of the subjects treated in Sec. 3.4, see Nakamura et al. (1987).

4 Periodic Sources

Important periodic or quasi-periodic sources of gravitational waves are compact binary systems, test bodies in close orbits around black holes, and rotating neutron stars with nonaxial ellipticity. In the first two cases, the waveforms are of the type of the first part of the wave of Fig.9 (for the detailed structure, e.g. see Detweiler (1978) and Kojima (1986)), and in the latter case, of its last part. The importance of the gravity-wave emission for the rotational motion of newborn, rapidly rotating neutron stars has

been pointed out by Ostriker and Gunn (1969). The detectability of these gravitational waves on Earth has been investigated by Piran and Nakamura (1988). They conclude their detectability from neutron stars in our Galaxy with present detectors and from neutron stars in the Virgo cluster with advanced detectors.

5 Stochastic Sources

Some words should be said also to the stochastic gravity-wave background. The most spectacular contribution to this background are surely the gravitational waves from the Planck era of the big bang, i.e. from the time 10^{-43}sec after the big bang where the last scattering of the gravitons has taken place (cf. hereto the time of last scattering of neutrinos and photons of about 0.1 sec and 10^6 years, respectively). Because of our rather crude knowledge of the physics of the big bang, including the inflationary epochs, the to be expected amplitudes and frequencies of the primordial waves at the present epoch are very uncertain. Important to mention is the fact that the interaction of the gravitational waves with the large-scale background curvature yields parametricly amplified states of the gravitational radiation field (Grishchuk 1975). The GUT and the electro-weak phase transitions in the early universe also contribute to the stochastic gravity-wave background (Vilenkin 1981; Witten 1984). For frequencies in the region $\lesssim 0.06$ Hz the stochastic gravity-wave background from white-dwarf binaries will complicate the measurement of the stochastic components mentioned above (Hils et al. 1990).

6 Memory Effects

If the initial and final amplitudes of a gravity-wave burst are different and constant over a time interval which is long compared to the duration of the burst, one speaks of a burst with memory (Braginsky and Grishchuk 1985). Examples are the gravitational waves from the scattering of two bodies (e.g. see Junker and Schäfer 1992) and from the test-body head-on infall into a black hole with finite velocity at infinity (Ferrari and Ruffini 1981), the gravity-wave emission through an asymmetric neutrino burst in supernova explosions (Epstein 1978), and the emission of gravitational waves through a burst of gravitational waves itself (Christodoulou 1991). A simple understanding of the memory structure comes upon from thinking in terms of Liénard-Wiechert-type gravitational potentials with particles or dust (in the cases of neutrino and gravity-wave bursts: null dust) as sources. A related understanding is given in terms of the Coulomb-type gravitational field of the source (Braginsky and Thorne 1987; Thorne 1992b). For a more

complete discussion, see Wiseman and Will (1991). The memory effects are low frequency effects which originate from the growth of the mean gravity-wave amplitude, i.e. their frequencies correspond roughly to the duration of the burst. In the case of coalescing neutron star binaries, the memory effect results simply from the addition of about 20% of the envelope of the inspiraling part of the wave, e.g. see Fig. 9, to the wave itself (Wiseman and Will 1991).

7 Detectability of Gravitational Waves

The detectors for gravitational waves can be devided into three classes (Thorne 1992a): Earth-based, high-frequency detectors operating in the region 1 Hz to 10 kHz; space-based, low-frequency detectors with main sensitivity between 10^{-5}Hz and 1 Hz; and astronomical detectors in the very low-frequency regime of 10^{-18}Hz to 10^{-5}Hz. To the first class belong bar and beam detectors (e.g. see Blair 1991; contributions by K. Danzmann, A. Rüdiger, and W. Winkler in the present proceedings), to the second class, LAGOS, the "LAser Gravitational-wave Observatory in Space" (e.g. see Stebbins et al. 1989), and Doppler tracking of spacecraft (e.g. see Estabrook and Wahlquist 1975), and to the third class, the timing of millisecond pulsars (e.g. see Stinebring et al. 1990), anisotropies in the cosmic microwave background (e.g. see Zel'dovich and Novikov 1983), and time delays between different images in gravitational lenses (Allen 1989).

The detectors in the third class give already interesting upper bounds on the stochastic gravity-wave background. LAGOS will be very important for the detection of periodic gravitational waves from short orbital-period binaries like i-Boo or even PSR 1534+12 (Wolszczan 1991). The periodic gravitational waves from newborn, and perhaps also older, rotating neutron stars will fall into the domain of the Earth-based detectors. The main importance of these detectors will lie, however, in the detection of gravitational waves from coalescing binaries and Type II supernova explosions.

The Fig.13 shows to be expected gravity-wave amplitudes and an advanced detector–sensitivity curve for a laserinterferometric Earth-based detector. $h_{\rm eff}$ denotes the effective gravity-wave amplitude, $h_{\rm eff} = h\,(n/2)^{1/2}$, where n specifies the number of oscillations near the frequency f and h is the true amplitude.

For coalescing neutron star binaries one expects several events per year in a sphere with radius 200 Mpc (Phinney 1991; Narayan et al. 1991) and for Type II supernovae the corresponding radius is expected to be about 15 Mpc (van den Bergh and Tammann 1991). 15 Mpc is the approximate distance to the Virgo cluster, the cluster nearest to our Galaxy. Because the rates of black–hole collapse and of non–axisymmetrical neutron star formation are still very uncertain the most secure conclusion we can draw

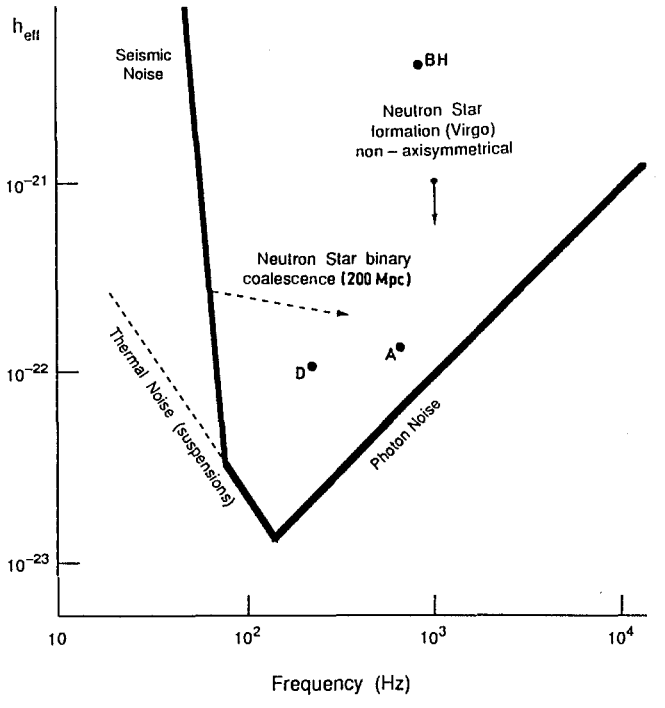

Fig. 13. Expected effective gravity-wave amplitudes and advanced detector sensitivity vs. gravity-wave frequency f. (See also Fig.13 of the contribution by Danzmann et al., these proceedings). The sources of the Models A and D (see Sec.3.1) are located at 1Mpc, about the edge of the Local Group. A $20M_{\odot}$ axisymmetric black–hole collapse at 15Mpc is indicated by BH, see Sec.3.2.

from Fig.13 is: to search for gravitational waves from coalescing neutron star binaries.

References

Allen, B. (1989): Phys. Rev. Lett. **63**, 2017.

Blair, D.G., Ed. (1990): "The Detection of Gravitational Waves", (Cambridge University Press, Cambridge).

Blanchet, L., Damour, T. (1989): Mon. Not. R. astr. Soc. **239**, 845.

Blanchet, L., Damour, T., Schäfer, G. (1990): Mon. Not. R. astr. Soc. **242**, 289.

Bondi, H. (1957): Nature **179**, 1072.

Braginsky, V.B., Grishchuk, L.P. (1985): Soviet Phys. JETP **62**, 427.
Braginsky, V.B., Thorne, K.S. (1987): Nature **327**, 123.
Christodoulou, D. (1991): Phys. Rev. Lett. **67**, 1486.
Cunningham, C.T., Price, R.H., Moncrief, V. (1978): Astrophys. J. **224**, 643; and (1979): Astrophys. J. **230**, 870; and (1980): Astrophys. J. **236**, 674.
Damour, T., Iyer, B.R. (1991): Ann. Inst. H. Poincaré **54**, 115.
Davis, M., Ruffini, R., Tiomno, J. (1972): Phys. Rev.D **5**, 2932.
Detweiler, S.L. (1978): Astrophys. J. **225**, 687.
Detweiler, S.L. (1979): In "Sources of Gravitational Radiation", ed. by L. Smarr (Cambridge University Press, Cambridge), p. 211.
Detweiler, S.L., Szedenits, E. (1979): Astrophys. J. **231**, 211.
Einstein, A. (1916): Preuss. Akad. Wiss. Berlin, Sitzber., p. 688; and (1918): Preuss. Akad. Wiss. Berlin, Sitzber., p. 154.
Epstein, R. (1978): Astrophys. J. **223**, 1037.
Estabrook, F.B., Wahlquist, H.D. (1975): Gen. Rel. Grav. **6**, 439.
Ferrari, V., Ruffini, R. (1981): Phys. Lett. **98B**, 381.
Finn, L.S., Evans, C.R. (1990): Astrophys. J. **351**, 588.
Grishchuk, L.P. (1975): Soviet Physics JETP **40**, 409.
Gunn, J., Ostriker, J. (1969): Phys. Rev. Lett. **22**, 728.
Hillebrandt, W., Wolff, R.G. (1985): In "Nucleosynthesis: Challenges and New Developments", ed. by D. Arnett, J.W. Truran (University of Chicago Press, Chicago), p. 131.
Hils, D., Bender, P.L., Webbink, R.F. (1990): Astrophys. J. **360**, 75.
Ipser, J.R., Managan, R.A. (1984): Astrophys. J. **282**, 287.
Junker, W., Schäfer, G. (1992): Mon. Not. R. astr. Soc. **254**, 146.
Kojima, Y. (1986): In "Dynamical Spacetimes and Numerical Relativity", ed. by J.M. Centrella (Cambridge University Press, Cambridge), p. 379.
Landau, L.D., Lifshitz, E.M. (1941): "Teoriya Polya" (Nauka, Moscow).
Lincoln, C.W., Will, C.M. (1990): Phys. Rev.D **42**, 1123.
Mönchmeyer, R., Müller, E. (1989): In "Timing Neutron Stars", NATO ASI C262 (Kluwer, Dordrecht), p. 549.
Mönchmeyer, R., Schäfer, G., Müller, E., Kates, R.E. (1991): Astron. Astrophys. **246**, 417.
Moncrief, V. (1979): Astrophys. J. **234**, 628.
Müller, E., Hillebrandt, W. (1981): Astron. Astrophys. **103**, 358.
Müller, E. (1982): Astron. Astrophys. **114**, 53.
Nakamura, T., Oohara, K., Kojima, Y. (1987): Prog. Theor. Phys. Suppl. **90**, 1.
Nakamura, T., Oohara, K. (1989): Prog. Theor. Phys. **82**, 1066; and (1991): Prog. Theor. Phys. **86**, 73.
Nakamura, T. and Oohara, K. (1989): In "Frontiers in Numerical Relativity", ed. by C.R. Evans, L.S. Finn, D.W. Hobill (Cambridge University Press, Cambridge), p. 254.
Narayan, R., Piran, T., Shemi, A. (1991): Astrophys. J. **379**, L17.
Novikov, I.D. (1975): Soviet Astron. **19**, 398.
Oohara, K. (1986): In "Dynamical Spacetimes and Numerical Relativity", ed. by J.M. Centrella (Cambridge University Press, Cambridge), p. 365.
Oohara, K., Nakamura, T. (1989): Prog. Theor. Phys. **82**, 535; and (1990): Prog. Theor. Phys. **83**, 906.

Ostriker, J., Gunn, J. (1969): Astrophys. J. **157**, 1395.

Peters, P.C., Mathews, J. (1963): Phys. Rev. **131**, 435.

Petrich, L.I., Shapiro, S.L., Wasserman, I. (1985): Astrophys. J. Suppl. **58**, 297.

Phinney, E.S. (1991): Astrophys. J. **380**, L17.

Piran, T., Stark, R.F. (1986): In "Dynamical Spacetimes and Numerical Relativity", ed. by J.M. Centrella (Cambridge University Press, Cambridge), p. 40.

Piran, T., Nakamura, T. (1988): Prog. Theor. Phys. **80**, 18.

Ruffini, R. (1978): In "Physics and Astrophysics of Neutron Stars and Black Holes", ed. by R. Giacconi and R. Ruffini (North–Holland, Amsterdam), p. 287.

Saenz, R.A., Shapiro, S.L. (1978): Astrophys. J. **221**, 286; and (1979): Astrophys. J. **229**, 1107; and (1981): Astrophys. J. **244**, 1033.

Schutz, B.F. (1986): Nature **323**, 310.

Seidel, E., Moore, T. (1987): Phys. Rev.D **35**, 2287.

Seidel, E., Myra, E.S., Moore, T. (1988): Phys. Rev.D **38**, 2349.

Shapiro, S.L. (1977): Astrophys. J. **214**, 566.

Smarr, L. (1979): In "Sources of Gravitational Radiation", ed. by L. Smarr (Cambridge University Press, Cambridge), p. 245.

Stark, R.F., Piran, T. (1985): Phys. Rev. Lett. **55**, 891; and **56**, 97.

Stebbins, R.T., Bender, P.L., Faller J.E., Hall, J.L., Hils, D., Vincent, M.A. (1989): In "Proc. of the 5th Marcel Grossmann Meeting on General Relativity", ed. by D.G. Blair, M.J. Buckingham (World Scientific, Singapore), p. 1759.

Stinebring, D.R., Ryba, M.F., Taylor, J.H. (1990): Phys. Rev. Lett. **65**, 285.

Taylor, J.H., Wolszczan, A., Damour, T., Weisberg, J.M. (1992): Nature **355**, 132.

Thorne, K.S. (1980): Rev. Mod. Phys. **52**, 299.

Thorne, K.S. (1983): In "Gravitational Radiation", ed. by N. Deruelle, T. Piran (North-Holland, Amsterdam), p.1.

Thorne, K.S. (1987): In "300 Years of Gravitation", ed. by S.W. Hawking, W. Israel (Cambridge University Press, Cambridge), p. 330.

Thorne, K.S. (1992a): In "Recent Advances in General Relativity", ed. by A. Janis, J. Porter (Birkhauser, Boston), p. 196.

Thorne, K.S. (1992b): Phys. Rev.D **45**, 520.

Turner, M.S., Wagoner, R.V. (1979): In "Sources of Gravitational Radiation", ed. by L. Smarr (Cambridge University Press, Cambridge), p. 383.

van den Bergh, S., Tammann, G.A. (1991): Annu. Rev. Astron. Astrophys **29**, 363.

Vilenkin, A. (1981): Phys. Lett.B **107**, 47.

Weber, J. (1960): Phys. Rev. **117**, 307.

Wiseman, A.G., Will, C.M. (1991): Phys. Rev.D **44**, R2945.

Witten, E. (1984): Phys. Rev. D **30**, 272.

Wolszczan, A. (1991): Nature **350**, 688.

Woosley, S.E., Weaver, T.A. (1986): Annu. Rev. Astron. Astrophys. **24**, 205.

Zel'dovich, Ya. B., Novikov, I.D. (1983): "Relativistic Astrophysics Vol. 2. The Structure and the Evolution of the Universe" (University of Chicago Press, Chicago).

The GEO – Project
A Long-Baseline Laser Interferometer for the Detection of Gravitational Waves

K. Danzmann, J. Chen, P.G. Nelson, T.M. Niebauer, A. Rüdiger, R. Schilling, L. Schnupp, K.A. Strain, H. Walther, and W. Winkler

Max-Planck-Institut für Quantenoptik, D-8046 Garching, Germany

J. Hough, A.M. Campbell, C.A. Cantley, J.E. Logan, B.J. Meers, E. Morrison, G.P. Newton, D.I. Robertson, N.A. Robertson, S. Rowan, K.D. Skeldon, P.J. Veitch, and H. Ward

Department of Physics, University of Glasgow, Glasgow, UK

H. Welling, P. Aufmuth, I. Kröpke and D. Ristau

Laser-Zentrum and Institut für Quantenoptik, Universität Hannover, D-3000 Hannover, Germany

J.E. Hall, J.R.J. Bennett, I.F. Corbett, B.W.H. Edwards, R.J. Elsey, and R.J.S. Greenhalgh

Rutherford Appleton Laboratory, Chilton, Didcot, UK

B.F. Schutz, D. Nicholson, and J.R. Shuttleworth

Department of Physics, University of Wales, Cardiff, UK

J. Ehlers, P. Kafka, and G. Schäfer

Max-Planck-Institut für Astrophysik, D-8046 Garching, Germany

H. Braun

Bauabteilung der Max-Planck-Gesellschaft, D-8000 München, Germany

V. Kose

Physikalisch-Technische Bundesanstalt, D-3300 Braunschweig and D-1000 Berlin, Germany

1 Introduction

More than 70 years ago, Gravitational Waves have been predicted as one of the consequences of Einstein's Theory of General Relativity. Einstein describes gravity as a curvature of space-time [1]. When the curvature is weak, it produces Newtonian gravity that we are so familiar with. But a strong curvature behaves in a very different, non-linear fashion. Actually curvature can produce curvature without the aid of any matter. Fast variations of the curvature in time, (due to stellar collapse or collisions, for example) should produce ripples in the fabric of space-time that propagate out at the speed of light and carry the information about the underlying cosmic events. Gravitational waves are clearly one of the fundamental building blocks of our theoretical picture of the universe and there is some circumstantial evidence pointing to their existence [2]. But in spite of numerous attempts over the last 30 years, their direct detection remains as one of the great unsolved problems of experimental physics.

1.1 The Birth of Gravitational Astronomy

Today, the technology seems to be in hand to finally tackle this problem and this article is aiming at outlining the current efforts to bring non-linear gravity into confrontation with experiment through the detection of gravitational waves. But the final aim is not a mere proof of the existence of gravity waves, rather to make them useful for observational astronomy through the creation of a world-wide network of detectors. We have to realize that the information carried by gravitational waves is complementary to the information carried by electro-magnetic radiation. Whereas electro-magnetic radiation is an incoherent superposition of radiation mostly emitted by thermally excited atoms and high-energy electrons, it is the coherent, bulk motion of huge amounts of mass that produces significant levels of gravity waves. Electro-magnetic radiation is easily scattered and absorbed, but gravitational radiation is transmitted almost undisturbed through all forms and amounts of intervening matter [3]. The introduction of Gravitational Astronomy would thus literally open a new window to the universe.

Nobody really knows with certainty how hard it will be to open this window. It is difficult to predict, from our present knowledge based on electromagnetic radiation, just how sensitive a detector has to be to begin to see gravitational waves. But once it does see waves, it will give us information about the universe that we have almost no hope of gaining in any other way. Optical and radio telescopes are sensitive to stellar atmospheres, interstellar dust or primordial gas, all things that a gravitational wave detector will

not see. Instead, we will learn about the inspiral and coalescence of black hole and neutron star binaries, their birth-rates and distribution in distant galaxies, the final collapse of asymmetric supernova cores, "cosmic strings", and the first millisecond of the big bang [4].

2 The Detection of Gravitational Waves

Gravitational waves change the metric of space-time. They can be detected through the strain in space created by their passage. Consider a gravity wave impinging perpendicular to the plane of a circle, see Fig.1. During the first half-cycle of the wave, the circle will be deformed into a standing ellipse and during the second half-cycle into a horizontal ellipse. It is the fractional change in diameter that is commonly quoted as a measure for the amplitude, $h = 2dL/L$, of a gravitational wave. The principle behind the detection of gravity waves is thus a simple length measurement. The problem is that the length change is so small. As an example, consider a supernova in a not too distant galaxy. This might produce a relative length change here on earth of 1 part in 10^{21}. Such a relative length change corresponds to the diameter of one hydrogen atom on the distance from here to the sun, or equivalently to a thousandth of a proton diameter on GEO's 3 km long detector arms. And this happens during a few milliseconds only.

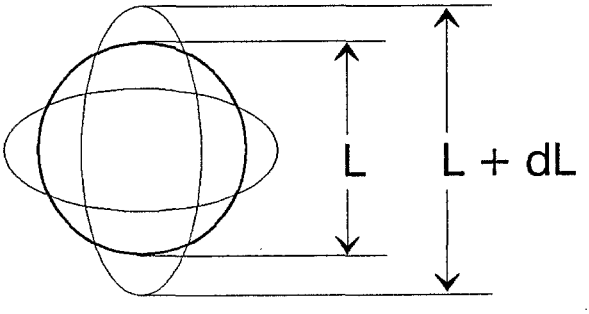

Fig. 1. Gravitational waves change distances by squeezing space.

2.1 Bar Antennas

The history of attempts to detect gravity waves began in the 1960s with the famous bar experiments of Joseph Weber [5]. Bar antennas, in principle, are very simple objects. Imagine a large cylindrical block of e.g. aluminum that during the passage of a gravitational wave gets excited similarly to being struck with a hammer. Even though these experiments have not yet

detected gravitational waves, they had the undisputed effect of alerting the scientific community to the possibility of experimentally detecting gravity waves. Bar detectors have in the meantime been developed and refined in several places all over the world. Being supercooled to mK temperatures and equipped with very sophisticated length transducers [6], they have now reached a sensitivity ($h = 10^{-18}$ for millisecond pulses) where they could expect to see the next supernova in our own galaxy, and they are likely to remain an important ingredient of the world-wide gravity wave-watch. But being resonant devices, they are in practice sensitive only in a relatively narrow band around their central frequency. They are also limited in their sensitivity through the quantum-mechanical uncertainty of their mechanical state, although this limitation may in priciple be overcome by QND techniques, and so their usefulness will in all likelihood remain very limited for the forseeable future.

2.2 Laser Interferometers

Although the seeds of the idea can already be found in early papers by Pirani [7] and Gertsenshtein and Pustovoit [8], it was really in the early 1970s when the idea emerged that laser interferometers might have a better chance of detecting gravity waves, mainly promoted by Weiss [9] and Forward [10]. Large interferometers would offer the additional advantage of having broadband sensitivity and they would not be limited by the uncertainty principle until well below a sensitivity of 10^{-23}. A Michelson interferometer measures the phase difference between two light fields having propagated up and down two perpendicular directions, i.e. essentially the length difference between the two arms. This is exactly the quantity that would be changed by the passage of a properly oriented gravitational wave, see Fig. 2.

Immediately obvious at this point is the need for long interferometer arms. The quantity measured is the absolute phase difference between the fields. But the gravity wave induces a fractional length change. So the phase difference measured can be increased by increasing the armlength or, equivalently, the interaction time of the light with the gravity wave. This works up to an optimum for an interaction time equal to half a gravity wave period. For a gravity wave frequency of 1 kHz this corresponds to half a millisecond or an armlength of 75 kilometers.

2.2.1 Long Light-Path

While it is clearly impractical to build such a large interferometer, there are ways to increase the interaction time without increasing the physical arm-length beyond reasonable limits (see the two following contributions by W. Winkler and A. Rüdiger). Historically, two approaches have emerged: storing the light in the arms in resonant optical Fabry-Perot cavities [11] and

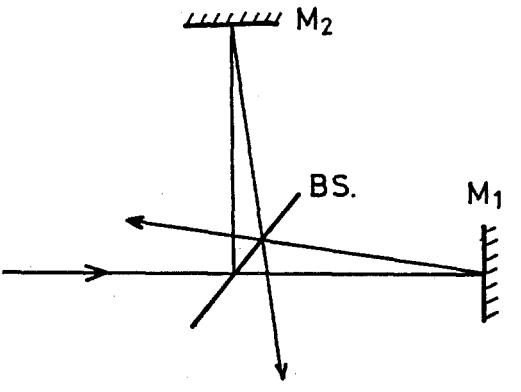

Fig. 2. Michelson interferometer

literally folding the light back and forth in optical delay lines [12]. Nowadays this distinction is beginning to disappear, because with the development of Dual Recycling [13], a new optical technique, to be discussed later, we now have a hybrid arrangement in our hands that combines the advantages of both approaches and more.

2.3 Prototypes

Several small prototypes of laser interferometric gravitational wave detectors have been developed in the world, a delay-line based interferometer with 30 m armlength at the Max-Planck-Institut für Quantenoptik in Garching, a Fabry-Perot based instrument with 10 m armlength at the University of Glasgow, a delay-line based instrument with 10 m armlength at the Institute for Space and Astronautical Science in Tokyo, and a Fabry-Perot based instrument with 40 m armlength at the California Institute of Technology. With arm-lengths on the order of a few tens of meters these prototypes are clearly too small to permit observations of real gravity waves. Their sensitivities have continually been improved over the years, and the larger ones have all reached sensitivities for millisecond pulses roughly equivalent to the best bar detectors, but in addition they are broad-band devices.

While the absolute sensitivity reached by the prototype detectors is certainly encouraging, it is much more important that they are well-understood devices. That is, the various physical processes creating noise sources at the various frequencies have to be identified in order to find ways to improve on those. In Fig. 3 we see such a noise analysis for the Garching 30-m prototype [14].

Shown is the spectral density of the apparent mirror displacement as expected from the most important noise sources. For comparison, the measured spectral density of displacement noise is also given. Very good agree-

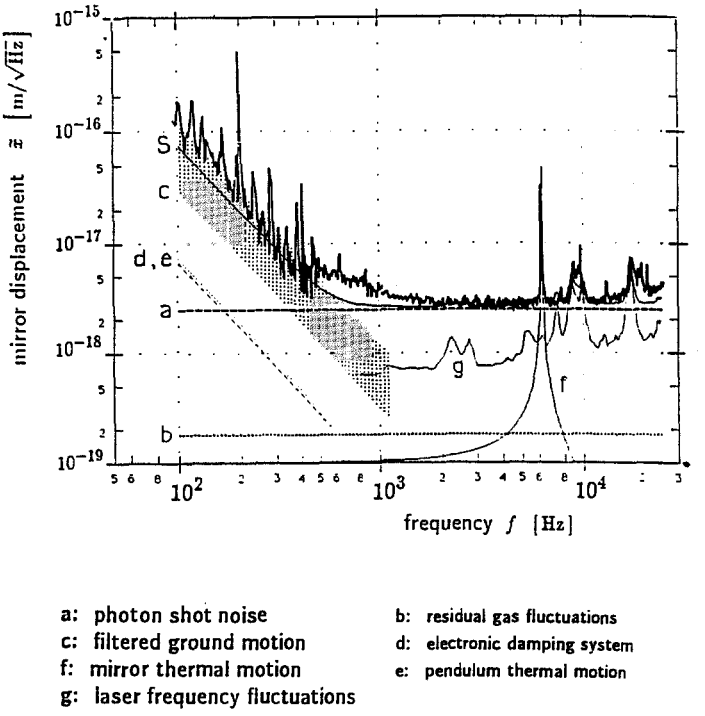

Fig. 3. Noise analysis of the Garching 30-m prototype

ment between the measured and the expected noise is found. The additional sharp peaks at frequencies of a few hundred Hertz are due to violin string resonances of the suspension wires holding the mirrors. The sensitivity of the prototype is presently limited by residual ground motion at frequencies below 1 kHz (labeled c in Fig. 3), by the photon shot noise corresponding to the available laser power at frequencies between 1 kHz and 6 kHz (labeled a), by the thermallly excited internal mechanical vibration of the mirrors in a narrow peak at 6 kHz (labeled f), and by residual frequency fluctuations of the laser at higher frequencies (labeled g). At the present level of sensitivity, the refractive index fluctuations due to the Brownian motion of the residual gas in the vacuum pipe (labeled b) are unimportant. Also unimportant at this level are the above-resonant wing of the thermally excited mirror suspension pendulum resonance (labeled e), the sub-resonant wing of the mirror internal mechanical resonance (labeled a), and the noise introduced by the electronic damping system for the mirror suspension (labeled d, see also Sect. 3.6.1).

3 The GEO Project

After almost two decades of research on small prototypes, the time had come to proceed towards the construction of full-scale interferometers with arm-lengths of several kilometers, and several such proposals were submitted at the end of the 1980s. Out of ten research groups in Germany and Britain, the GEO collaboration was formed [15], aiming at the construction of a laser interferometer with 3 km arm-length near Hannover in the German state of Niedersachsen. The research groups involved are listed at the beginning of this contribution.

In the following, the main problems encountered in the design of such an interferometer and the envisioned solutions are highlighted using the GEO project as an example. But it should be emphasized that the problems are common to all projects and that there actually is a very strong coordination and sharing of tasks between the various collaborations, especially between the German-British GEO project and the French-Italian VIRGO project. An overview of the current status (spring of 1992) of the world-wide efforts is given at the end.

3.1 The Laser Source

The sensitivity of a simple Michelson interferometer with optimized arm-length to gravitational wave bursts is limited by the photon shot noise to

$$h_{DL} \approx 2.4 \times 10^{-21} \left[\frac{\epsilon I_o}{50\,\mathrm{W}}\right]^{-1/2} \left[\frac{f}{1\,\mathrm{kHz}}\right]^{3/2}, \qquad (1)$$

where ϵ is the quantum efficiency of the detector, I_o is the laser output power, and f is the center frequency of the burst. Green light and a bandwidth of half the center frequency have been assumed. The first problem to be solved is thus the construction of a laser with sufficient output power in a stable single transverse and longitudinal mode. Moreover, frequency as well as amplitude of the laser have to be stabilized to unprecedented values.

3.1.1 Output Power

Currently all operating prototypes use Argon ion lasers as light sources. The most powerful commercially available lasers of this type offer a single-mode output power of about 5 W. The coherent addition of several such lasers phase-locked to a master oscillator to reach higher output power has been demonstrated experimentally [16]. But this approach does not seem promising because of the poor energy efficiency of these lasers (only about 10^{-4}), their complexity and high cost of operation and their poor free-running noise that requires elaborate means for stabilization. The laser source under development for the GEO detector is an all-solid state YAG laser pumped

by laser diodes. YAG lasers have traditionally been pumped by discharge lamps, but the dramatic advances in the development of laser diodes in the last few years have now made it possible to replace the noisy and inefficient lamps with diode lasers [17].

This laser will be based on a diode-pumped miniature monolithic ring laser oscillator (see Fig. 4) with an output power of a few hundred milliwatts. The oscillator incorporates an electro-optic phase modulator to permit fast tuning and easy frequency stabilization. The output of this oscillator is then amplified in diode-pumped YAG slabs enclosed in discrete ring resonators. The oscillator is operational and by the end of 1992 the final laser system is expected to deliver a single-mode output power of more than 50 W at a wavelength of 1064 nm. Although the fundamental wavelength of this laser could be used in an interferometer (and there may actually be advantages as far as the fundamental absorption in optical components is concerned) we are investigating the option of doubling the frequency to obtain light in the green at 532 nm. Because of the higher energy per photon in the green only half the power is required to reach the same shot-noise limited sensitivity. Also the optical components can be smaller because the diffraction-limited beam-diameter is smaller in the green by the square-root of two. Doubling efficiencies around 50 percent have been achieved by others for powers around 10 W and much more seems possible [18]. This option will be investigated over the next two years.

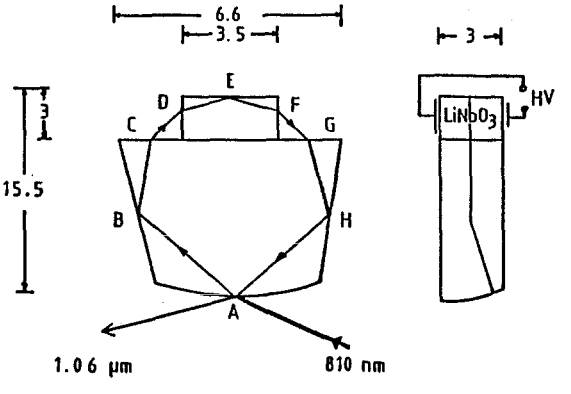

Fig. 4. Monolithic Nd-YAG ringlaser (dimensions in mm)

3.1.2 Frequency Noise

A perfect interferometer is entirely insensitive to frequency fluctuations of the laser. But in a real interferometer, noise signals can be created if at the output of the interferometer there is interference of lightbeams that have a different history. Such a situation can arise if the storage times in the

arms are not identical, or if stray light can reach the output through a path different from that of the main beam. If the relative amplitude of stray light capable of interfering with the main beam is σ, then the arm-length change $\frac{\delta\ell}{\ell}$ simulated by a frequency change $\frac{\delta\nu}{\nu}$ of the laser is

$$\frac{\delta\ell}{\ell} = \sigma\frac{\delta\nu}{\nu}, \tag{2}$$

if stray-light with a path-difference of one full round-trip dominates.

The prototypes use relatively noisy Argon lasers that require sophisticated frequency stabilization techniques. Figure 5 shows as an example the unstabilized frequency noise of the laser previously used on the Garching prototype, curve (a). Curve (b) shows the noise after prestabilization onto a rigid Fabry-Perot reference resonator, and curve (c) shows the noise after final stabilization onto the average armlength of the 30-m interferometer. A lowest value of about $5 \times 10^{-3}\mathrm{Hz}/\sqrt{\mathrm{Hz}}$ is reached in the relevant frequency range. The Argon laser in the Glasgow prototype, being stabilized onto the 10 m long Fabry-Perot cavity in one of the interferometer arms using similar techniques, reaches a frequency noise of a few times $10^{-5}\mathrm{Hz}/\sqrt{\mathrm{Hz}}$.

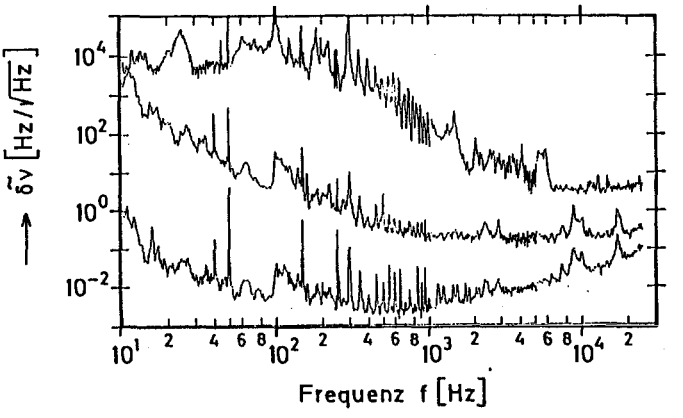

Fig. 5. Frequency noise of the Garching Innova 90-5 laser:
(a) upper curve: unstabilized laser,
(b) middle curve: laser stabilized onto a 25-cm rigid Fabry-Perot,
(c) lower curve: laser stabilized onto the interferometer armlength.

For a full-scale detector, the frequency-noise out of the laser must be smaller than $10^{-6}\mathrm{Hz}/\sqrt{\mathrm{Hz}}$. But for the diode-pumped YAG laser the unstabilized frequency noise is orders of magnitude smaller than for an Argon laser. So achieving the same gain in the feed-back loop as for the Argon laser is already enough to reach the desired stability goal.

3.2 Recycling

Clearly the sensitivity of a simple Michelson interferometer is not sufficient, even if very strong lasers are used. Two techniques have been developed that improve the interferometer sensitivity to a level that would allow the detection of gravitational wave signals with high confidence. These techniques are known as Power Recycling and Signal Recycling, and the combination of both as Dual Recycling [13].

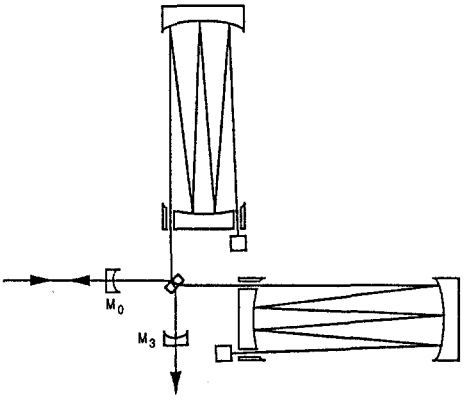

Fig. 6. Dual recycled interferometer

Power Recycling makes use of the fact that the interferometer output is held on a dark fringe by a feed-back loop and almost all the light goes back towards the input, i.e. the locked interferometer behaves as a mirror. By placing a mirror in the input of the interferometer, a resonant optical cavity can be formed that uses the whole locked interferometer as an end mirror. So the circulating light power inside the interferometer will be higher than the laser power by the inverse of the losses in the interferometer.

Signal Recycling works similarly, except that it leads to a resonant enhancement of the signal instead of the light. A gravity wave shaking the mirrors will phase modulate the reflected laser light, or in other words create side-bands of the laser frequency. These side-bands exit through the output port of the interferometer. By placing another mirror there, a resonant cavity for the signal-containing side-bands is formed. Depending on the reflectivity of this mirror, the detector can be made to operate narrow-band or broad-band, and by changing the position of this mirror the interferometer can be tuned. As an additional advantage, this configuration greatly reduces the power losses due to bad interference, because the light can no longer escape the interferometer through the output port, which is now closed by the signal recycling mirror. Just as the signal recycling cavity enhances light at its resonant mode and frequency, it suppresses light

which, through aberrations, got diffracted into non-resonant modes of the light field.

The shot noise-limited sensitivity of a Dual Recycled interferometer to gravitational wave bursts is given by

$$h_{\mathrm{DL}} \approx 10^{-22} \left[\frac{f}{1\,\mathrm{kHz}}\right] \left[\frac{\epsilon I_o}{50\,\mathrm{W}}\right]^{-1/2} \left[\frac{1-R}{5\times 10^{-5}}\right]^{1/2} \left[\frac{\ell}{3\,\mathrm{km}}\right]^{-1/2}, \quad (2)$$

where f is the center frequency of the burst, ϵ is the quantum efficiency of the detector, I_o is the laser output power, ℓ is the arm-length, and R is the mirror reflectivity. Green light and a bandwidth of half the center frequency have been assumed.

3.3 Mirror Losses

In order to make these recycling techniques work, we need mirrors with extremely small losses. This requires substrates with a microroughness on the order of an Angstrom and reflective coatings with very small scatter and absorption. Fortunately, the last few years have brought us tremendous advances in the art of making superpolishes and supercoatings. Mirrors with reflection losses of much less than 50 parts per million are now available from several sources.

But it is not just the linear reflection losses that are responsible for the total losses in the Power Recycling cavity. One of the "mirrors" of this cavity is actually a very complicated object, - an interferometer locked to a dark fringe. So any effect that degrades the interfering wavefronts will let light leak out the wrong port of the interferometer. Though some of this light can be recovered by the Signal Recycling mirror, this is a serious loss process for the Power Recycling cavity.

There are two main reasons for a degradation of the interference in the interferometer: wavefront deformation because of imperfections of optical elements and because of thermal effects in the optical elements. Both processes can be addressed through optical modeling and numerical wavefront propagation calculations.

3.3.1 Thermal Distortions

Thermal effects can arise because of absorption in the optical coatings or in the bulk of elements used in transmission. This will cause a deformation of the substrates and/or a lensing effect through thermally induced refractive index changes. Figure 7 is the result of a model calculation including thermal effects for the main mirrors [19]. It shows the interference minimum as a function of the thermally induced mirror deformation. For increasing light power, the interference quality deteriorates in a very non-linear fashion, even approaching chaotic looking behaviour above a certain threshold.

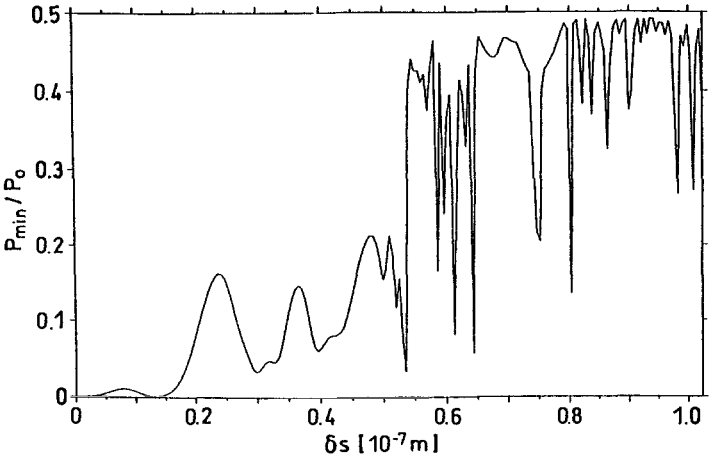

Fig. 7. Power loss due to thermal effects:
interference minimum as a function of local mirror deformation.

We have studied the absorption in coatings produced in ion beam sputtering chambers (one of which we have access to inside the collaboration at LZH). Fortunately, the absorption in modern coatings is only on the order of very few parts per million and thermal effects seem very tractable. Absorption in the bulk of the beam-splitter will probably be the limiting process at circulating powers of many kilowatts.

3.3.2 Mirror Quality

Existing supermirrors with almost negligible losses have been developed for applications using spot sizes of a millimeter or so. For large laser interferometers the beam size will be on the order of several centimeters and the high demands on the surface quality now extend to much larger lateral dimensions. Surface deformations on length scales of several centimeters will behave just as the microroughness does for the smaller beams. But it is not the substrate alone that determines the surface quality. The reflective dielectric multilayer stack coated onto the mirror can show variations of its effective thickness that may overwhelm the surface variations of the substrate.

3.3.3 Optical Modeling

Using a numerical wavefront-propagation code running on the Garching Cray Y-MP, we are studying the effects of surface deformations on the light fields in the interferometer. Currently this code can propagate optical wavefronts on a 4000 x 4000 grid. So it is possible to even include scattering

to angles large enough for the light to miss the end mirror entirely and to hit the vacuum tube after a few hundred meters. This light will then be converted into the diffuse background of scattered light. The code takes as its input assumed or measured surface profiles of the relevant mirrors. Surface deviations from the ideal shape can be measured with today's state-of-the-art to an accuracy of a few Angstroms over a field of a quarter of a meter with a lateral resolution of half a millimeter [20].

As an example, Fig. 8 shows the output field expected for a 3-km interferometer with a mirror distortion of $\lambda/1000$ amplitude on a length scale equal to the beam diameter. The code permits us to simulate all kinds of mirror distortions and to predict the performance of a specific mirror in practice. Through an R+D contract with Zeiss we are addressing the mirror manufacturing problem and we are confident that in 1992 the first prototype mirrors with the required quality will be finished.

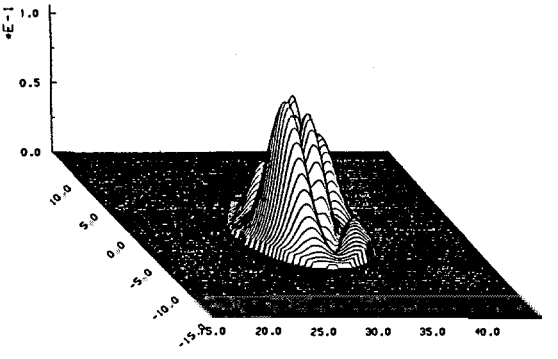

Fig. 8. Output field for mirror with $\lambda/1000$ distortion

3.4 Thermal Noise

3.4.1 Mirror Internal Noise

The mirrors are macroscopic objects and, correspondingly, have internal mechanical vibration modes. Even for perfectly well isolated mirrors, these resonances will still get thermally excited. At the resonant frequencies the mechanical vibration amplitudes are too large and will overwhelm any mirror motion due to a possible gravitational wave. The only solution is to shift all mechanical resonances out of the frequency range of interest by a suitable choice of mirror shape and material. The best compromise is obtained for cylindrical mirrors with a length about half the diameter. For synthetic quartz this yields resonant frequencies of a few kHz, - above the interesting frequency window.

But even though the resonances are outside the observation window, the sub-resonant wings of those resonances do cause a stochastic motion of the mirror surface. The strain spectral density due to these motions is given by

$$\tilde{h} \approx \sqrt{\frac{16kT}{\pi^3 \rho {v_s}^3 Q_{\mathrm{INT}} \ell^2}}, \tag{3}$$

where ρ is the density, ℓ the arm-length, v_s the sound velocity, and Q_{INT} the mechanical quality factor of the mirror material. Note that this expression is independent of mirror size. It is mandatory to use a material with very low internal damping (very high Q). Single crystal Silicon suggests itself for the non-transmitting components with a Q of up to 10^8, but even synthetic quartz gives a Q of a few hundred thousand, whereas low-expansion Zerodur only has a Q of about 1000. Special attention has to be paid to the way of suspending the mirrors, because any way of dissipating energy, like friction of a rubbing suspension wire, will immediately destroy an internal Q as high as this. The problem of material Q as a function of experimental parameters is currently being investigated by us [21].

It should be noted that the noise density due to thermal mirror noise will be strictly constant in frequency only if the internal damping mechanism has viscosity-like behaviour. Sub-resonant thermal noise like this has never been measured directly, but it is highly likely that the mechanical mirror resonance will behave much more like a harmonic oscillator with a complex spring constant [22]. In this case the thermal noise would actually increase from the quoted level towards lower frequencies like the inverse of the square-root of the frequency.

3.4.2 Suspension Noise

Thermal noise is also present in the last stage of the vibration isolation system. This will be a simple wire pendulum made from a sling supporting the mirror. In this case the resonant frequency is about 1 Hz, and the frequency window of observation is on the above-resonant wing. The spectral density of apparent strain noise due to this effect is given by

$$\tilde{h} \approx \sqrt{\frac{16kT\omega_o}{mQ_S\omega^4\ell^2}}, \tag{4}$$

where m is the mirror mass, ω_o the resonant frequency and Q_S the mechanical quality factor of the suspension pendulum. This Q can be much higher than the internal Q of the wire material because most of the energy of the pendulum is in the form of potential and kinetic energy of the swinging bob and not in the elastic energy of a bent wire. But it is extremely important that the wire support points be properly designed to avoid friction. Pendulum Qs as high as 10^7 have been experimentally observed [25],

(meaning that a 1 Hz pendulum will oscillate for more than 4 months before its amplitude has decreased to one-third).

3.5 Vibration Isolation

We are trying to measure very small length changes due to the action of gravitational waves and so it is extremely important to ensure that no other effect moves the test masses. By far the largest disturbance is the random ground motion because of the natural seismic activity. The spectral density of displacement due to seismic ground motion typically falls of as

$$\tilde{x}(f) \approx 10^{-7}\mathrm{m}/\sqrt{\mathrm{Hz}}\left[\frac{1\mathrm{Hz}}{f}\right]^2. \qquad (5)$$

At a frequency of 1 kHz, this is about 10^7 times larger than the effect we are trying to measure and we require an effective way of vibration isolating the test masses.

Passive vibration isolation is, in principle, straightforward [23]. The object to be isolated gets suspended by a pendulum. If a disturbance shakes the suspension point of the pendulum with a frequency below its resonant frequency, then the pendulum will transmit the disturbance unattenuated. But a disturbance above the resonant frequency f_o will get attenuated with the ratio $(f_o/f)^2$ up to a frequency Qf_o, where the frequency dependence changes to a linear slope. The resonant frequency f_o should thus be as low as possible. Due to practical limitations on the pendulum length to around one meter, horizontal resonant frequencies are usually limited to around 1 Hz. Vertical resonant frequencies are normally a bit higher because the spring has to be stiff enough to support the full load. Several of these stages can be cascaded to achieve a very steep fall-off above the highest normal mode of the coupled oscillator system.

3.5.1 Stacks

A very simple, yet very effective way of achieving vibration isolation at moderately high frequencies (above 100 Hz) is the use of vibration isolation stacks. A detailed analysis of stack systems has recently been carried out by Cantley et al [26]. Stacks are basically alternating layers of a high-density material (like lead) and rubber. Each layer acts as a 3-dimensional pendulum, although with a fairly low Q. But since it is easy to use several such layers, a rather steep fall-off towards higher frequencies can be achieved. As an example, Fig. 9 shows a comparison between calculated and measured transmissibility of a small 4-layer lead-rubber stack intended for the Garching prototype.

Because of their low mechanical Q-factor, such stacks show a rather high thermal noise. The last two stages of the suspension should accordingly be very high-Q wire pendulums of low resonant frequency. These will

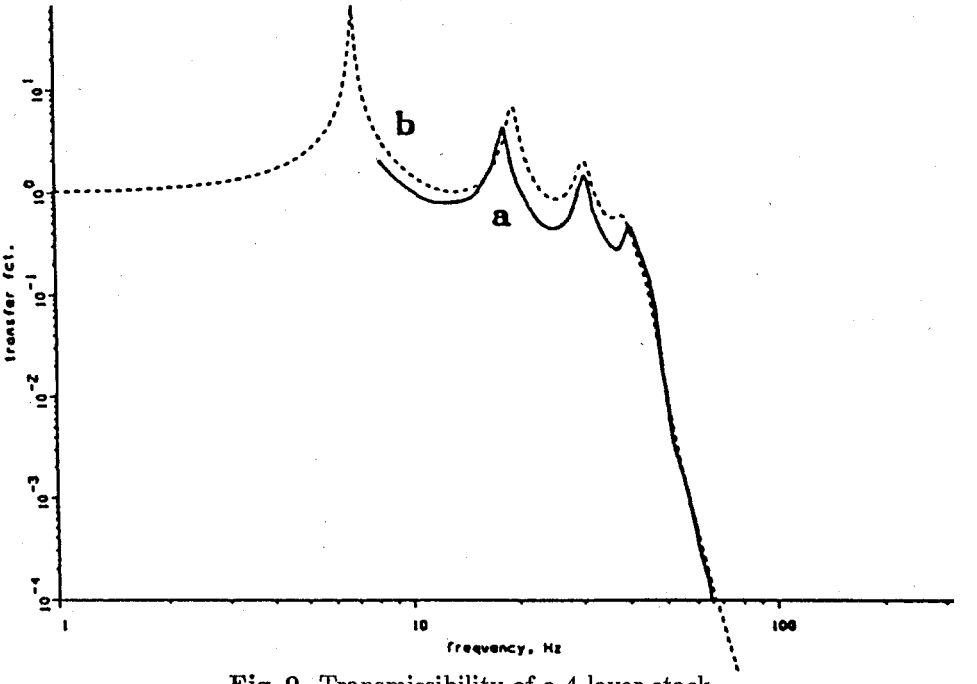

Fig. 9. Transmissibility of a 4-layer stack

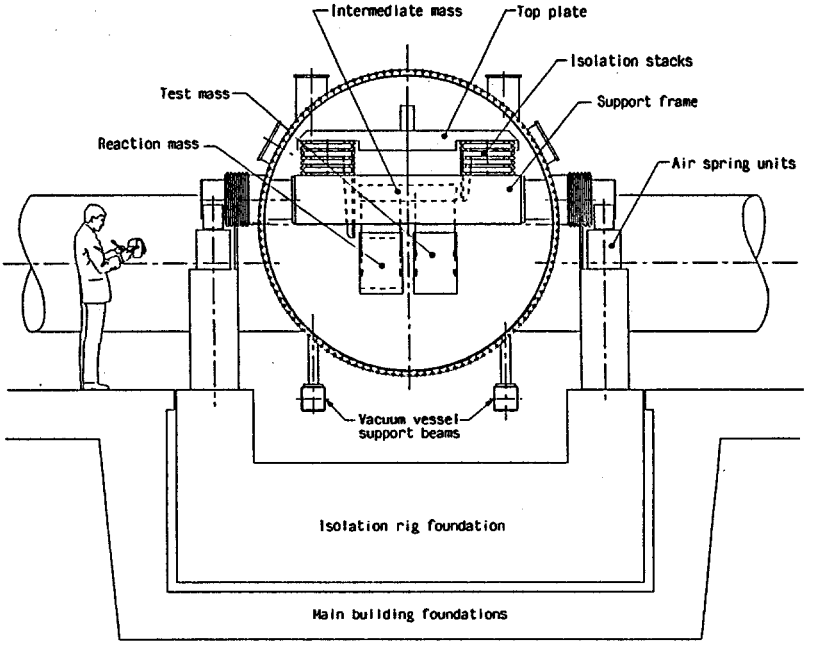

Fig. 10. GEO suspension

also give an additional $1/f^4$ filtering in the horizontal direction where it is most critical (this is the direction of the laser beam; in principle, vibration isolation in the vertical direction is only required because of the unavoidable cross-coupling between the two degrees of freedom).

The design chosen for the GEO suspension is shown in Figs. 10 and 11. It consists of an active airspring system as a first stage to give some very low-frequency isolation. Supported by this system is a five-layer isolation stack supporting the top plate. Each of the layers has a horizontal resonant frequency of 10 Hz, a vertical resonance at 30 Hz and a Q of about 2. The top plate is a hollow structure filled with damping material to suppress structural resonances. The mirror is suspended from the top plate through a double wire pendulum with an intermediate mass. Each of these pendulums has a resonance at 1 Hz. Some small components that require fast feedback (see below) are paired up with reaction masses. The total isolation as expected from model calculations should be more than sufficient at all frequencies above 100 Hz.

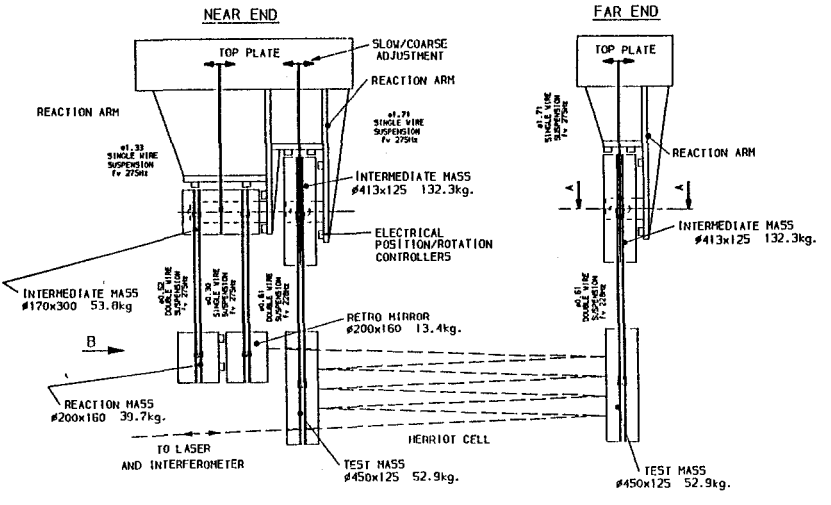

Fig. 11. GEO suspension

3.5.2 Low-Frequency Isolation

Extending vibration isolation down to lower frequencies becomes increasingly difficult. The ground noise spectrum increases towards lower frequencies like $1/f^2$ and the isolation decreases because one is moving closer to the highest normal mode of the pendulum chain. Some very promising work on low frequency passive isolation has been done in Pisa [34]. The 12 m-high Superattenuator uses a cascade of an inverted pendulum, 7 gas springs, and

a wire pendulum to achieve efficient vibration isolation above 10 Hz. If the control problems associated with this approach can be solved, it could be an alternative to the use of stacks.

Another way of achieving isolation at very low frequencies would be the use of active or combined active/passive vibration isolation systems. Some very promising work, aiming at achieving isolation all the way down to 1 Hz, is going on at the Joint Institute for Laboratory Astrophysics in Boulder, Colorado [28].

3.6 Position Control and Feed-Back

An interferometer with as many components and degrees of freedom as a gravitational wave detector requires sophisticated control systems to keep all parameters at their optimal operating points. The guiding principle behind the position control of the optical components is easily stated: At low frequencies the optical components must be rigidly held relative to each other to keep them from drifting and to prevent the interferometer and the recycling cavities from losing lock. On the other hand, at higher frequencies, where gravitational wave signals could be detected (above 100 Hz), the test masses must be totally free and the system used to prevent them from moving at low frequencies must not exert any residual forces at high frequencies. Three main classes of control systems are used:

3.6.1 Local Controls

The suspension of the optical components via high-Q pendulums is an efficient way to isolate them from high-frequency vibrations and pendulum thermal noise. But at the resonant frequencies of the undamped suspension, large vibration amplitudes can occur that will greatly exceed the dynamic range of the detector (a pendulum with a Q of 10^7 can, at its resonance, lead to an amplification of up to 10^7). The resonance must thus be damped. But damping it in the usual dissipative way will degrade the Q and introduce thermal noise. The damping is instead done in an active and frequency-selective way. This technique is routinely used on the prototypes. The position of the masses is sensed with a low-noise local sensor. This signal is then electronically filtered to a narrow frequency range around the resonance and then fed back to a force transducer acting on the mass selectively in this band around the resonant frequency only. On the prototypes the position sensing is usually done via shadow sensors consisting of a LED-photodetector pair with the light path partly interrupted by a movable vane mounted on the test mass. The force is applied through a coil acting on a magnet on the mass. Such systems typically show a sensing noise of around $10^{-11}\mathrm{m}/\sqrt{\mathrm{Hz}}$. In the full scale detector we are trying to measure displacements smaller than $10^{-20}\mathrm{m}/\sqrt{\mathrm{Hz}}$. So the servo gain would have to roll off by

9 orders of magnitude in the small frequency range from a few Hz to 100 Hz, which is clearly a formidable task. The problem can be solved by sensing the motion of, and applying feed-back to, a higher stage in the suspension system and using the passive $1/f^2$ filtering of each stage. Such systems are currently being tested in the prototypes.

3.6.2 Global Controls

In order to optimally align an optical interferometer, and keep it aligned regardless of drift or stability of the optical components, some kind of automatic alignment system is required. The guiding principle is that the alignment signals for such a system should be derived directly from the existing interfering beams without introducing additional components into the high-sensitivity/high-intensity part of the interferometer. Suitable feed-back should then be applied to all the relevant optical components.

For example, consider the case of two interfering beams, where a differential high frequency phase modulation is applied and the overall phase difference is determined by coherently demodulating the intensity of the interfered output. Relative angular misalignment introduces a differential phase gradient between the two beams which can be sensed using a split photodiode and coherently demodulating as before. Lateral misalignment may be detected using another split photodiode and a suitable lens arrangement to cause laterally offset beams to converge.

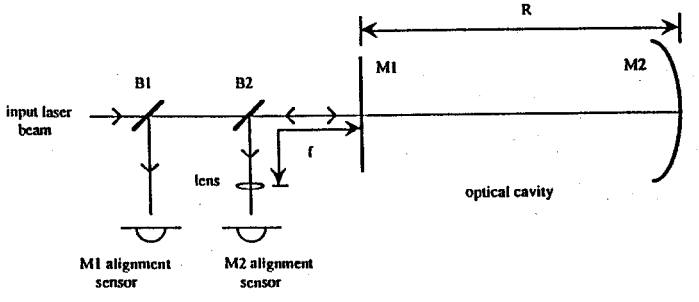

Fig. 12. Automatic alignment of a Fabry-Perot cavity

An illustration of this techniques for a Fabry-Perot cavity is given in Fig. 12. All 4 degrees of freedom necessary for alignment can be extracted. An extension of this technique can be used to automatically align all degrees of freedom for all components of the interferometer. A system similar to the automatic alignment system intended for the large interferometer is being

installed and tested in the Glasgow prototype [27] and will be installed in the Garching prototype after it is reconstructed.

3.6.3 Fast Feed-Back

While high-frequency ground motions can be efficiently suppressed by the suspension system, there are very-low frequency seismic motions (1 Hz and less) that are very difficult to isolate against. This very-low frequency seismic motion may lead to rms differential arm-length changes of several microns. Unsuppressed, these would lead to a severe limitation of the sensitivity of the detector, because an interferometer operating away from its null fringe by an amount δx becomes sensitive to the intensity fluctuations δI of the laser according to

$$h_{\text{noise}} = \frac{\delta x}{\ell} \frac{\delta I}{I} \tag{6}$$

The sensing in this case is no problem, because the main interferometer output itself provides the signal. But these deviations from the null have to be suppressed by at least 10^6 which requires a fast servo with a bandwidth of order a kHz. These correction signals will be applied to the small optical components in the 3-km interferometer such as beam-splitter and retro-mirror. This feed-back is best applied relative to a reaction mass suspended from the same isolation system to avoid coupling ground motion back in through the actuator (see Fig. 11).

3.7 Vacuum System

At the required sensitivity, the whole detector must operate in a high vacuum to reduce noise from gas molecules. The limiting criterion is the fluctuation in the refractive index due to changes in the average density of the gas through which the laser beams pass, resulting in phase shifts of the interfering light beams. If these changes occur in the frequency window of observation (from a few tens of Hz to a few kHz) they can mask gravitational events.

Pressure fluctuations can arise from the random motion of the gas molecules in the vacuum tubes and chambers. The amplitude of a fluctuation is proportional to the square root of the pressure. To reach the proposed sensitivity of the detector, a pressure smaller than 10^{-8}mbar is required for Hydrogen. Since other gases have higher refractive indices, the sum of their partial pressures should be smaller than 10^{-9}mbar. Finally, the system should be hydrocarbon free as far as technically possible.

Pressure fluctuations can also arise from changes in the pumping speed. There are mechanisms which could cause fluctuations in pumping speed for most UHV pumps. Ion pumps have noisy discharge currents, cryogenic pumps are likely to emit bursts of gas and even turbomolecular pumps can

show variations of their pumping speed due to asymmetries and fluctuations in the rotation or variations in the backing line. We are currently investigating the subject of fast pressure bursts due to vacuum pumps.

However, it is difficult to think of mechanisms for fluctuations in the pumping speed of non-evaporable getter (NEG) pumps. Also, their operation is completely vibration-free. As a result, 21 NEG pumps of 14 000 l/s capacity each have been chosen in our conceptual design to provide the main UHV pumping [24].

The vacuum tubing carrying the laser beams up and down the two arms will have a total length of 6 km and a diameter of 1.4 m and will be one of the largest UHV systems in the world. It will be one of the most costly elements of the detector. It is thus worthwhile to look for unconventional designs to save costs on this component. We are proposing to use a thin-walled (0.7 mm) tube made from 316L stainless steel sheet material. To provide stability, it would have a continuous corrugation with a height of 40 mm peak-to-peak over the whole length of the tube. Such a tube, if continuously manufactured and welded on site, would cost only a small fraction of a more traditional thick-walled and flanged tube. Vacuum tests on a 5 m-long test section showed very promising results, especially a very low outgassing rate of 3×10^{-13} mbar l/s/cm^2 after baking it at only 150 C [24].

Before deciding on this design for the final detector, more tests have to be done on a much longer test section manufactured under field conditions. These tests are currently being prepared by the European research groups.

4 Detector Realization

The proposed detector consists of two perpendicular arms, each of 3 km length. The vacuum system is designed to accommodate two simultaneously operating interferometers which can be optimized to different parameters. Commissioning is planned to proceed in steps from a simple Michelson interferometer to a full Dual-Recycled system and will take several years after the installation of the first interferometer. After reaching a burst sensitivity of 10^{-21} we expect to split the available detector time roughly equally between observation runs and experimental work to push the sensitivity further towards a value of 10^{-22}. The significance of the various noise sources for the sensitivity of a fully optimized interferometer to broad-band gravitational wave bursts is shown in Fig. 13. The curve labeled seismic noise is valid for the described stack system. With a Pisa-type suspension, the curve would be moved to the left edge of the figure and the low-frequency sensitivity would be limited by thermal suspension noise. The thermal noise in Fig. 13 has been calculated assuming viscosity-like internal damping for mirrors and suspension.

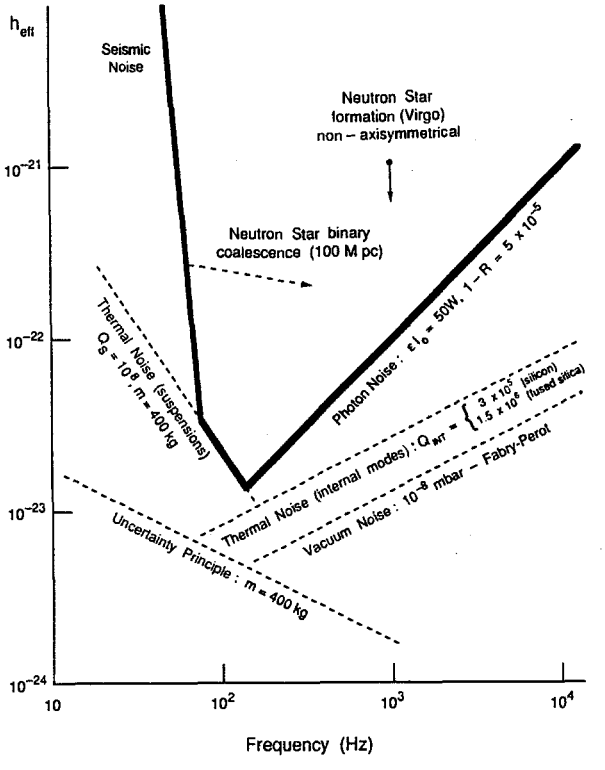

Fig. 13. Noise sources relevant for the detector

The site selection survey was finished in 1991 and two suitable sites were identified near Hannover in the German state of Niedersachsen. One of them is a flat, dry and uninhabited piece of land owned by the state government. Here, the tunnel housing would be a concrete structure with 3 m internal height and a width of 4 m submerged in a trench close to the surface. The control building as well as the laboratory buildings containing the vacuum tanks and the optics would be conventional structures above ground. The other possible site is a mountain site also owned by the state government. Here, the arms would be in a 4-m diameter underground tunnel cut through bedrock. The vacuum tanks and optics at the vertex and the ends of the arms would be in 35-m diameter underground caverns, but each of the three corner caverns would be accessible from the outside through short horizontal access tunnels. Only the control building would be visible above ground. Geological, hydrological and seismological investigations found both sites very suitable; the mountain site having an advantage with respect to seismic noise, of course. Costings and architectural designs have been prepared. Construction at the underground site has been found to be more expensive

by about 15 percent. The final decision between the two sites will be made in 1992, largely based on necessary additional environmental impact studies. An impression of the main control building is given in Fig. 14.

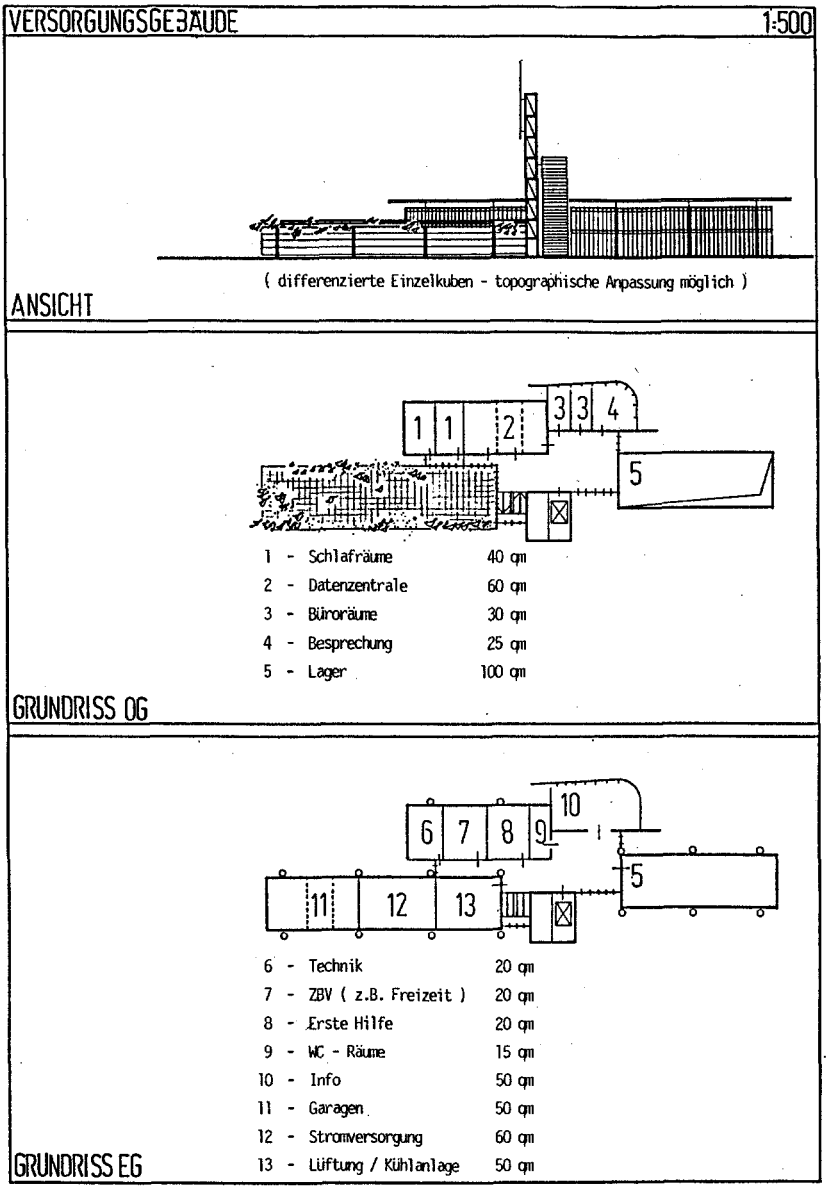

Fig. 14. GEO main building

5 Status of Efforts in the World

5.1 Proposals

The American LIGO proposal [29] calls for the construction of two detectors with 4 km arm-length. It was approved in the fall of 1991 and funds have been appropriated by Congress. Two sites were selected in the spring of 1992, one in Hanford, Washington and the other in Livingston, Louisiana. Construction is expected to begin in 1993 and commissioning may start as early as 1997.

An Australian proposal (AIGO) for a 3-km detector [30] has been submitted, but has not been able to obtain approval.

In Japan, a proposal for an intermediate 100-m interferometer (TENKO-100) [31] has been funded. Construction may be finished in 1994. In parallel, plans are being developed for a 3-km interferometer.

The French-Italian VIRGO collaboration [32] has proposed a 3-km interferometer to be built near Pisa and the German-British GEO collaboration [15] has proposed a 3-km interferometer to be built near Hannover. Both interferometers are being coordinated under the EUROGRAV framework. A decision from all the relevant governments in Europe is expected in the summer of 1992.

5.2 A World-Wide Network

All these projects are not in competition with each other. On the contrary, each of the projects is crucially dependent on the others. To sort out gravitational wave events from the ever-present noise background requires observation in coincidence of several detectors. So two gravitational wave detectors are the absolute minimum to even prove the existence of gravitational waves. But to fully unravel the information contained in the signals with respect to the source direction, time structure and polarization requires a world-wide network of four detectors [33].

If all goes well, this network can be in place by the end of this decade, and at the beginning of the next millenium we may be able to mark the beginning of the age of Gravitational Astronomy.

References

1. C.W. Misner, K.S. Thorne, and J.A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973).
2. J.H. Taylor and J.M. Weisberg, Astrophys. J. **345**, 434 (1989).
3. K.S. Thorne in *Gravitational Radiation*, edited by N. Deruelle and T. Piran (North Holland, Dordrecht, 1983), pp. 1-54.
4. K.S. Thorne in *Recent Advances in General Relativity*, edited by A. Janis and J. Porter (Birkhauser, Boston, 1992), pp. 196-229.

5. J. Weber, Phys. Rev. **117**, 306 (1960).
6. P.F. Michelson, J.C. Price, R.C. Taber, Science **237**, 150 (1987).
7. F.A.E. Pirani, Acta Physica Polonica **15**, 389 (1956).
8. M.E. Gertsenshtein and V.I. Pustovoit, JETP **16**, 433 (1963).
9. R. Weiss, Quarterly Progress Report of RLE, MIT **105**, 54 (1972).
10. G.E. Moss, L.R. Miller, R.L. Forward, Applied Optics **10**, 2495 (1971).
11. R.W.P. Drever in *The Detection of Gravitational Waves*, edited by D. Blair (Cambridge University Press, Cambridge, 1990), pp. 306-328.
12. W. Winkler, ibid.
13. B.J. Meers, Phys. Rev. D **38**, 2317 (1988); K.A. Strain and B.J. Meers, Phys. Rev. Lett. **66**, 1391 (1991).
14. D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger and A. Rüdiger, Phys. Rev. D **38**, 423 (1988); T.M. Niebauer, R. Schilling, K. Danzmann, A. Rüdiger, and W. Winkler, Phys. Rev. A **43**, 5022 (1991).
15. J. Hough et al., *Poposal for a Joint German-British Interferometric Gravitational Wave Detector*, Max-Planck-Institut für Quantenoptik, Report No. MPQ 147 (1989), unpublished.
16. G.A. Kerr and J. Hough, Appl.Phys. B **49**, 491 (1989).
17. I. Schütz, H. Welling and R. Wallenstein, in *Advanced Solid State Lasers* (Salt Lake City, 1990).
18. R.L. Byer, private communication.
19. W. Winkler, K. Danzmann, A. Rüdiger, and R. Schilling, Phys. Rev. A **44**, 7022 (1991).
20. K. Freischlad, M. Küchel, W. Wiedmann, W. Kaiser, and M. Mayer, Proc. SPIE **1332**, 8 (1990).
21. J.E. Logan, N.A. Robertson, J. Hough, and P.J. Veitch, Phys. Lett. A **161**, 101 (1991).
22. P.R. Saulson, Phys. Rev. D **42**, 2437 (1991).
23. N.A. Robertson in *The Detection of Gravitational Waves*, edited by D. Blair (Cambridge University Press, Cambridge, 1990).
24. J.R.J. Bennett and R.J. Elsey, Vacuum **43**, 35 (1992).
25. W. Martin, Ph.D. Thesis, Glasgow (1978), unpublished.
26. C. Cantley et al., Rev. Sci. Instr. **63**, 2210 (1992).
27. E. Morrison et al., to be published.
28. P.G. Nelson, Rev. Sci. Instr. **62**, 2069, 1991.
29. R.E. Vogt, R.W. Drever, F.J. Raab, K.S. Thorne, and R. Weiss, *Laser Interferometer Gravitational-Wave Observatory*, proposal to the National Science Foundation, (California Institute of Technology, December 1989), unpublished; A. Abramovici et al., Science (1992), in press.
30. R.J. Sandeman, D.G. Blair, and J. Collett, *Australian International Gravitational Research Centre*, proposal to the CRC, (Australian National University, 1991), unpublished; D.E. McClelland et al. in *Gravitational Astronomy*, edited by D.E. McClelland and H.A. Bachor (World Scientific, Singapore, 1991).
31. N. Kawashima in *Gravitational Astronomy*, edited by D.E. McClelland and H.A. Bachor (World Scientific, Singapore, 1991); M. Fujimoto, M. Ohashi, N. Mio, and K. Tsubono, ibid.

32. C. Bradascia et al. in *Gravitational Astronomy*, edited by D.E. McClelland and H.A. Bachor (World Scientific, Singapore, 1991).
33. Y. Gürsel and M. Tinto, Phys. Rev. D **40**, 3884 (1990).
34. R. DelFabbro et al., Rev. Sci. Instr. **59**, 292 (1988); C. Bradascia et al., Phys. Lett. A **137**, 329 (1989).

The Optics of an Interferometric Gravitational-Wave Antenna

W. Winkler and colleagues of GEO [1]

Max-Planck-Institut für Quantenoptik, D-8046 Garching, Germany

[1] *Names and affiliations of all authors are given at the end of the paper*

Abstract: The basic concept of an interferometric gravitational wave detector, the realization of the long light path with optical delay lines or with Fabry-Perot cavities, and the need for high light power are described. The techniques for improving the sensitivity, recycling and squeezed states of light, are considered and the consequences on the specifications of the optical components are shown. The specifications are explicitly given and particularly the influence of thermal effects is treated quantitatively.

1 Introduction

All realistic sources for gravitational waves that have been thought of so far are expected to provide us with extremely small signals. The most efficient emission of radiation from supernova core implosions or from coalescing binaries lasts only a few milliseconds, and for an event rate of several per month the strain in space introduced by gravitational waves (that is a change in relative distance between testparticles) may be on the order of only $h = \frac{1}{2}\frac{\delta L}{L} \approx 10^{-22}$.

As described in the contribution of K. Danzmann in this issue, a Michelson interferometer is an adequate tool to look for such tiny strains. For an optimization of the sensitivity one has to optimize both – the effective light path L corresponding in some sense to the interaction time between the gravitational wave and the light inside the interferometer – and the resolution for the path difference δL. With techniques available today it should be possible to reach the required sensitivity level, as we will see in the following sections.

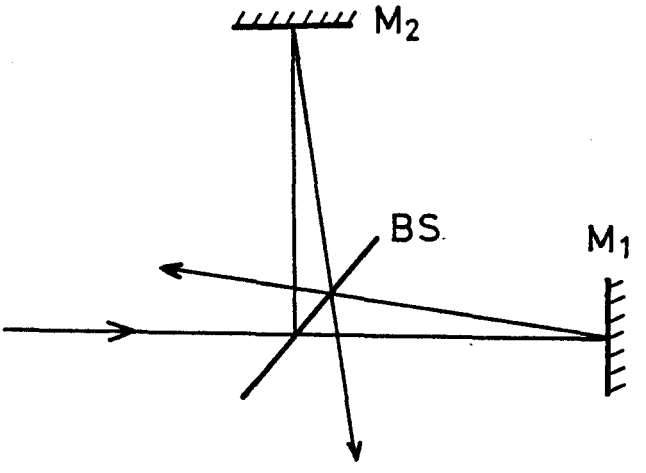

Fig. 1. Michelson interferometer with two mirrors M_1 and M_2 and a beam-splitter BS

2 The long Light Path

A gravitational wave passing the interferometer introduces a strain in space with opposite sign in two orthogonal directions. The path difference δL to be measured in a Michelson interferometer is a maximum if the storage time of the light in the arms is matched to the period of the gravitational wave. For periods of milliseconds the optical light path is optimally of the order of 100 km. In order to realize this long light path the light is sent back and forth in each arm several times before it is superposed with the other beam at the beam-splitter. The number of beams cannot be chosen very high, otherwise residual mirror motions like the thermally driven eigenmodes of the mirror substrate would limit the sensitivity of the antenna. An armlength of a few kilometers seems to be a good choice as a compromise between optimal performance of the interferometer and financial investment.

In the prototypes of gravitational wave antennas built and tested in the past, two possibilities for realizing long light paths have been investigated: optical delay lines and Fabry-Perot cavities.

2.1 Fabry-Perot Cavities

In optical cavities the many beams running between the terminating mirrors are all superposed to form one intense beam [1]. The far mirror is coated for highest possible reflectivity, whereas the transmission of the near mirror determines the effective path length inside the cavity:

$$L_{\text{eff}} \approx \frac{\ell}{T}, \tag{1}$$

where ℓ denotes the mirror separation and T the power transmittance of the coupling mirror. Thus, the effective path length inside the cavity is adjustable by the proper transmittance of the input mirror. Ideally all of the light eventually goes back to the beam-splitter. Variations in path length result in a phase shift which is detected by superposition with the light from the other arm.

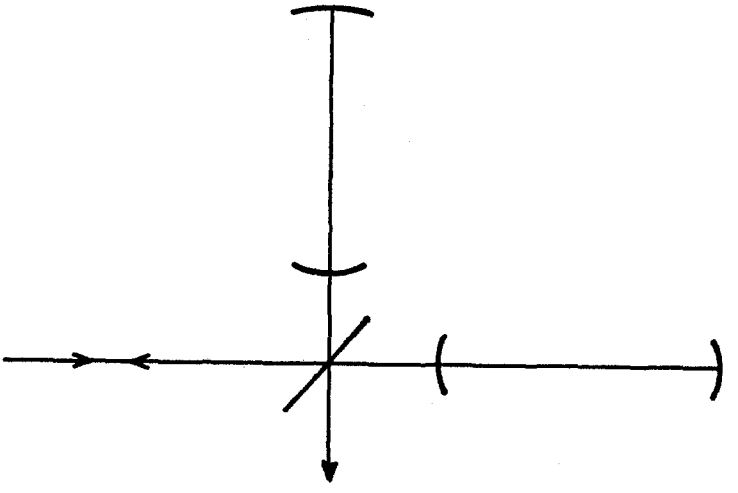

Fig. 2. Michelson interferometer with Fabry-Perot cavities

A cavity has to be designed according to the properties of the light beam, in our case a diffraction limited laser beam. Such a Gaussian beam is already determined by position and diameter $2w_o$ of its focus. The beam radius w is defined by the distance between the beam axis and the point where the intensity is e^{-2} of its maximum value.

There is a characteristic length b, the so-called Rayleigh range, for each Gaussian beam, defined by

$$b = \frac{2\pi}{\lambda} w_o^2, \tag{2}$$

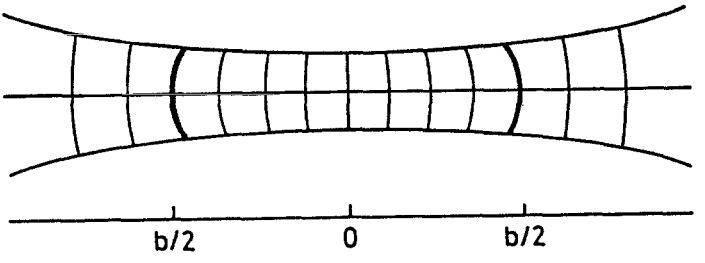

Fig. 3. Wavefronts in a diffraction limited laser beam

with λ the wavelength of the light. The radius w in the radial intensity distribution

$$I = I_o \exp(-\frac{2r^2}{w^2}) \tag{3}$$

varies with distance z to the focus according to

$$w = w_o \left(1 + \left(\frac{z}{b/2}\right)^2\right)^{1/2} . \tag{4}$$

The equivalent relation for the radius of curvature R_w of the wavefront reads as

$$R_w = z \left(1 + \left(\frac{b/2}{z}\right)^2\right) . \tag{5}$$

b is the smallest radius of curvature of the wavefronts in a beam with focal diameter $2w_o$, occurring at a separation of $b/2$ from the focus.

A cavity can be formed by inserting mirrors in the light path with a curvature of the surface equal to the curvature of the wavefront. The beam is then reproduced by the subsequent reflections. A possible arrangement could be a flat mirror in the focus and a concave mirror at a distance z, with radius of curvature corresponding to the last equation. The average beam size is smallest in a near confocal cavity, that is a symmetric arrangement with radii of curvature close to the mirror separation:

$$2w = 2\left(\frac{\lambda \ell}{\pi}\right)^{1/2} . \tag{6}$$

For visible light and km-dimensions the beam size as defined above is a few cm. In order to avoid diffraction losses, that is parts of the beam falling off the mirror edge, one has to assign a space to each reflection by a factor S bigger in diameter than the beam diameter. Usually $S \approx 3$ is assumed to be

sufficient. For more details see for instance the contribution of R.W.P. Drever in [2].

2.2 Optical Delay Lines

The second possibility for a multireflection scheme to realize the long light path is an optical delay line [3]. Here the subsequent reflections are more or less well separated from each other, and the light path is well defined as $L = N\ell$ with N the number of beams and ℓ the mirror separation. The simplest case of a delay line is formed by two equally curved spherical mirrors facing each other at a separation ℓ equal to the radius of curvature R. This is the so called confocal separation, as the focal points of the two mirrors coincide.

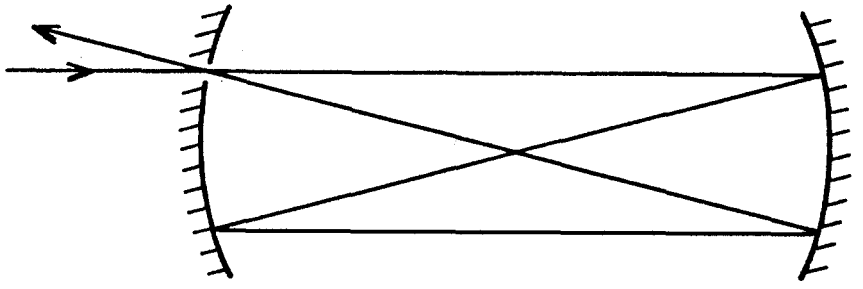

Fig. 4. Confocal delay line

The beam usually enters the delay line through a coupling hole in the near mirror. The far mirror produces an image of the coupling hole onto the near mirror symmetrically to the optical axis. (The optical axis can be defined as the line connecting the centers of curvature of the two mirrors). This image can now again be considered to represent an object, which in turn is imaged by the far mirror into the coupling hole – independent of the position of the coupling hole and the orientation of the input beam. In order to get more than four beams, the mirror separation is changed by $\Delta\ell$. As a result, the fourth beam is shifted with respect to the coupling hole by an amount proportional to $\Delta\ell$, hits the reflecting surrounding of the hole and starts a new round trip. For a particular value of $\Delta\ell$ the beam falls into the coupling hole after N reflections. Usually the reflection spots are circularly arranged by a proper orientation of the input beam (see Fig. 5).

The coordinates of the reflections at the mirrors obey the following equations:

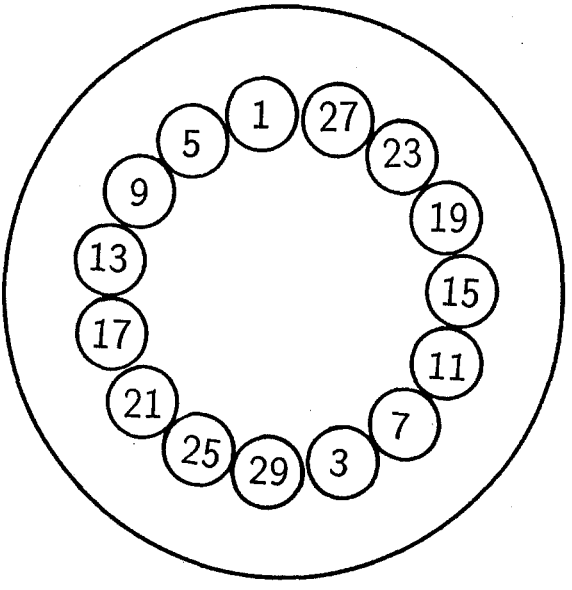

Fig. 5. Circular arrangement of reflection spots at one delay line mirror for $N = 30$ beams

$$x_n = x_o \cos n\Theta + \sqrt{\frac{\ell}{2R-\ell}}\,(x_o + Rx'_o)\sin n\Theta \tag{7}$$

$$y_n = y_o \cos n\Theta + \sqrt{\frac{\ell}{2R-\ell}}\,(y_o + Ry'_o)\sin n\Theta \tag{8}$$

(x_o, x'_o) and (y_o, y'_o) are position and slope of the input beam in x- and y-direction, respectively.

In a circular arrangement the subsequent reflections seem to be rotated by Θ from one mirror to the next, whereas the reflections at one mirror are rotated by 2Θ. Θ is defined in the usual way [3] by:

$$\cos\Theta = 1 - \frac{\ell}{R}. \tag{9}$$

(For the confocal mirror separation we have $\Theta = \pi/2$). The mirror separation can be chosen such that after N reflections the beam arrives at the same spot where it entered the delay line, in our case the coupling hole. This is the so-called reentrance condition. It is fulfilled if $N\Theta$ is equal to an integer multiple of 2π.

Delay lines have some practical advantages: position and orientation of the output beam are independent of rotations and tilts of the far mirror.

Thus the interference quality is not spoiled by angular motions of that mirror which may be kilometers away. For practical reasons it is sometimes desirable to have a well defined optical path length – for instance to tune the storage time to a particular signal frequency, or to equalize the light paths in the two arms in order to be less susceptible to frequency fluctuations. In a delay line the optical path is $N\ell$; as the mirror separation is rigidly connected to the radii of curvature of the mirrors, it is given with the same precision as these. In a cavity the effective light path depends on the reflection losses, and these are more difficult to control.

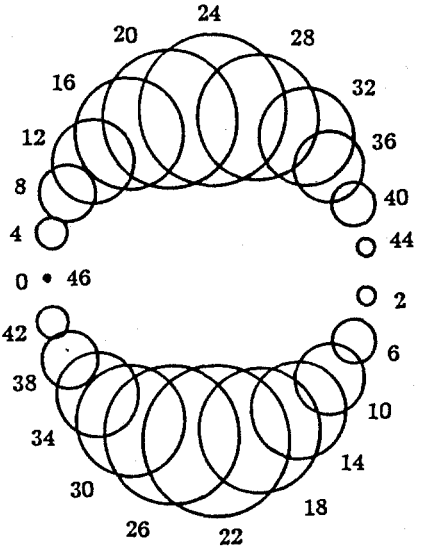

Fig. 6. Shape of the reflection spots in a delay line with minimal mirror diameter

On the other hand, there is the disadvantage of the large mirrors needed in delay lines. In a cavity the mirror diameter D has to be chosen to be $D = S \cdot 2w$ with a safety factor S of at least 3. A mirror diameter of the order of 20 cm is therefore sufficient for cavities with km-dimensions and light in the near infrared.

In delay lines the different reflections have to be more or less well separated from each other, the beam has to fit through the coupling hole and no light should fall from the edges of the mirrors. For the so-called matched case where all the reflection spots have the same size, the mirror diameter would become too big, especially for a large number of beams. The mirror diameter can be minimized if the diameter of the input beam is reduced,

allowing for a smaller coupling hole and closer packed reflections, see Fig. 6. This minimal mirror diameter D is given by

$$D = S\frac{2+\sqrt{2}}{\pi}\sqrt{\lambda L}\,. \tag{10}$$

Thus, D depends only on the wavelength of the light and the total light path. For green light, a path length of 100 km and a safety factor of 3 the mirror diameter would have to be about 70 cm.

A large mirror diameter is disadvantageous because of several reasons: the thermally driven eigenresonances of the mirror substrate come close to the frequency window of interest, and it is difficult to manufacture large mirrors with the required high quality (see below). For more details see for instance the contribution about delay lines and interferometric detection of gravitational waves [4].

2.3 Shot noise limit

The long light path is one condition for optimized performance of an interferometric gravitational wave detector. The other one is a good resolution of changes in path difference within the frequency window of observation. Provided the influence of all noise sources can be kept small enough, the limiting noise source is the fluctuation in photo current as it is determined by the statistics of the light at the output of the interferometer. Conventional laser light at its best is described by Poisson statistics, leading to the well known $\sqrt{n}$ law: if n photons are detected within a given resolution time, the uncertainty in that number is given by $\sqrt{n}$. On the other hand, a signal related to a changing path difference increases proportional to the number of photons available. Therefore, the signal-to-noise ratio improves with the square root of the number of photons, that is with the square root of the laser power. The spectral density of the mean-square fluctuation in path difference, simulated by the shot noise in an otherwise perfect interferometer, is given by

$$S_{\delta L}(f) = \frac{\hbar c}{\pi}\frac{\lambda}{\eta P}\,. \tag{11}$$

Here $\hbar$ is Planck's constant divided by 2π, c the speed of light, P the light power at the beam splitter and η the quantum efficiency of the photo diode. The spectral density is normalized to the bandwidth used; the units therefore are m^2/Hz. In order to get a linear measure, usually the square root of this quantity is given, with the unit $\mathrm{m}/\sqrt{\mathrm{Hz}}$.

Once a frequency window for observation is chosen, the relevant fluctuation in path difference can be obtained by integrating the spectral density of the noise over the frequency band in question.

The strain in space simulated by the shot noise in a perfect interferometer with an optical path length L, an effective light power P, an observational bandwidth Δf and green light is described by

$$\frac{\delta L}{L} = 1 \times 10^{-22} \left(\frac{50\,\mathrm{kW}}{\eta P} \frac{\Delta f}{1\,\mathrm{kHz}} \right)^{1/2} \frac{100\,\mathrm{km}}{L} . \tag{12}$$

This relation is valid for a path-length smaller than half a wavelength of the gravitational wave. A path-length of 100 km would be optimal for ms timescales.

Discouraging is the huge light power of 50 kW occuring in the last relation. It would drop to a more decent 50 W for the same strain sensitivity, if the frequency band of observation would be shifted from the region around 1 kHz to around 100 Hz, using an optical path of 1000 km and a bandwidth of 100 Hz. The optimal choice of the parameters certainly depends on the characteristic timescales of the gravitational waves. Since signals occurring at a reasonable rate of about one per month are expected to have an amplitude of 10^{-21} at most, and since detailed information about the waveform of a gravitational wave is of basic scientific concern, one will try to improve the sensitivity as far as possible. There have been several proposals in the past, how one could proceed.

All of the concepts are attractive in themselves, and have been proven to work in principle. But in order to work properly and reliably, they pose quite stringent demands on the quality of the optical components.

3 Techniques for improved sensitivities

Mainly three possibilities to improve the sensitivity of an interferometer are currently under investigation: power recycling, signal recycling and the use of squeezed states of light.

3.1 Power recycling

As we have seen in (11), a high light power sensing the relative position of the interferometer mirrors keeps the shot noise limit of the measurement process low. It seems therefore desirable to increase the light power beyond the level provided by the illuminating laser. One possibility is to add up coherently the output power of several lasers. In order to run all the lasers at the same frequency, a small fraction of the light of a well stabilized master laser is injected into the cavities of the other lasers; these lasers oscillate at the frequency of the injected light, if it is close enough to a possible eigenfrequency [5,6].

Another solution to the high power problem is to implement so-called power recycling. The idea behind it is the following: The interferometer is operated at a dark fringe at the signal output port. A fast servo loop maintains that condition, and a signal appears as a voltage or a current applied to the positioning elements. Ideally no light leaves through the signal

output port – it leaves the interferometer through the other output port towards the laser. As seen from the laser, the interferometer acts as a mirror. By inserting another mirror (M_3 in Fig. 7) between the laser and the interferometer, a Fabry-Perot cavity is formed.

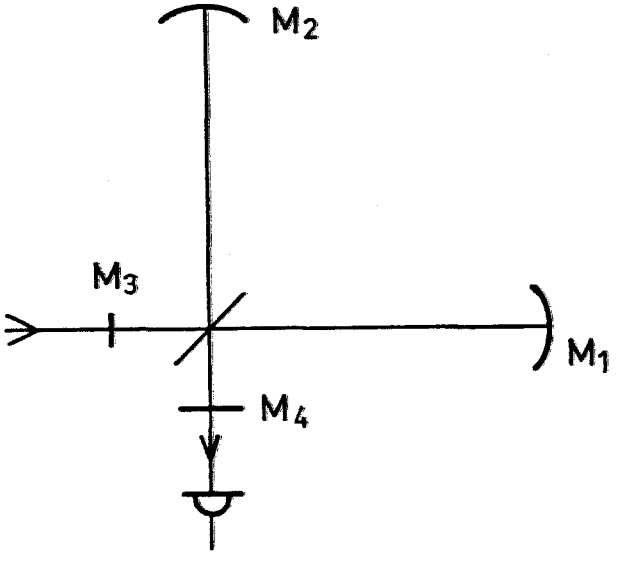

Fig. 7. Recycling the light in an interferometer: mirror M_3 for power recycling, mirror M_4 for signal recycling

The power-recycling cavity is tuned to the laser frequency, and the light power circulating inside the interferometer (and thus sensing the relative position of the mirrors) may be substantially increased if the overall power losses ΔP can be kept small. The power enhancement G is given by

$$G = \frac{P}{\Delta P}. \tag{13}$$

G is the inverse relative power loss per total round trip inside the interferometer. Losses occur because of scattering, absorption, residual transmission of the mirrors, and poor interference.

3.2 Signal recycling

A similar argument holds for the second recycling technique. As just mentioned, the interferometer is operated such that in the absence of a signal all light leaving the interferometer ideally goes back towards the laser. Analogously, light sent into the interferometer from the other side of the beam splitter leaves through the output port on that side. The interferometer therefore also acts like a mirror if one looks at it from this side of the beam splitter. By inserting a mirror (M_4 in Fig. 7) between interferometer and photo diode, a cavity for the signal is formed [7]. In general, this cavity will have to be tuned to a frequency different from the laser frequency: the effect of a gravitational wave is to introduce a time dependent strain in space; light being underway in a region underlying a time dependent strain in the direction of propagation can be considered as being frequency modulated – just as if the index of refraction would be modulated during the propagation through that medium. Superposition with the reference beam from the other arm can be considered as a superposition of light of the original frequency – the carrier – and of light of a frequency shifted by the frequency of the gravitational wave – the sidebands. For the carrier the interferometer is still in its condition of giving a dark fringe at the signal output port. But for the sidebands the interferometer is detuned and light is leaving towards the photodiode. Therefore the signal recycling cavity has to be tuned to the frequency of the sidebands.

The position of M_4 determines the resonance frequency, and the reflectivity defines the bandwidth (together with the losses inside the interferometer, of course). In general, for a very narrow resonance it is possible to resonate one sideband only. For broader resonance it may be possible to cover both sidebands by tuning to the carrier frequency.

This technique allows to optimize the interaction time of a gravitational wave with the light, even when the light path in the interferometer arms is relatively short. It is therefore, for instance, not necessary to realize a light path of half a gravitational wavelength in each arm for optimum performance; the longer light path may equally well be established by means of signal recycling. This fact helps to overcome the problem of too large a mirror size in delay lines.

If the losses and wavefront distortions can be kept small enough, a very high recycling gain for the power and for the signal may be possible. A multipath scheme like delay lines or Fabry-Perots in the arms may no longer be necessary; the arrangement of Fig. 7 with four mirrors in total could be sufficient. A practical difficulty arises from the high light power which in this case has to be transmitted through the beam-splitter. Thermal lensing is then very likely to limit the performance of the system. For this practical reason one will therefore choose at least a few bounces in each arm before the light goes back to the beam-splitter.

Signal recycling provides the possibility to track a signal of variable frequency, once it is detected and the subsequent shape is predictable within certain limits, like in the case of a coalescing binary. For these special sources this combines the convenience of broadband observation with the sensitivity of narrow band observation.

A very important consequence of signal recycling is the ability to regain at least part of the light that otherwise would leave the interferometer because of bad interference: this light is composed of other geometrical modes of the light-field than the fundamental one. The signal recycling cavity is tuned only to the fundamental mode; this field distribution builds up inside. The other modes are not resonant and are therefore suppressed by a factor up to the reflectivity of the signal recycling mirror. Very high order modes, originating from surface deformations with spatial wavelengths much smaller than the beam diameter, lead to diffraction losses, which certainly can not be regained.

3.3 Squeezed states of light

What limits the sensitivity of measuring phase differences in a Michelson interferometer – provided all other spurious signals are reduced sufficiently – is the noise due to the statistics of the light hitting the photodiode. Light of an ideal laser is described by Poisson statistics. This statistics leads to the theoretical sensitivity limit defined by (11). Surprisingly enough, it is not the fluctuations of the illuminating light that limit the performance of the system. Fluctuations of the laser light may take place either in amplitude or in phase. Fluctuations in amplitude only show up at the output of the interferometer if there is a deviation from the dark fringe at the measurement output port. A dark output stays dark, even when the illuminating light varies in amplitude.

A similar argument holds for the phase fluctuations of the laser light. In an interferometer phase fluctuations of the light give rise to a signal only if there is a path difference between the two interfering beams. For zero path difference, that is in the minimum of order zero, there is no light in the output, no matter which colour is used to illuminate the apparatus. A phase fluctuation is equivalent to a superposition of several frequencies, and thus on the zero order dark fringe no signal can be expected due to phase fluctuations in the illuminating light.

Where does the noise in the output come from when there seems to be no relation to the fluctuations of the input beam? The crucial point is the beam splitter. As Carlton Caves has pointed out, the noise due to photon statistics at the output of the interferometer can be described as originating from the zero point fluctuations of the vacuum field entering the interferometer through the normally unused input port symmetrically to the input illuminated by the laser [8] (see Fig. 8).

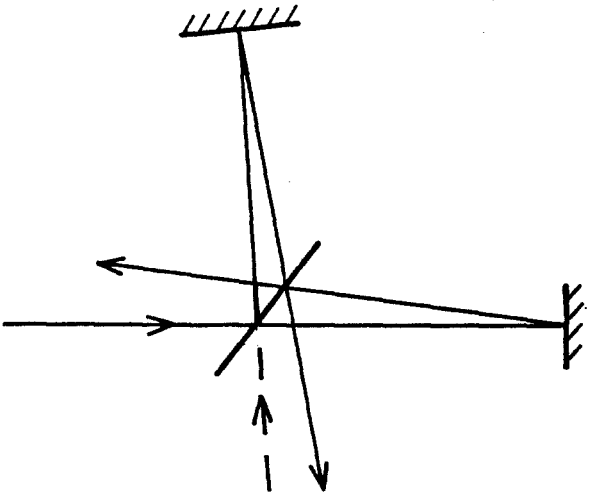

Fig. 8. Vacuum fluctuations entering the interferometer as indicated with dashed lines

These fluctuations are superposed with the light circulating inside the interferometer. The phase fluctuations are responsible for the uncertainty in the number of photons in the output, and the amplitude fluctuations give rise to the back action of the measuring process on the system via the time dependent differential light pressure on the mirrors. Caves proposed to send a particularly prepared state of the electromagnetic field into the second input port, replacing the usual zero point fluctuations, and thus reducing the fluctuations in the number of photons at the output. It is not necessary for this particular field to contain any real photons – it is sufficient to have less phase fluctuations than the vacuum field (or an ideal laser, which has the same uncertainty).

A reduction of the fluctuation in the number of photons in the output is connected with an increased fluctuation in differential light pressure onto the mirrors. But with realistic light powers and masses of several hundred kilograms this effect is negligible. Squeezed states of light have been realized in the meantime, and even the performance of an interferometer has been shown to improve by a few dB using squeezed states [9, 10]. Unfortunately the technique of squeezed states is not yet ready to be used in interferometric detection of gravitational waves; it will take some time to make it reliable and to overcome all the effects that tend to make the photon statistics Poissonian again.

The possible improvement of the sensitivity by use of squeezed states of light is limited by non-ideal conditions, particularly by losses, in a rather drastic way [11]. Equation (11) for a perfect interferometer can be rewritten as

$$\delta\phi_{\min} = 1/\sqrt{n}\,, \tag{14}$$

where n is the number of photons sensing the position of the mirrors within the chosen resolution time. For an ideal interferometer the gain in sensitivity

due to the application of squeezed light is proportional to the amount of squeezing, $1 - e^{-s}$. In this picture $s = 0$ means no squeezing. An imperfect visibility V limits the possible improvement to

$$\delta\phi_{\min} = \left(\frac{[2(1-V)]^{1/2} + e^{-2s}}{n} \right)^{1/2} . \tag{15}$$

(The visibility is related to minimum $P_{\min}$ and maximum $P_{\max}$ output power of the interferometer via $V = (P_{\max} - P_{\min})/(P_{\max} + P_{\min})$).

Thus, even for an arbitrarily high degree of squeezing the gain in sensitivity would be limited to $[2(1-V)]^{1/4}$. This means that an interference minimum of 10^{-4}, which seems to be an optimistic assumption as we will see soon, would allow a factor of 7 at most to be gained in sensitivity by utilizing perfectly squeezed light.

4 Quality of the optical components

There are several criteria that have to be fulfilled by the optical components. Some of them will be mentioned in the section 5.3 dealing with substrate materials. Here we will concentrate on the optical requirements. Most important is the request for low losses. Losses occur because of scattering, absorption and bad interference. We will now consider the losses introduced by a limited surface quality.

4.1 Surface quality

The ideal shape of the optical surfaces, especially of the mirrors, is in most of the cases either a sphere or a plane. Deviations from the ideal shape deform the wavefront of the passing beam. Depending on the spatial wavelength Λ of these deformations, there are three regions to be considered: Λ larger, equal to, or smaller than the beam diameter.

4.1.1 Aberrations

Deformations of the mirror surface on a scale larger than the beam diameter are called aberrations. They lead to a displacement and a tilt of the beams inside the interferometer. In a Fabry-Perot cavity the beam can be readjusted by properly orienting the mirrors. In a two mirror delay line not all distortions can be compensated by mirror alignment, for instance the effect of an astigmatism (that is different curvature of the mirrors in different directions). In this case the output beam in general is shifted with respect to its ideal position in radial as well as in tangential direction. Only the tangential displacement can be compensated by adjustment of the mirrors.

The visibility at the output of the 3 km interferometer can be kept better than 99%, if the error in slope of the mirror surface stays below 10^{-7} radian. If more than two mirrors are used in each interferometer arm, then position and orientation of the output beam are adjustable as well. One possibility is the use of so called retro-mirrors (see Fig. 9), where the beam leaves the delay line through a second coupling hole, hits a retro-mirror perpendicularly and retraces its original path.

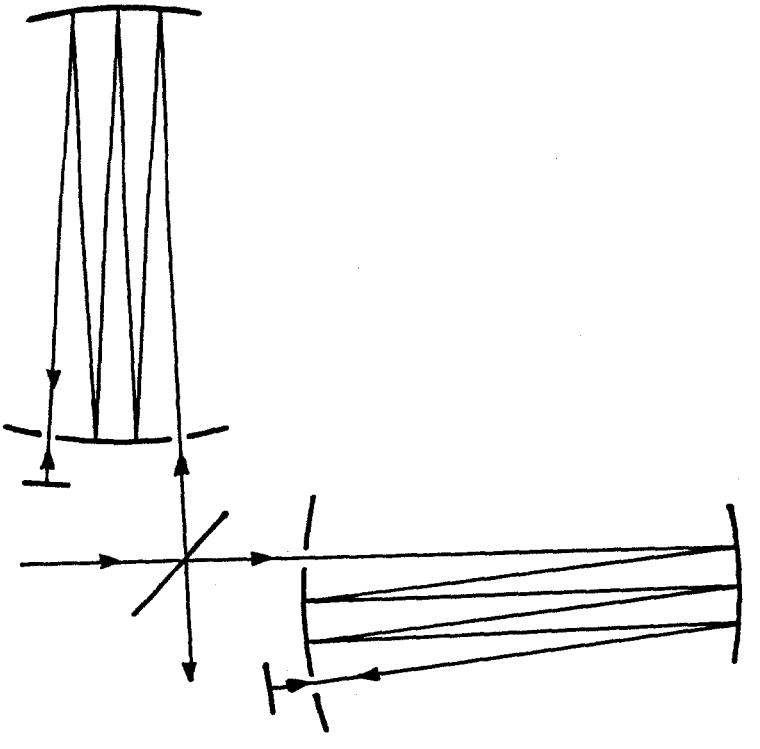

Fig. 9. Delay lines with retro-mirrors

There are several advantages of inserting a retro-mirror: for a given number of beams the number of reflection spots is smaller than in a two mirror delay line and therefore the mirrors can be kept smaller, and the recycling schemes can be realized by inserting one mirror only, without the necessity of complicated reorienting the output beam for good superposition with the input beam.

4.1.2 Ripple

Surface deformations with lateral dimensions in the order of centimeters are historically called ripple. In our case the beam diameter is just of that magnitude. The main effect of a ripple on a reflecting surface is then to deform the wavefront. Computer simulations have been performed for simplified conditions [12]: all surfaces and wavefronts have been assumed to be locally spherical, and deviations in curvature have been described by a deviation δs of the sagitta s at the mirror (see also Fig. 10):

$$s = \frac{w^2}{2R} \tag{16}$$

where w denotes the beam radius and R the radius of curvature of the wavefront. In a near confocal arrangement, where the mirror separation equals the radius of curvature, the sagitta is only a fraction of a wavelength:

$$s \approx \frac{\lambda}{2\pi}. \tag{17}$$

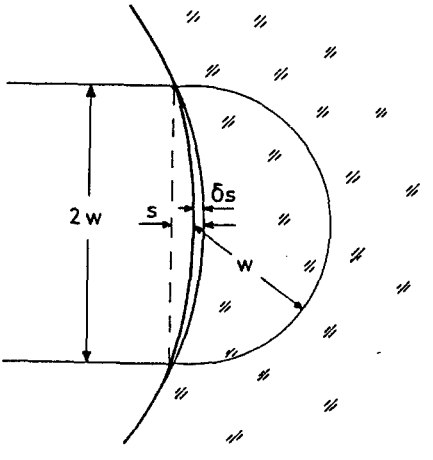

Fig. 10. A locally spherical deformation of the mirror surface and the corresponding change in sagitta

A reflection at a surface deformed by δs spoils the relative minimum at the output of an otherwise perfect interferometer to

$$\frac{P_{\min}}{P} \approx 10^{-3} \left(\frac{\delta s}{\lambda/100} \right)^2. \tag{18}$$

In a real setup there are deformations at each reflection. If we assume a statistical deformation of the relevant surfaces with a standard deviation s_d of the Gaussian - distributed surface amplitudes, then after N reflections the relative minimum at the output deteriorates to

$$\frac{P_{min}}{P} \approx 5 \times 10^{-2} \frac{N}{34} \left(\frac{s_d}{\lambda/100} \right)^2 . \tag{19}$$

This relation tells us that the wavefront deformation produced by reflections at a statistically deformed surface on the average grows proportional to the square root of the number of reflections N, and the minimum deteriorates with N. If we assume 34 reflections, then the deviations of the optical components from their ideal shape on a scale in the order of the beam size has to stay below $\lambda/200$ in order to allow a power enhancement by a factor of 100 via power recycling. This is a fairly hard condition, as the beam size in the final detectors will be somewhere between 5 and 10 cm (depending on the particular layout), and the amplitudes of the surface irregularities usually grow with the spatial wavelengths considered. Today's state of the art is very close to these goals. Deviations from an ideal shape for an aperture of 250 mm have been measured with a lateral resolution of 1/2 mm and an accuracy of a few Angstroms [13].

The demands on the surface quality at scales comparable to the beam size can be somewhat reduced if the technique of signal recycling is implemented (see the remarks in section 3.2 on signal recycling). Most of the light power leaving the interferometer through the output port is due to bad interference. It hits the signal recycling mirror and is sent back into the interferometer. But only the fundamental mode is resonant; higher modes are suppressed corresponding to the reflectivity of the signal recycling mirror.

4.1.3 Micro-roughness

Surface deformations with spatial wavelengths Λ smaller than a few mm – that is the order of magnitude of the beam size for small optical setups – are named micro-roughness in the usual terminology. In the large scale interferometers the beam size comes close to 10 cm, and therefore we have to include surface deformations up to spatial wavelengths of a few cm. Micro-roughness is one of the most important reasons for light scattering, and scattering is mostly the dominant loss mechanism in optical experiments. For a quantitative description of the scattering losses the quantity TIS (total integrated scattering) is used. It gives the fraction of light that leaves the main beam by propagating into arbitrary directions:

$$TIS = \frac{\delta P}{P} = \left(4\pi \frac{\delta s_{\mathrm{rms}}}{\lambda} \right)^2 \tag{20}$$

It is exactly the microroughness that defines the relevant surface deformation in (20). If we require the scattering losses to be small enough to allow a recycling gain of a factor of 100 in an interferometer with 34 reflections, then the last equation tells us that δs_{rms} has to be kept below $\lambda/730$, that is better than a nanometer for visible light. Within the last few years there has been an impressive progress in manufacturing high quality optical components. The rms-value δs_{rms} of the surface deformations for Λ smaller than a few mm can now be kept well below one tenth of a nanometer. It is again a hard (but manageable) condition, to extend the range of spatial wavelengths up to the required few cm.

4.2 Coating

In order to minimize reflection losses, the optical components are coated with dielectric layers. The art of producing high quality coatings has been brought to a very high standard within the last few years. Mirrors for visible light are available with overall reflection losses of less than 10 ppm (parts per million), and an absorption of hardly measurable 2 ppm. For infrared light, mirrors with even better performance have been made.

Some development work is still left to be done with respect to the size of the components involved. The very high quality mirrors made so far have a size of a few cm only, whereas for an interferometric gravitational wave antenna components with a size of several decimeters will be needed. Again, just as with the grinding procedure, it is more difficult to get extremely smooth surfaces up to spatial wavelengths in the decimeter region. As an example in Fig. 11 a scanline across a mirror used in the Garching prototype is shown, before and after coating, respectively.

No attempt has been made to produce a coating with constant thickness across the whole mirror, since only a circular area is used for the reflection spots. The coating therefore was done by sputtering from an off-axis ion source, while the mirror was rotated about its symmetry axis. As one can see from the pictures, the shape of the same surface (measured as deviation from an ideal sphere) looks totally different for the coated and for the uncoated case on a scale of several tens of nanometers. This is not surprising, since the overall thickness of the coating is about 5 micrometers, and a variation of 50 nanometers corresponds to a relative change of 1% only.

4.3 Density fluctuations

Wavefront deformations do not only occur when the beam is reflected at a non-ideal surface, but also when it is transmitted through a not perfectly homogeneous material, one that shows a gradient of the index of refraction across the light path. These variations result from density fluctuations, inhomogeneous distribution of the components composing the substrate, or from

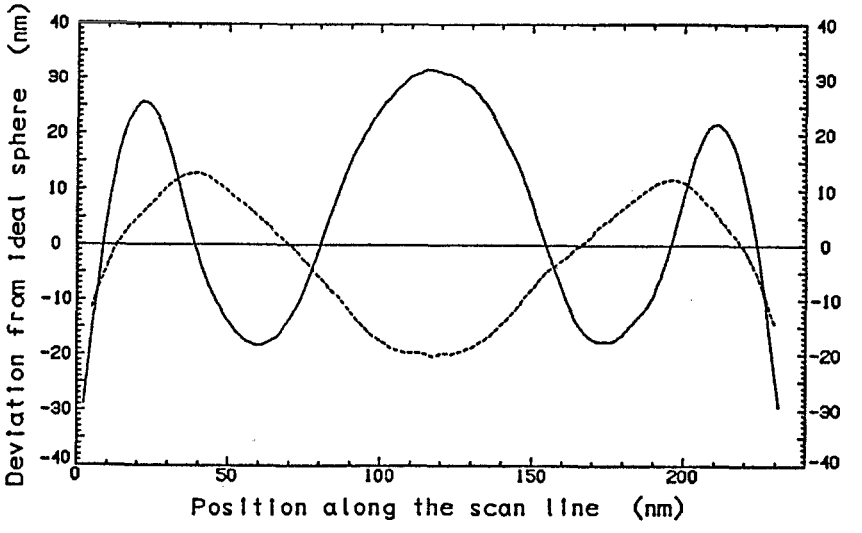

Fig. 11. Scanlines across a mirror; dashed line: before and solid line: after coating

temperature gradients (see the following section). In inhomogeneous materials different parts of the wavefront will see different optical path-lengths, leading to a deformation of the wavefront. Fluctuations of the index of refraction therefore have the same effect on the wavefront as a deformed mirror surface.

Just as in the case of surface deformations, the amplitudes of the fluctuations in the index of refraction grow with growing spatial wavelengths. The specifications for large scale interferometers are therefore harder to meet also in this respect. To give an example: The most homogeneous fused silica (a commonly used material for high quality optics) shows fluctuations of the index of refraction by 5×10^{-7} on a scale on the order of decimeters, that is just the beam diameter in the large interferometers. The path difference between the beam axis and outer parts of the beam, as inferred by these inhomogeneities, may then be up to 5×10^{-8} m for a 10 cm thick component, that is $\lambda/10$ for green light. Such a wavefront deformation would drastically limit the possible power enhancement by power recycling (see for instance (18)). The difference in light path that different parts of the wavefront see on their way through the interferometer could at least partly be compensated for by properly shaped surfaces of the components or compensation plates. (It is planned to include this topic in a research and development contract with Zeiss, Oberkochen). In addition, as already mentioned, the

implementation of signal recycling also reduces the losses by bad interference. This statement holds in particular for the lowest order modes, that are preferrably produced by the density fluctuations, and that are most effectively suppressed by signal recycling. But in any case the homogeneity of the substrates for the large antennas will have to be extremely high.

5 Thermal effects

Absorption at the components also contributes to the losses inside the interferometer. It is not the loss of light that worries us. To allow a power recycling gain of 100, the relative loss at each of the assumed 34 reflections would have to stay below 300 ppm only. This is no problem, since an absorption level in the ppm region is already possible. More severe are the thermal effects produced by the absorbed light power: thermal expansion and thermal lensing.

5.1 Thermal expansion

Let us consider a reflection at a mirror with an absorbing coating. For all relevant materials the heat will be removed by conduction rather than by radiation. The temperature profile and the related deformation by thermal expansion have been calculated [14], but the formalism and the results are somewhat complicated to handle. A good estimate of the effects in question can be obtained by the following consideration [12]: in a Gaussian beam most of the power is contained inside a circular area with radius w around the beam axis. Most of the heat produced at the reflection spot is therefore also limited to that area. The heat is removed by thermal conduction into the substrate. The steepest temperature gradient occurs in the hemisphere inside the substrate with its center at the beam axis, and its radius equal to the beam radius. Eventually the whole substrate is heated (by a small amount), and radiates the power away. In the equation for the heat conduction the relevant quantities are: the absorbed light power P_a, the heat conductivity κ, the temperature drop δT across the hemisphere, the area $A = 2\pi w^2$ through which the heat is transported, and finally α, the coefficient of thermal expansion.

$$P_{\mathrm{a}} = \kappa A \nabla T \approx \kappa 2\pi w^2 \frac{\delta T}{w}, \tag{21}$$

The temperature drop δT causes the hemisphere to expand by

$$\delta s \approx \alpha w \, \delta T/2 . \tag{22}$$

A combination of the last two equations gives

$$\delta s \approx \frac{\alpha}{4\pi\kappa} P_a \,. \tag{23}$$

This equation tells us that the thermal deformation δs does not depend on the beam size – it is the same for small and for large scale interferometers. On the other hand, the sagitta s is close to $\lambda/2\pi$ for all cases (see Equations (16) and (17)). Therefore, heating effects are of same magnitude for large and for small interferometers.

In order to minimize the influence of thermal deformation onto the performance of the interferometer, the absorption has to be kept very small. The high quality coatings available today show an absorption in the ppm range for visible and for infrared light. In addition one has to use substrate materials with a small ratio between expansion coefficient and thermal conductivity. For a description of the effect of such a deformation on the performance of the interferometer we can use the relations described under the heading *Ripple* of the last section. To give an example: 600 mW of absorbed light power at *one* reflection spot on a fused silica substrate degrades a perfect interference minimum to 1%, allowing a power recycling gain of less than 100 only.

5.2 Thermal lensing

The wavefront of a beam passing through a locally heated substrate is possibly deformed because of the temperature dependence $\beta = \partial n/\partial T$ of the index of refraction. A temperature gradient may occur because of absorption in the coating, or absorption along the light path in the substrate. The optical path along the beam axis differs from the optical path of outer parts of the beam by

$$\delta s \approx \frac{\beta}{4\pi\kappa} P_a \,. \tag{24}$$

P_a is the light power absorbed either in the coating or inside the substrate. Thermal lensing puts more stringent limitations on the tolerable level of absorbed light power than thermal expansion, since for all relevant materials β is larger than α. Again the example of fused silica: 20 mW of absorbed light power are sufficient to deteriorate a perfect minimum to 1% by thermal lensing. For this reason all substrate materials through which the light is transmitted have to be extremely pure and homogeneous in order to keep the absorption small.

The low absorption levels reached today can no longer be measured with calorimeteric techniques. More sensitive measurements have been proposed by Olmstead *et al.* [15] and Boccara *et al.* [16]. They are based on the effects of thermal expansion and thermal lensing. A strong laser with the wavelength in question is sent to the component to be investigated. The change in orientation of a weak and narrow probe beam, reflected at or transmitted through the locally heated substrate, and scanned across the

heated area, is monitored with a position sensitive photodiode. Absorbed light powers on the order of 10^{-7} watts are detectable. The Orsay and the Garching groups have used this technique to test their optical components [17, 18].

5.3 Substrate materials

The substrate materials have to meet several conditions. They have to be stable in shape, even on a scale of about one hundredth of a wavelength over a scale comparable to the beam diameter. Next, the mechanical quality factor has to be high, that is the mechanical internal damping has to be small in order to concentrate the thermally driven motions to a narrow bandwidth around the resonance frequencies. The resonances in turn have to be kept well outside of the frequency window of observation. In this way the tails of the resonances extending to the frequencies of observation can be kept small enough. This topic is covered in more detail in the contribution of A. Rüdiger in this issue.

The optical components should not be magnetic, even magnetic impurities have to be avoided, in order to exclude motions introduced by fluctuating electromagnetic fields.

As stated in the last sections, for minimum thermal deformation the ratio between the coefficients of thermal expansion and thermal conductivity has to be small. If the beam is transmitted through the substrate, then for minimum thermal lensing the absorption inside the material and the ratio between temperature dependence of the index of refraction and thermal conductivity has to be small.

There are several materials which meet the requirements.

Table 1.

material	$\alpha/\kappa\,(10^{-8}\mathrm{m/W})$	$\beta/\kappa\,(10^{-8}\mathrm{m/W})$
fused silica	33	1000
ULE	± 2.3	850
silicon	1.28	-
sapphire	28	60
diamond	0.13	1

First of all, there are the materials normally used for optical applications, like fused silica, ULE and Zerodur. Unfortunately Zerodur, a special glass ceramic made for very low thermal expansion, is ruled out because of its high mechanical damping. Much better in this respect are ULE (also a material with low thermal expansion) and fused silica. They are possible candidates for mirror substrates. Despite their fairly strong thermal lensing, the very

low absorbing versions may be used for the beam-splitter or even for the coupling mirror of a cavity, as long as the light power is not too high.

Components at which the light is only reflected and not transmitted, like delay-line mirrors or end mirrors of cavities, are allowed to be made from opaque materials. Silicon would be a very good choice: it has a high mechanical quality factor, is very unsusceptible to thermal deformation and can be made in large pieces.

Sapphire also has a high quality factor and shows less thermal lensing than fused silica. But so far it cannot be made in large and very homogeneous pieces. There are also problems related to its birefringence.

Beryllium oxide would also be well suited, but it is not listed here because of the toxic dust produced during the grinding procedure.

The ideal substrate would be diamond because of its extraordinarily high thermal conductivity. At present it is certainly unrealistic to count on that material. But slices of polycrystalline diamond with decimeter diameter and fractions of a millimeter thick can already be made. So it is not totally impossible that at some time sufficiently large pieces of artificial diamond will be available.

6 Performance limitations due to thermal effects

During the last few years an impressive progress in manufacturing high quality optics took place. It is therefore sound to assume that the optical components of an interferometer can be produced with the required specifications to get an appropriate interference quality. But, after all, the wavefronts may eventually be deformed by thermal effects. In order to quantify their influence on the performance of an interferometer, computer simulations based on the simplified models described above have been carried out [12]. An interferometer with power recycling has been assumed. The implementation of signal recycling would improve (but also complicate) the situation somewhat. The order of magnitude of the effects becomes already clear in the plots of Figs. 12 to 14. Here the input power to be delivered by the laser is plotted against the power circulating in an interferometer with power recycling. Figure 12 relates to an interferometer with a Fabry-Perot cavity in each arm, and Fig. 13 describes a system with the same parameters, but with delay lines in the arms. In both cases Rayleigh scattering of 200 ppm was assumed (as occurring in 10 cm of fused silica and green light), 20 ppm absorption inside components traversed by the light, 34 reflections in each arm (or an equivalent finesse in the case of Fabry-Perots), 30 ppm total loss and 5 ppm absorption in the coating. The losses were assumed to vary between the different components or between the different reflections on the average by 10%.

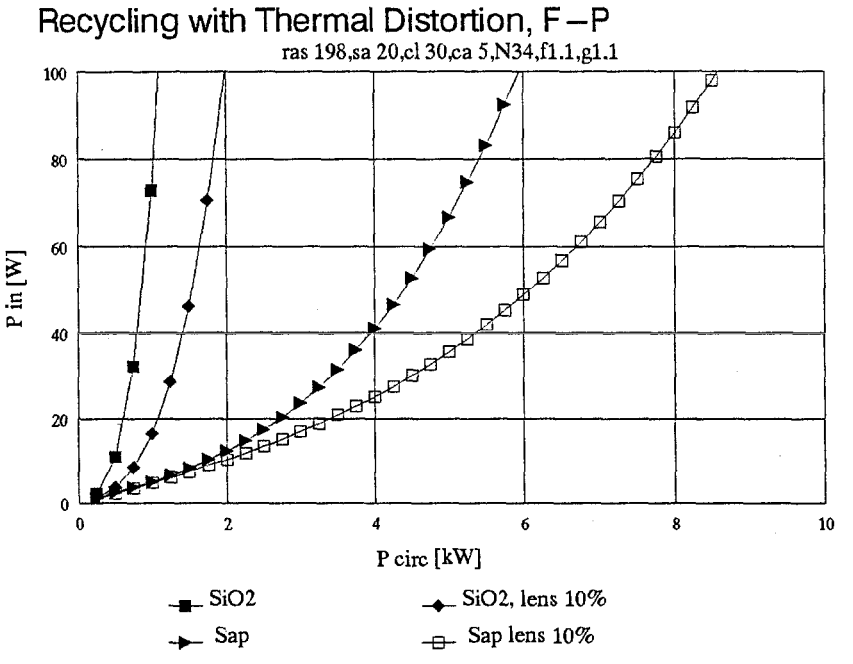

Fig. 12. Required input power P_{in} as a function of the circulating power P_{circ} in a thermally distorted interferometer containing Fabry-Perots in the arms.
filled squares: all components made from fused silica
filled diamonds: the same, but the thermal lens in the beam-splitter is compensated by a compensation plate down to 10%
filled triangles: all components made from sapphire
empty squares: the same, but compensation of the thermal lens in the beam-splitter

The most critical point for Fabry-Perot systems (Fig. 12) is the thermal lens in the coupling mirror, produced by absorption in the coating which is exposed to the enhanced power inside the cavity. This explains the striking improvement from the curve with filled diamonds to that with filled triangles, where the thermal lens in the substrate is reduced by using sapphire. The thermal lens in the beam-splitter is of minor importance, as its reduction down to 10% by use of a compensation plate only leads from the curve with filled squares to that with filled diamonds.

Delay line systems (Fig. 13) are less susceptible to thermal deformations than those with Fabry-Perots. Here the dominating effect is the thermal lens in the beam-splitter; its successively better compensation gives a correspondingly better performance, as shown in the figure. Starting with a fused silica beam-splitter (filled squares), a compensation of the thermal lens down to 10% (filled diamonds), a sapphire beam-splitter (filled triangles) and finally a compensated sapphire beam-splitter (empty squares) have

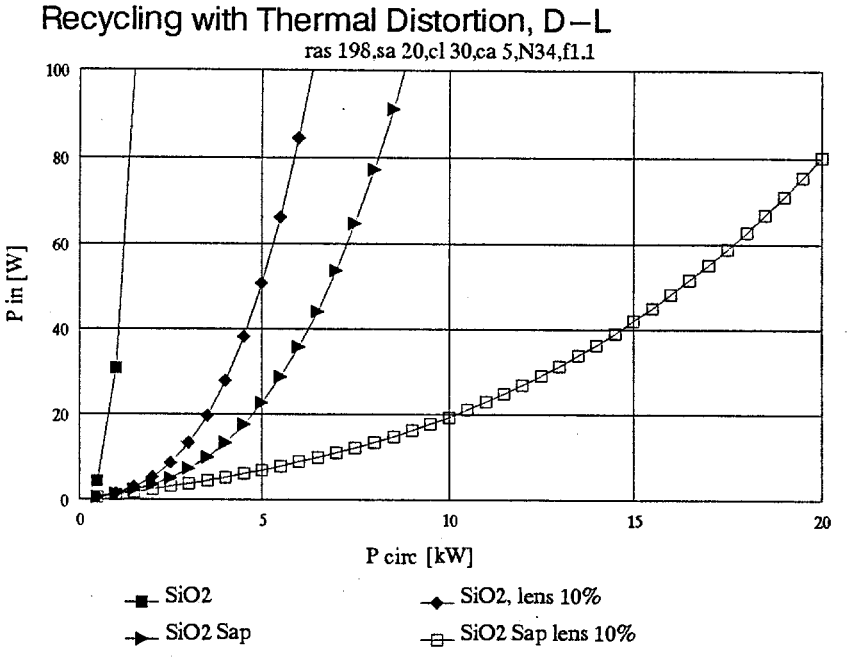

Fig. 13. Required input power P_{in} as a function of the circulating power P_{circ} in a thermally distorted interferometer containing delay-lines.
filled squares: all components made from fused silica
filled diamonds: the same, but the thermal lens in the beam-splitter is compensated by a compensation plate down to 10%
filled triangles: fused silica mirrors and sapphire beam-splitter
empty squares: the same, but compensation of the thermal lens in the beam-splitter

been assumed. The thermal expansion of the mirrors is of minor importance, as for all four lines mirrors made from fused silica have been assumed.

Finally in Fig. 14 some optimistic assumptions have been made as far as the materials are concerned. But they show that even light powers of 100 kW at the beam-splitter are conceivable. The common parameters for all four lines are: Rayleigh scattering 200 ppm, absorption in the substrate 10 ppm, coating loss 10 ppm, coating absorption 1 ppm, 34 beams, a variation of the losses of 10%, and a compensation of the thermal lens in the beam-splitter down to 10%.

It should be mentioned that Figs. 12 to 14 are based on green light. For infrared light the situation is still more promising, since the losses (scattering and absorption) usually are lower there.

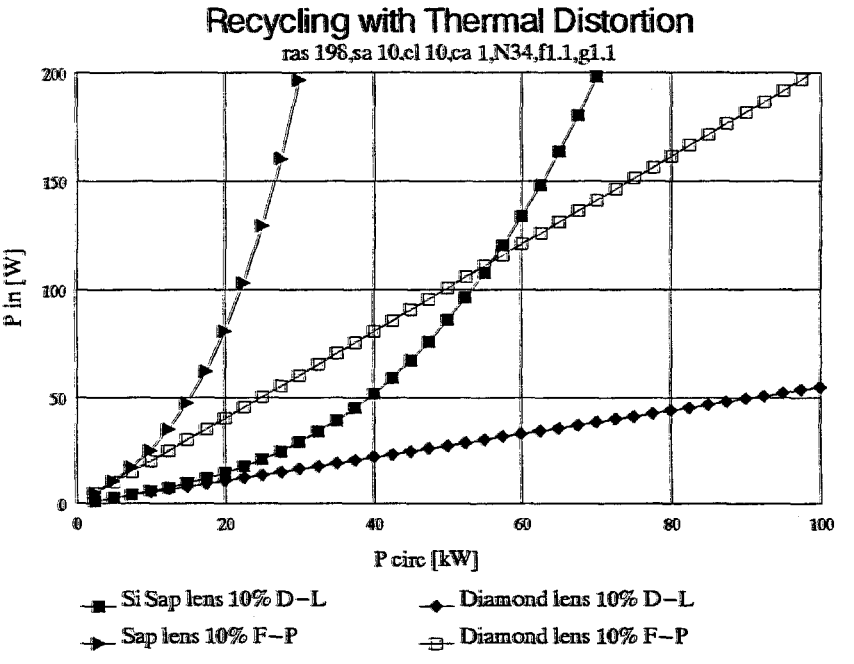

Fig. 14. Required input power P_{in} as a function of the circulating power P_{circ} in a thermally distorted interferometer.
filled triangles: a Fabry-Perot system made of sapphire
filled squares: a delay-line system with silicon mirrors and sapphire beam-splitter
empty squares: a Fabry-Perot system made of diamond
filled diamonds: a delay-line system made of diamond

7 Stabilization of the light

For completeness one should also mention the optics that is needed to stabilize the laser light in frequency and in its geometrical properties, before it is sent into the interferometer. The necessity for these kinds of stabilization arises because fluctuations in the beam parameters produce spurious signals in connection with small, but practically unavoidable asymmetries inside the interferometer.

7.1 Frequency stabilization

One example of such a mechanism is the signals produced by frequency fluctuations in connection with a path difference ΔL between the interfering beams. The fluctuations $\delta\phi$ in phase difference as simulated by the frequency noise $\delta\nu$ of the light are given by

$$\delta\phi = 2\pi\frac{\Delta L}{c}\delta\nu\,. \tag{25}$$

The path difference ΔL depends on the level of symmetry between the two arms. As described above in Sect. 2.2, the path length L in a delay line is rigidly related to the radius of curvature of the mirrors. With existing technology one can hope to get the radius of curvature of the 3 km mirrors equal to about 10^{-3}. So with 30 beams one has therefore to expect a path difference of about 100 m. This path difference can in principle be made to disappear if retro-mirrors are used. But there is another effect which poses demands on the frequency stability in the same order of magnitude as the path difference between the main beams [4]: scattered light with a huge path difference with respect to the main beam and eventually interfering with it also leads to spurious signals.

A small part of the laser light is therefore sent to a stable Fabry-Perot cavity. The laser frequency is servoed to maintain resonance of the light in that cavity. *Stable* means in this context that in the frequency window of observation of the interferometer the cavity mirrors are allowed to show very small motions only. Even thermally driven motions at eigenfrequencies are too large. The frequency servo would then make the frequency follow these motions. The cavity is therefore either made rigid enough to have all eigenfrequencies above the frequency window in question, or uses suspended mirrors just like the interferometer. In a second stage the frequency is stabilized with respect to the light path of the interferometer itself – the quietest reference we have. In practice this can be done by stabilizing the laser frequency to the length of the power recycling cavity.

7.2 Mode cleaner

Another mechanism for spurious signals to appear because of unstabilities of a real laser beam are fluctuations in the beam geometry combined with proper asymmetries between the interferometer arms. A simple example is a lateral displacement δy of the beam combined with an angular misalignment α of the beam-splitter. The path-difference simulated in this case is

$$\delta L = 4\alpha\delta y. \tag{26}$$

To give a numerical example: at the relevant frequencies the beam of an Argon ion laser shows fluctuations in beam position of about $10^{-10}\mathrm{m}/\sqrt{\mathrm{Hz}}$. This would require the angle of misalignment to stay below 10^{-9} radian – certainly not easy to fulfil, even for a servo system to maintain the alignment.

Similar relations hold for other kinds of beam fluctuations, for instance a pulsation in beam width combined with differently curved wavefronts of the interfering beams. Spurious signals of that sort can be suppressed by inserting a so called mode cleaner between the laser and the interferometer [19]. This mode cleaner is essentially a non-confocal Fabry-Perot cavity, where the mirror separation is chosen to resonate only one eigenmode of the

electromagnetic field. Any fluctuation in beam geometry can be described as admixtures of other modes, which are not resonant in the cavity – they are not transmitted, they are reflected. The transmitted beam in consequence fluctuates in power – but power fluctuations are made ineffective in the signal output anyhow by use of a nulling method.

Besides mode cleaners, single mode fibers have also been used in the prototype experiments to suppress geometrical beam fluctuations. In addition, it is easier to handle position and orientation of the beam when fibers are used. But so far no fibers have been produced to stand more than one Watt of transmitted single-mode light power. For higher laser power it will therefore be necessary to use mode cleaners again.

8 Conclusion

The specifications for the optical components in a large baseline interferometric gravitational wave detector are at the limits of the well advanced technology today: surfaces smooth to better than 10^{-10} m on a scale of up to a few centimeters; on a larger scale the deviations of the surface shape with respect to an ideal sphere have to stay in the nanometer region. The losses, particularly the absorption, are tolerable only at a level of a few ppm. The beam-splitter, and particularly the coupling mirror of a Fabry-Perot, have to be made of extremely homogeneous and pure materials. Tests and first results have shown that the problems are manageable.

List of Authors

W. Winkler, J. Chen, K. Danzmann, P.G. Nelson, T.M. Niebauer, A. Rüdiger, R. Schilling, K.A. Strain, L. Schnupp, H. Walther, *Max-Planck-Institut für Quantenoptik, D-8046 Garching, Germany*; J. Hough, A.M. Campbell, C.A. Cantley, J.E. Logan, B.J. Meers, E. Morrison, G.P. Newton, D.I. Robertson, N.A. Robertson, S. Rowan, K.D. Skeldon, P.J. Veitch, H. Ward, *Department of Physics, University of Glasgow, Glasgow, UK*; H. Welling, P. Aufmuth, I. Kröpke, D. Ristau, *Laser-Zentrum and Institut für Quantenoptik, Universität Hannover, D-3000 Hannover, Germany*; J.E. Hall, J.R.J. Bennett, I.F. Corbett, B.W.H. Edwards, R.J. Elsey, R.J.S. Greenhalgh, *Rutherford Appleton Laboratory, Chilton, Didcot, UK*; B.F. Schutz, D. Nicholson, J.R. Shuttleworth, *Department of Physics, University of Wales, Cardiff, UK*; J. Ehlers, P. Kafka, G. Schäfer, *Max-Planck-Institut für Astrophysik, D-8046 Garching, Germany*; H. Braun, *Bauabteilung der Max-Planck-Gesellschaft, D-8000 München, Germany*; V. Kose, *Physikalisch-Technische Bundesanstalt, D-3300 Braunschweig and D-1000 Berlin, Germany.*

References

1. A. E. Siegmann, *Lasers* (University Science Books, Mill Valley, 1986).
2. *The Detection of Gravitational Radiation*, edited by D. Blair (Cambridge University Press, 1991).
3. D. Herriot, H. Kogelnik, and R. Kompfner, Appl. Opt. **3**, 523 (1964).
4. W. Winkler, in *The Detection of Gravitational Radiation*, edited by D. Blair (Cambridge University Press, 1991).
5. C.N. Man, A. Brillet, Lecture Notes in Physics **212**, 222 (1984).
6. C.N. Man, A. Brillet, Opt. Lett. **9**, 333 (1984).
7. B. J. Meers, Phys. Rev. D **38**, 2317 (1988).
8. C. M. Caves, Phys. Rev. D **23**, 1693 (1981).
9. M. Xiao, L. A. Wu, and H. J. Kimble, Phys. Rev. Lett. **59**, 278 (1987).
10. P. Grangier, R. E. Slusher, B. Yurke, and A. La Porta, Phys. Rev. Lett. **59**, 2153 (1987).
11. J. Gea-Banacloche and G. Leuchs, J. Mod. Opt. **36**, 1277 (1989).
12. W. Winkler, K. Danzmann, A. Rüdiger, and R. Schilling, Phys. Rev. A, **44**, 7022 (1991).
13. K. Freischlad, M. Küchel, W. Wiedmann, W. Kaiser and M. Mayer, Proc. SPIE **1332**, 8, (1990).
14. P. Hello and J. Y. Vinet, J. Phys. (France) **51**, 2243 (1990).
15. M. A. Olmstead, N. M. Amer, and S. Kohn, Appl. Phys. A **32**, 141 (1983).
16. A. C. Boccara, D. Fournier, W. Jackson, and N. M. Amer, Opt. Lett. **5**, 377 (1980).
17. N. Man, private communication
18. M. Engl, Diplomarbeit, (Garching, 1991), unpublished.
19. A. Rüdiger, R. Schilling, L. Schnupp, W. Winkler, H. Billing, and K. Maischberger, Optica Acta **28**, 641 (1981).

Mechanical Aspects in Interferometric Gravity Wave Detectors

A. Rüdiger and colleagues of GEO [1]

Max-Planck-Institut für Quantenoptik, D-8046 Garching, Germany

[1] *Names and affiliations of all authors are given at the end of the paper*

Abstract: In order to measure the tiny effects of gravitational waves, strains in space (i.e. relative changes in distance) of as little as 10^{-21} or even less have to be detected, at frequencies ranging from 100 Hz to several kHz.

Large laser interferometers are the most promising approach to reach such extreme sensitivities. This 'straightforward' road is, however, obstructed by a multitude of effects that cause (or fake) such fluctuations in distance. Among these are seismic motions, thermal vibrations of optical components, pressure fluctuations of the residual gas in the vacuum tubes, and fundamental effects such as Heisenberg's uncertainty relation.

What all of these noise sources have in common is that their effects can be reduced by the choice of sufficiently large arm lengths. This is what dictates the (very expensive) choice of arm lengths of 3 to 4 km in the currently proposed gravitational wave detectors (USA, D–GB, F–I, AUS, JAP).

Introduction

0.1 Objective of talk

This is the third in our series of four talks on laser interferometric gravitational wave detection. The other talks are presented by G. Schäfer [1], K. Danzmann [2], and W. Winkler [3].

This talk will address the limitations in sensitivity imposed by mechanical noise of various kinds. It will discuss the origin of these noise sources and the methods to reduce them or suppress their effects. In particular, it will make clear why the long arm lengths envisaged are an absolute necessity.

0.2 Layout of talk

The talk will be divided into three parts. A first part will set the scene, briefly recapitulating astrophysical background, proposals, basic parameters, appropriate representation of noise, and it will introduce our noise "yardstick", the shot noise.

The second part will treat what can be termed the "internal noise " sources, of rather fundamental nature: Heisenberg's uncertainty relation, thermal vibrations, and index fluctuations of the residual gas in thermal equilibrium.

And only then will the third part discuss the mechanical noise that one would normally think of first: seismic vibrations, as an example of "external" noise sources, and the ways to cope with them.

This sequence was chosen to be able to cover the very important topic of seismic isolation more broadly, and to be able to include a series of investigations that had just recently been performed.

1 Setting the scene

1.1 Astrophysical Background

The talk of G. Schäfer treated the most likely sources of gravitational radiation, and again recalled how extremely small their strain amplitudes, h, are. Let us just repeat in a few words the conclusions about the two main candidates for detection: supernovae and coalescing binaries. One should, however, keep in mind that each new observational window on the universe has brought totally unexpected discoveries, and we expect that the same will come true also in gravitational wave astronomy.

1.1.1 Supernovae

Burst sources such as supernovae, are rare events, and to wait for one of them to occur in our own galaxy, the Milky Way, may take a normal physicist's working life. For a higher rate of events, perhaps a few per month, we would have to look as far out as the Virgo cluster, to a distance of more than 10 Mpc.

Carefully performed ab initio calculations of the collapse processes [4] reveal that a rotationally symmetric collapse (at a distance of 10 Mpc) cannot be expected to produce peak strain amplitudes of more than 10^{-22}. Calculations of similar reliability have not yet been performed for non-axisymmetric scenarios, but it is hoped that such cases might have up to an order of magnitude higher signals.

We conclude, therefore, that for burst sources a design sensitivity of 10^{-21} is the absolute minimum requirement, and clearly an upgrade to 10^{-22}

must remain a long term goal. The typical frequency range for burst sources is the region from a few hundred hertz up to a few kHz.

1.1.2 Coalescing binaries

The other most likely source of gravitational radiation is the spiralling together of close binaries composed of highly condensed partners: neutron stars or – even more efficient – black holes.

The rate at which such events are to be expected is somewhat controversial; the statistics of such binaries is still very sparse. The general opinion is, however, that the rate is so low that a coverage of the universe even deeper into space than the Virgo cluster is required, perhaps as far as 100 Mpc.

The signal of such coalescing binaries would consist of a quasi-continuous wave of slowly rising frequency (a 'chirp'), and also of slowly rising amplitude. The signals are difficult to detect until they come into the range of, say, 100 Hz. Within a few seconds, the evolving coalescence will reach frequencies of 200 Hz, and then only fractions of a second until final splash-down.

A sensitivity of 10^{-22} (already taking into account the longer observation time) is the design goal for such coalescing binaries, and being able to measure down to, say, 100 Hz becomes even more important here.

1.2 The proposals

The presentation of K. Danzmann gave an overview of the proposals made worldwide for building such laser-interferometric gravitational wave detectors of sufficient sensitivity. Even though some of the design details may differ between the three most advanced designs (GEO [5], LIGO [6], VIRGO [7]), there are some features that are very similar.

The most notable (and noticeable) common feature is the proposed length of the interferometer arms: 3 km each in VIRGO and GEO, 4 km in the wealthier and less populated United States (LIGO), and again 3 km in the Australian design [8] and the more recent Japanese concept [9].

This important design parameter, the arm length ℓ, turns out to be the major cost factor; the cost of civil engineering and of the vacuum system is approximately proportional to the length ℓ, and they make up close to 70 % of the total cost. Thus, a reduction in arm length would cut down the detector cost considerably.

There have been suggestions from various researchers on how one could build interferometric gravitational wave detectors having much smaller dimensions, perhaps even of 'table top' size. One main objective of my talk is to state the physical facts that rule out such possibilities. The choice of arm lengths in the order of 3 km is not a reckless use of taxpayers' money, nor an attempt to build impressive monuments for posterity, but rather it is governed by physical necessities, if the dream of a gravitational wave astronomy is to become true.

The suggested solutions of smaller sized detectors usually concentrated on how one can (or *hopes to*) defeat one particular noise effect, but then typically disregarded the other noise sources that also dictate the choice of long (km-sized) arms.

1.3 Basic parameters

In this subsection we will define some of the variables that will be used repeatedly in the sections to come, and give their typical range of values.

Of particular concern will be the arm length ℓ, i.e. the geometric distance between the mirrors in each of the interferometer arms. When we discuss *optical delay lines* as proposed in the GEO project, the total optical path L is given by multiplying ℓ with the number of passes, N:

$$L = N\,\ell\,. \tag{1}$$

In GEO, in order to obtain light travel times $\tau = L\,/\,c$ that are appropriate for kHz signals ($\tau = 0.3\,\mathrm{sec}$), the total path L needs to be in the order of 100 km, so with $\ell = 3\,\mathrm{km}$ we would need $N \approx 30$ passes in the delay line.

In LIGO and VIRGO, long light storage times are realized with *Fabry-Perot cavities*. The sensitivity with which gravitational wave signals can be detected is determined by the phase sensitivity, $\mathrm{d}\Phi/\mathrm{d}\ell$, and we see the finesse $\mathcal{F}$ play a similar rôle as the number of passes, N, in the delay line.

More recent ideas, such as the concept of "signal recycling" [10], make the distinction between the delay line scheme and the Fabry-Perot scheme less pronounced.

The various noise sources will be described with the specific GEO configuration in mind, but most of these are easily extrapolated to the configurations of LIGO and VIRGO. As will be seen, many of the effects can be discussed without having to make very specific assumptions about the particular design.

1.4 Noise Representation

The noise types to be treated here are all broadband, and of stochastic nature. A stochastic noise variable $v(t)$, of, say, dimension in meters, is then best represented by the spectral density (of the square) of the fluctuating variable, and it has become customary to give the *linear* spectral density denoted by

$$\tilde{v}(f), \qquad \text{of dimension } [\mathrm{m}\,/\,\sqrt{\mathrm{Hz}}] \tag{2}$$

such that the *rms* value in a given frequency band $\Delta f = f_u - f_l$ is given by

$$v_{\mathrm{rms}} = \left(\int_{f_l}^{f_u} \tilde{v}^2(f)\,\mathrm{d}f \right)^{\frac{1}{2}}, \qquad \text{again of dimension } [\mathrm{m}]\,. \tag{3}$$

It is important to keep in mind that the goal is to detect gravitational radiation in a frequency range that does not necessarily extend to very low frequencies.

In most cases in the following, we will assume a relatively large band width Δf of the interferometer, *i.e.* on the order of the median frequency f for which the interferometer is optimized. When a choice of Δf has been made, the sensitivity obtainable can be expressed as a function of the ***design*** frequency f. It is this type of representation that is chosen for plotting the noise contributions in Figure 1, where a bandwidth of $\Delta f = f/2$ is assumed throughout. The figure, taken from [5], also shows a rough indication of the magnitude of the expected signals.

1.5 Shot noise – the limit?

As is well known, the existence of shot noise gives rise to a very fundamental limit in sensitivity. The traditional representation is that shot noise fakes a fluctuation in phase of the measured output signal, or, in other words, an apparent fluctuation in total path length difference, $\widetilde{\delta L}$, as expressed in Equation (11) of W. Winkler's contribution [3]. When, as we will do here throughout, we express the noise as the attainable strain h, we find for the shot noise

$$h_{\mathrm{SN}} \approx 2.4 \times 10^{-21} \left[\frac{\epsilon I_0}{50\mathrm{W}}\right]^{-1/2} \left[\frac{L}{100\,\mathrm{km}}\right]^{-1} \left[\frac{f}{1\mathrm{kHz}}\right]^{3/2}, \qquad (4)$$

where ϵ is the quantum efficiency of the detector, I_0 is the laser output power, and f is the center frequency of the burst. It is noteworthy that this limit does not depend on the choice of the arm length ℓ, but rather on the total path length $L = N\ell$, regardless of how this is realized by the two factors N and ℓ.

In W. Winkler's contribution, the consequences of shot noise have been made quite clear. Even for obtaining our more modest goal of $h < 10^{-21}$ at 1 kHz, a light power of close to 1 kW would be required (of highly stabilized, single mode light); and for the eventual goal of $h < 10^{-22}$ a truly prohibitive value of almost 100 kW. No light source (laser) of such high power, which also satisfies all the other requirements, is anywhere in sight.

Fortunately, the light power inside the interferometer can be enhanced considerably by what is known as *power recycling*. The interferometer output is measured in the dark fringe of the interference. If for the moment we neglect the non-zero interferometer minimum, all light that is not lost due to the finite reflectivity loss, $(1-R)$, will be available for recycling (re-injecting) into the interferometer. The longer the arm length ℓ, the fewer passes N in the arms are required. This reduces the light loss due to the mirrors, and thus allows better power recycling. In this way, the strain sensitivity might reach best values of as little as

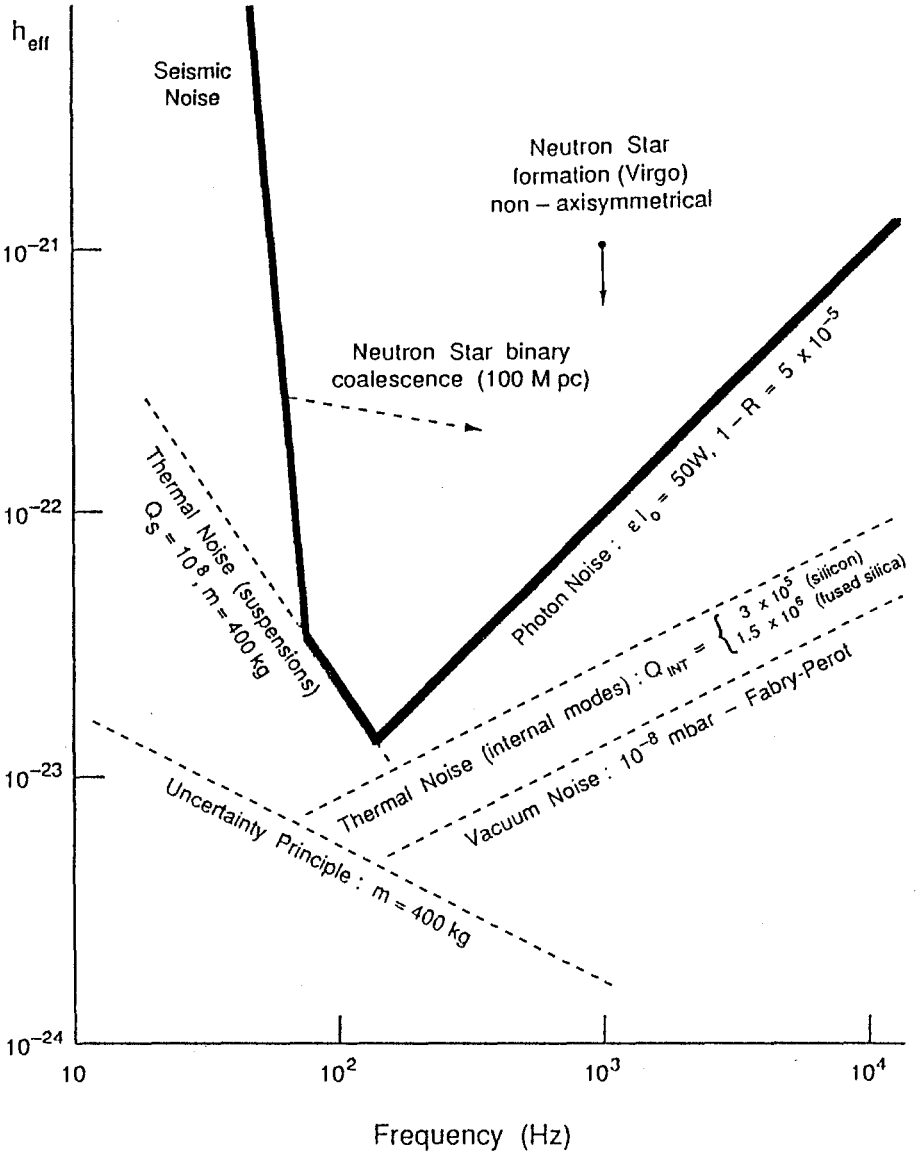

Fig. 1. This figure, taken from the German-British proposal [5], compares the strengths of two typical burst sources with the noise limitations imposed by the most prominent noise sources. For signals that allow observation over several oscillation periods, the effective amplitude h_{eff} is approximately $h\sqrt{n/2}$, where h is the true amplitude and n is the number of cycles of the waveform over which the signal can be integrated.

$$h_{\rm SN} \approx 10^{-22} \left[\frac{f}{1\,{\rm kHz}}\right] \left[\frac{\epsilon I_o}{50\,{\rm W}}\right]^{-1/2} \left[\frac{1-R}{5\times 10^{-5}}\right]^{1/2} \left[\frac{\ell}{3\,{\rm km}}\right]^{-1/2}, \quad (5)$$

and we see that the sensitivity $h_{\rm SN}$ improves (*i.e.*: drops) with the square root of the arm length ℓ.

So, even before we get to the actual mechanical noise sources, we find that our "measuring stick", the shot noise limit, is dependent on the arm length ℓ. In Figure 1, the heavy line denoted 'Photon Noise' assumes the design length of $\ell = 3\,{\rm km}$. The sensitivity would deteriorate (the line would move upward) if arm lengths shorter than 3 km were chosen, jeopardizing the detection of such events as those indicated in Figure 1: supernovae and coalescing binaries.

1.5.1 Squeezed light

A way around the shot noise limitation could be found if non-classical states of light were to be used in the interferometer. Caves [11] pointed out that the photon counting noise in an ideal interferometer can be interpreted as stemming from the ground state vacuum fluctuations entering through the unused input port of the interferometer.

If one succeeded in replacing them by a specially prepared light of particularly small phase fluctuations, the photon counting noise could be reduced considerably. Such *squeezed states* of light have successfully been generated, but their usage in gravitational wave interferometers is still not in sight. Furthermore, the gain in signal-to-noise ratio will be very limited, as is pointed out in the contribution of W. Winkler [3].

2 Internal Mechanical Noise

2.1 The Heisenberg Uncertainty Principle

The indeterminacy in the simultaneous measurement of the position x and the associated momentum p_x, as expressed in Heisenberg's uncertainty relation

$$\Delta x \Delta p_x \geq \hbar/2, \quad (6)$$

gives a lower limit down to which a measurement of the current mirror displacement $\delta\ell$ is possible. One easily derives a (squared) spectral density

$$\tilde{h}^2 \approx \frac{8\hbar}{m\,\omega^2\,\ell^2} \quad (7)$$

and, again with $\Delta f = f/2$, we arrive at a sensitivity as shown in Figure 1 by the dotted line marked 'uncertainty principle'. From (7) we see that the linear sensitivity limitation is inversely proportional to the arm length ℓ.

The straight line is still safely below the heavy polygon that, as we will see, determines the sensitivity limitation of the GEO design. This is reassuring, and it is one of the great advantages of the interferometer detector over the resonant bars. Unlike in resonant bars (when we want to achieve even the more modest goal of 10^{-21}), there is no necessity here to resort to such hard-to-realize schemes as "quantum non-demolition" or "back-action-evading" ... unless one wants to cut down on arm length. We clearly have here another good argument for km sized arms.

2.2 Thermal Noise

But things become worse if we consider yet other noise contributions. Let us take, as another important example, and also a very fundamental one, thermal vibrations. This example, too, has the advantage that its discussion needs no assumptions about the actual experimental implementation.

2.2.1 Internal thermal motion of mirrors

The thermal motion in the test masses introduces a vibration of the mirror surfaces that – for each relevant mode – can be described by a simple harmonic oscillator, of resonant frequency $f_o = \omega_o/(2\pi)$. The damping is normally assumed to be proportional to velocity, and can be expressed by the quality factor Q. The (linear) spectral density of these motions can be written as

$$\tilde{\delta\ell}(f) = \left(\frac{4kT}{M\omega_o^3}\right)^{\frac{1}{2}} \cdot \left[Q\left(1-\left(\frac{\omega}{\omega_o}\right)^2\right)^2 + \frac{1}{Q}\left(\frac{\omega}{\omega_o}\right)^2\right]^{-\frac{1}{2}}, \qquad (8)$$

with a frequency dependence of the square bracket as shown in Figure 2. It is obvious that we do not want the resonant peak at f_o to occur inside our frequency window of interest, since such a peak would have noise signals that are many orders of magnitude above the signal we want to measure.

We have to make sure that all resonant modes of the mirror substrate are well above our signal-frequency window. We then have to consider only the unavoidable low-frequency tails of theses modes, each of which has a white noise with a spectral density of

$$\tilde{\delta\ell}(f) = \left(\frac{4kT}{MQ\omega_o^3}\right)^{\frac{1}{2}} = \left(\frac{4kT}{\pi^3\rho v_s^3 Q}\right)^{\frac{1}{2}}. \qquad (9)$$

Even with favorable assumptions, $M = 400$ kg, a high mechanical Q of 10^6 (silicon), a resonant frequency of $f_o = 3$ kHz, and a bandwidth of $\Delta f = 1$ kHz, we get a $\delta\ell$ of 10^{-19} m. Considering the number of mirrors and relevant modes involved, a total thermal motion of 10^{-18} m seems a

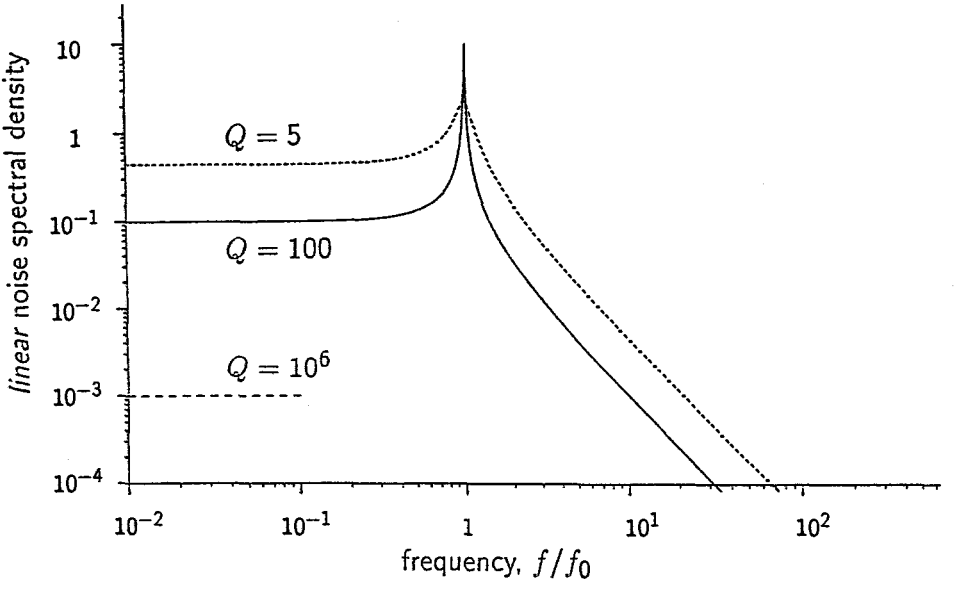

Fig. 2. *Linear* thermal noise spectral densities of oscillators with identical resonant frequency f_0, but different quality factors Q. For the extremely high Q of 10^6, only the level of the subresonant white-noise tail is indicated.

reasonably safe upper estimate, and from this we deduce that we can get near the required goal of $\delta\ell/\ell = 10^{-21}$ only with arm lengths ℓ of a few kilometers, say, 3 km.

In order to avoid the relatively low frequencies of the bending modes, the substrate aspect ratios (thickness over radius) will be chosen close to unity. It is interesting to note that the white noise of (9) then becomes independent of the size of the mirror substrate, the length dependencies in M and ω^3 cancel each other.

Expressing the noise as a (squared) strain spectral density,

$$\tilde{h}^2 \approx \frac{16\,kT}{\pi^3\,\rho\,v_s{}^3\,Q\,\ell^2}\,, \tag{10}$$

and again using the bandwidth $\Delta f = f/2$ and a safety margin of 2.5 to take into account the presence of several modes, we find a sensitivity that is shown in Figure 1 as the dotted line marked 'Thermal Noise (internal modes)'. We can clearly see that for the materials assumed one can just barely keep below the shot noise limit at low frequencies, whereas we have a rather comfortable margin at higher frequencies.

2.2.2 Choice of materials

A look at (10) gives us some indication about what characteristics of substrate materials to look for. The velocity of sound, v_s, enters very strongly, more so than the specific weight ρ. Many substrate materials have much better figures of ρv_s^3 than fused silica, and at the same time also better quality factors Q. Silicon and sapphire (both already available in rather large single crystal ingots), would both have good characteristics, even better (because of sound velocities v_s above 10 km/sec) would be beryllium (toxic!), and – better yet – beryllia and diamond. Beryllia crystals have so far been a militarily classified material, so not very much is known about large specimens. Diamond is being grown by novel processes now, but so far only in thin layers (up to 1 mm). These alternative materials also have advantages in their thermal properties which make them less susceptible to problems associated with heating due to the light beam, as discussed in W. Winkler's contribution [3]. Only the future can tell whether these ideal materials will become available in the sizes required.

2.2.3 Table top interferometer ?

What if we wanted to attain the same sensitivity with a short interferometer of, say, 1 m in length ? Let us look at the second expression in Equation (9). Not very much more than a factor of ten can be gained via the factor ρv_s^3, so for more drastic improvements we would have only the parameters T and Q to play with, both entering under the square root.

We could cool the substrate, by six powers of ten, to $T = 300\mu$K. This certainly is not an *attractive* solution: One would have to make sure that the vacuum system as a whole is even colder than the mirrors, unless we want our top quality mirror surfaces to trap the residual gases. But worse yet, this scheme is *impossible*: Even assuming mirror coatings with absorption losses of only 0.1 ppm, *i.e.* ten times less than the best ones made today, the dissipation on the mirror surfaces could be radiated away only at substrate temperatures of, say, 10 K. Trying to remove the dissipated light power by heat conduction is hard to reconcile with the high demand on seismic isolation.

We could hope for materials with higher Q. The fused silica which we consider for our mirrors is already a material of very high Q ($Q \approx 10^5$), but some better materials are known, particularly if one cools them down to or below liquid helium temperatures. Pure single crystal sapphire is known to have an extremely high Q, perhaps up to something like 10^9. So with the combined efforts (and huge expenses) of sapphire end masses and cryogenics, one might marginally get to the required 6 powers of ten. But with that we have defeated only one enemy, although admittedly a very prominent one.

2.2.4 Thermal motion of pendulums

A very different regime of the resonant curve of Figure 2 applies when we consider the thermal noise of the pendulum suspension. Here the resonant frequency of the 'pendulation mode' is way below our frequency range of measurement, and only the high-frequency tail enters. The sensitivity limit due to the suspension noise is then determined by the noise spectral density

$$\tilde{h}^2 \approx \frac{16\,k\,T\,\omega_0}{m\,Q_S\,\omega^4\,\ell^2}\,. \tag{11}$$

At higher frequencies, its contributions are negligible, but not so at frequencies around 100 Hz. This is seen, again with our GEO specifications in mind, from the dotted (and partly heavy) line marked 'Thermal Noise (suspension)'. For this line, the very favorable assumption of a pendulation Q of as high as 10^8 was assumed.

Such a high Q is not only difficult to obtain, it is also extremely difficult to measure or verify. The Q gives the number of oscillations after which the amplitude of the motion has fallen by a factor $1/e$. At a period in the order of 1 second, 10^8 oscillations will take about 3 years. Not only would a measurement be stretching the patience of the experimenter, also any seismic influences that might add to (or subtract from) the present state of oscillation must be strictly avoided.

Values of Q up to 10^7 have been measured [12]. Although some groups have proposed to measure the decay more directly (by going to yet lower frequencies) the Q assumed here is not a well-established figure, but rather one that can be derived by physical arguments, considering the heavy end mass, and the very tiny area (near the top of the suspension wire) where any dissipation is to be expected.

As we see from Figure 1, this suspension noise already affects the attainable sensitivity, so clearly any reduction in arm length will increase the noise limit, with the inverse of ℓ. No reduction can be afforded, particularly not if detection is intended to reach into the frequency range of 100 Hz. Particularly for the VIRGO project, with its declared aim of measuring down to frequencies below 100 Hz, this suspension noise would pose a serious problem.

2.2.5 'Imaginary spring constant' damping

Recently, the question of how best to represent mechanical damping has been looked at more closely by P. Saulson [13] and others. These researchers have proposed (and observed) a different damping law, and it is not unlikely that this law also applies to the mechanical systems we are dealing with here. Rather than describing the damping by a velocity-proportional term in the differential equation, it is described (in the frequency domain) by

an imaginary (and possibly frequency dependent) component in the spring constant k:

$$k = k_0 \, [1 + i\phi(\omega)] \, . \tag{12}$$

Such a modified friction law will not influence the characteristics of the seismic isolation very substantially, but it does have serious implications on thermal noise. In the standard damping model the thermal displacement noise of the mass, as shown in Figure 2, has the same shape as the transfer function of the harmonic oscillator; this is because the oscillator is driven by a 'white' random force. This case is plotted as the solid line in Figure 3.

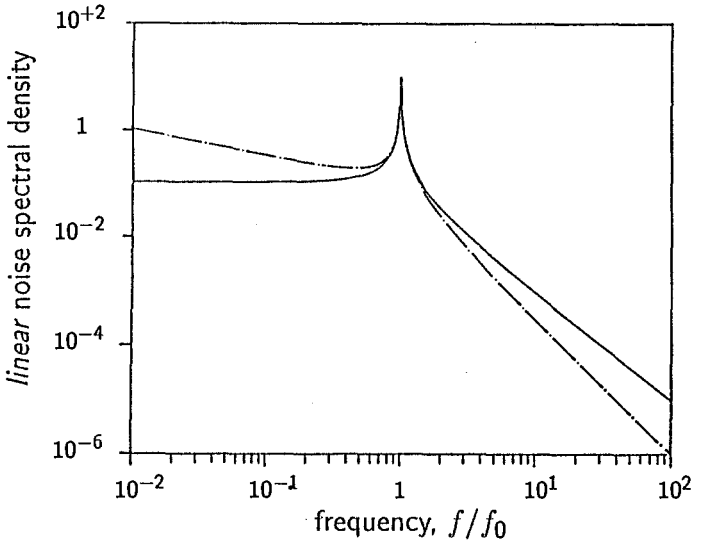

Fig. 3. Thermal noise spectrum of a harmonic oscillator: for velocity-proportional damping (solid curve) and for 'imaginary spring constant' damping (dashed curve), $Q = 100$, after Saulson [13]. *Linear* spectral densities as in Fig. 2.

In the case of the 'imaginary spring constant' damping, however, the thermal driving force turns out to be frequency dependent, and the spectral density of the thermal noise rises towards low frequencies, but falls off more rapidly above resonance [13]. This is shown as the dotted line in Figure 3.

If this modified damping law turns out to be a better description, then the subresonant tails of the substrate vibrations will become a greater problem. In Figure 1, this would make the right-hand line named 'Thermal noise (internal modes)' shallower, and higher at the low-frequency side. The line could then possibly slightly exceed the shot noise curve in the deep trough where photon noise crosses the suspension noise.

If this loss mechanism applies to the pendulum suspension, the 'above-resonant' roll-off in Figure 1 will become even steeper, relaxing the requirements for the suspension's thermal noise. This fact would help the VIRGO project at low frequencies.

2.3 Refractive index fluctuations in the residual gas

The light will travel between the mirrors in highly evacuated tubes, but local changes in the number of molecules remaining in the light path will lead to fluctuations in the refractive index. This changes the optical path between the mirrors without actually changing their positions.

Sudden changes (only these would cause signals in our frequency range of interest) could be caused by local gas eruptions from the tube walls, and even more so from the surfaces of the pumps, particularly getter pumps. A recent research program at PTB in Berlin is making headway in investigating these phenomena at the pressures and time scales that are of interest here.

An estimate of the *lower bound* of the refractive index noise can be made by treating the residual gas as being in thermodynamic equilibrium. Each single molecule traversing the beam of width $2w$ will retard the light phase by an amount that depends on the relative field strength at its location. By averaging over all possible positions and flight directions (and thus interaction times), assuming a Maxwell distribution of the molecule velocity v, and weighting with the (gaussian) field strength, one can calculate the autocorrelation function $R(\tau)$ of the effect of these atoms. It was a pleasant surprise that the analytical treatment, after many integrations of higher transcendental functions (Dawson's integral, modified Bessel functions) of complicated arguments, finally led to the simplest autocorrelation function imaginable, a Lorentzian of the form

$$R(\tau) = \frac{1}{1+(\tau/t_{\mathrm{R}})^2}\,, \tag{13}$$

where the mean interaction time t_{R} is given by $t_{\mathrm{R}} = \sqrt{2}\,w/v_{\mathrm{m}}$ and where the thermal velocity v_{m} stands for the "most probable velocity" $v_{\mathrm{m}} = \sqrt{2kT/m}$ in the Maxwell distribution

$$\frac{4}{\sqrt{\pi}}\,\frac{v}{v_{\mathrm{m}}}\,\mathrm{e}^{-(v/v_{\mathrm{m}})^2}\,. \tag{14}$$

The autocorrelation function $R(\tau)$ leads to a (*single*-sided) power spectral density equalling *twice* the Fourier transform,

$$S(f) \;:=\; 2\int_{-\infty}^{\infty} R(\tau)\,\mathrm{e}^{i\omega\tau}\,\mathrm{d}\tau \;\;=\;\; \frac{1}{f_{\mathrm{c}}}\cdot\mathrm{e}^{-f/f_{\mathrm{c}}}\,. \tag{15}$$

For frequencies well below the 'cut-off frequency' $f_{\mathrm{c}} = 1/(2\pi t_{\mathrm{R}})$ the 'vacuum noise' (15) is almost frequency independent (white). With the usual GEO characteristics, we find the (squared) strain noise to be

$$\tilde{h}^2 \;\approx\; \frac{2\sqrt{2}\,(n_0-1)^2}{N_0\,v_0\cdot\ell\,\bar{w}}\left(\frac{p}{p_0}\right), \tag{16}$$

where n_0 and v_0 are the refractive index and the mean velocity of the gas in question, and N_0 is the number of molecules per unit volume at standard pressure p_0: $(2.7 \times 10^{25}$ molecules/m$^3)$.

The fluctuations (16) determine the maximum allowable gas pressure p for a given strain sensitivity h. To be on the safe side, the vacuum specifications for GEO were laid down as: $< 10^{-8}$ mbar for hydrogen and $< 10^{-9}$ mbar for heavier molecules, such as water and nitrogen. This leads to the line marked "Vacuum Noise" in Figure 1.

The beam width $2w$ is a function of arm length ℓ and wavelength λ, $w \approx \sqrt{\ell\lambda/\pi}$, so in the end we find $\widetilde{h}$ to be proportional to $\ell^{-3/4}$, and for any arm length ℓ below 1 km, the vacuum specifications would have to be tightened beyond what can be done at reasonable cost.

Thus, also the effect of index fluctuations gets reduced by the choice of a longer arm length, mainly due to the better 'averaging' over the longer beam length ℓ and the wider beam diameter $2w$.

3 Seismic Isolation

The third part of this talk deals with the motions of the ground at the site of the interferometer, a noise that is generally termed 'seismic', even though it is not necessarily and exclusively of geophysical origin. Discussing the efforts required to suppress noise due to these seismic motions (see also N. Robertson [14]) will again bring to our attention how extremely small the gravitational wave effects are that we want to measure.

3.1 Seismic noise

3.1.1 The frequency range

We will have to cope with seismic noise over a wide frequency spectrum, a few kHz at the high end, and down to semidiurnal tidal deformation of the earth's crust, or even seasonal variations, at the low end.

Although the gravitational waves are to be measured in a rather limited frequency band only, from, say, 100 Hz to a few kHz, it is nonetheless necessary to consider the effects due to high-amplitude motions (slow drifts) at the extremely low frequencies. They can make the interferometer deviate too far from its ideal point of operation.

It takes a wide spectrum of measuring devices to cover this vast frequency range: (piezo-type) accelerometers for frequencies from kHz downward to 10 Hz; seismometers (mostly velocity-proportional, dip-coil) from 100 Hz down to 1 Hz, in astatized seismometers down to 10^{-1} Hz; and strain measurements (mechanical and laser-interferometric) between two measuring points down to seasonal and secular variations.

3.1.2 The model

For the purposes of designing a seismic isolation system, it has become customary to model the seismic motion of the ground by a (linear) spectral density of the displacement

$$\tilde{x}(f) = 10^{-7} \frac{\mathrm{m}}{\sqrt{\mathrm{Hz}}} \cdot \left[\frac{1\,\mathrm{Hz}}{f}\right]^2 , \tag{17}$$

which describes the frequency dependence quite well over a wide frequency spectrum, certainly from kHz downward to, say, the microseismic frequencies at $\approx 10^{-1}$ Hz. This equation represents a *worst case* motion at the sites being considered, at very quiet sites the amplitudes might be by a factor of ten lower.

3.1.3 Site selection

It is quite natural that in selecting a site for the experiment one will not want to pick a particularly noisy place. Man-made noise arising from traffic, industry, agriculture, etc. contributes most strongly in the frequency range from a few Hz up to, say, 100 Hz. Sufficient distance (one to several km) from busy roads, railroad tracks, heavy industry, mining, (and more than 10 km from military artillery ranges) must be guaranteed.

Coastal regions have strong ground noise contributions from the surf and swell of the sea. The microseismic phenomena, on the other hand, and in particular the 'microseismic storms', at frequencies of 0.1 to 0.15 Hz are believed to stem mostly from the swell of heavy sea, and to travel far into the mainland. There is no obvious way to escape from them, at least not inside Europe.

If one excludes the obviously unappropriate sites, the ground noise (in the range from 1 Hz to 100 Hz) still shows a wide variation depending on the geological formation.

In an earlier series of measurements, Steinwachs [15] had established that the seismic noise increases monotonically with the height of loose rock or scree above bedrock. Furthermore, the man-made noise propagates mainly as a surface wave, and its amplitudes drop rapidly as one goes far enough (at least 10 m, better 100 m) below surface.

Such arguments have again raised the interest in below-ground sites, and particularly into tunnels driven into hard bedrock. With modern techniques, the additional cost over surface installations no longer seems prohibitive. Sites in (seismologically stable) mountain ranges in the German state of Niedersachsen are being considered, and recent seismic measurements [16] in abandoned mines there have established their excellent usefulness.

3.1.4 Isolation required

The degree of isolation required depends on the design parameters of the detector: the sensitivity h_D aimed at, the design frequency $\bar{f}$, and the bandwidth Δf.

From the rather steep roll-off of the ground motion, Equation (17), and from the fact that all methods of isolation improve with frequency, we can expect that isolation at higher frequencies must be a relatively easy job.

Let us take as a first example a modest design sensitivity of $h_D = 10^{-21}$ at $\bar{f} = 1\,\mathrm{kHz}$, $\Delta f = 500\,\mathrm{Hz}$, and $\ell = 3\,\mathrm{km}$. The *rms* ground motion in this band is about $x_{\mathrm{rms}} = 2.5 \cdot 10^{-12}\,\mathrm{m}$. With 4 mirrors involved in the measurement, we see that even for this case a suppression of the ground noise by 10 powers of ten is required to arrive at $h_D = 10^{-21}$.

For the more ambitious goal of $h_D = 10^{-22}$ at 100 Hz, $\Delta f = 100\,\mathrm{Hz}$, we find $x_{\mathrm{rms}} = 1.6 \cdot 10^{-10}$, so we need a suppression by 13 powers of ten, which is much more difficult, particularly at these lower frequencies.

The next sections will give some examples of isolation methods with which one can hope to achieve these suppression values.

3.2 Isolation by pendulums

3.2.1 Single pendulum suspension

The simplest way to isolate a mirror from high-frequency ground motion is to suspend it as a pendulum by one or several thin wires. The method chosen at Garching was to hold the mirror by a thin steel wire sling, as shown schematically in Figure 4.

The damping is typically very low for such a suspension system, particularly if care is taken to avoid friction at the lift-off point of the wire from the mirror, as well as at the suspension point.

The typical frequency response of the mirror motion $\tilde{z}(f)$ for a given 'ground' motion $\tilde{x}(f)$ of the suspension point, the transfer function $\tilde{H}(f) = \tilde{z}(f)/\tilde{x}(f)$, is given by the simple resonant curve of the shape already shown in Figure 2. At frequencies well above the pendulum's resonant frequency f_P, this transfer function (also called transmissibility) rolls off as $(f_P/f)^2$.

For reasonable wire lengths l_P, on the order of 1 m, the resonant frequency is near

$$f_P = \frac{1}{2\pi}\sqrt{\frac{g}{l_P}} \approx 0.5\,\mathrm{Hz}\,. \tag{18}$$

Although at 1 kHz we then have a transmissibility of less than 10^{-6}, this still falls short of even our modest goal of a reduction by 10^{-10}. Much longer pendulums are not practicable, and they could never provide the missing four powers of 10.

3.2.2 Wire resonances

Another obstacle in reaching the desired suppression is the fact that the suspension wires are not massless and thus have their own resonances. The high amplitudes at these resonances are transformed (though reduced by the mass ratio μ/m of effective wire mass μ to mirror mass m) into motions of the mirrors.

The wire pendulum can be treated in close analogy to an electrical transmission line terminated with an inductance (to represent the inertial termination by the impedance $Z_P = i\omega m$ of the pendulum mass m). The characteristic impedance $Z = \sqrt{mg\gamma}$ of the mechanical transmission line is given by the tensile force mg on the wire and the linear mass density γ. The propagation constant $k = \omega/v_{tr}$ is determined by the velocity $v_{tr} = \sqrt{mg/\gamma}$ with which a transverse motion propagates along the wire.

As in an electrical transmission line, the displacement x_P at the termination (pendulum mass) is transformed to the front end (suspension point) via a transformation

$$x_0 = x_P \left(\frac{Z_P}{Z} i \sin kl + \cos kl \right) , \tag{19}$$

and one arrives at the transfer function

$$H(f) := \frac{x_P}{x_0} = \frac{1}{\cos kl - \frac{\omega m}{Z} \sin kl} . \tag{20}$$

The gravest resonance $\omega_P = \sqrt{g/l}$ (the pendulation mode) and the low-frequency transfer function $H(f) = \left(1 - (f/f_P)^2\right)^{-1}$ are easily derived by expanding for $kl \ll 1$. All further resonances (the 'violin string' resonances f_n) can be found from the approximation $kl \approx n\pi$, leading to

$$f_n \approx n\pi f_P \sqrt{\frac{m}{\mu}} , \tag{21}$$

with $\mu = \gamma \cdot l$ the mass of the wire sling (two wires).

Figure 4 shows the suspension system used in the 1986 Garching prototype [17], and the theoretical and experimental transfer functions. For the values used ($m = 1.1$ kg, steel wire 0.1 mm in diameter), the mass ratio m/μ is about 12 500, and the wire resonances are in very good agreement with the measured peaks at multiples of $f_1 \approx 212$ Hz. At these frequencies, the pendulum suspension not only loses its isolation feature, it even enhances the motion of the pendulum over that of the ground.

In between these resonances, the transfer function $H(f)$ provides an isolation that is at best

$$H(f) \approx \frac{Z}{\omega m} = \frac{f_P}{f} \sqrt{\frac{\mu}{m}} , \tag{22}$$

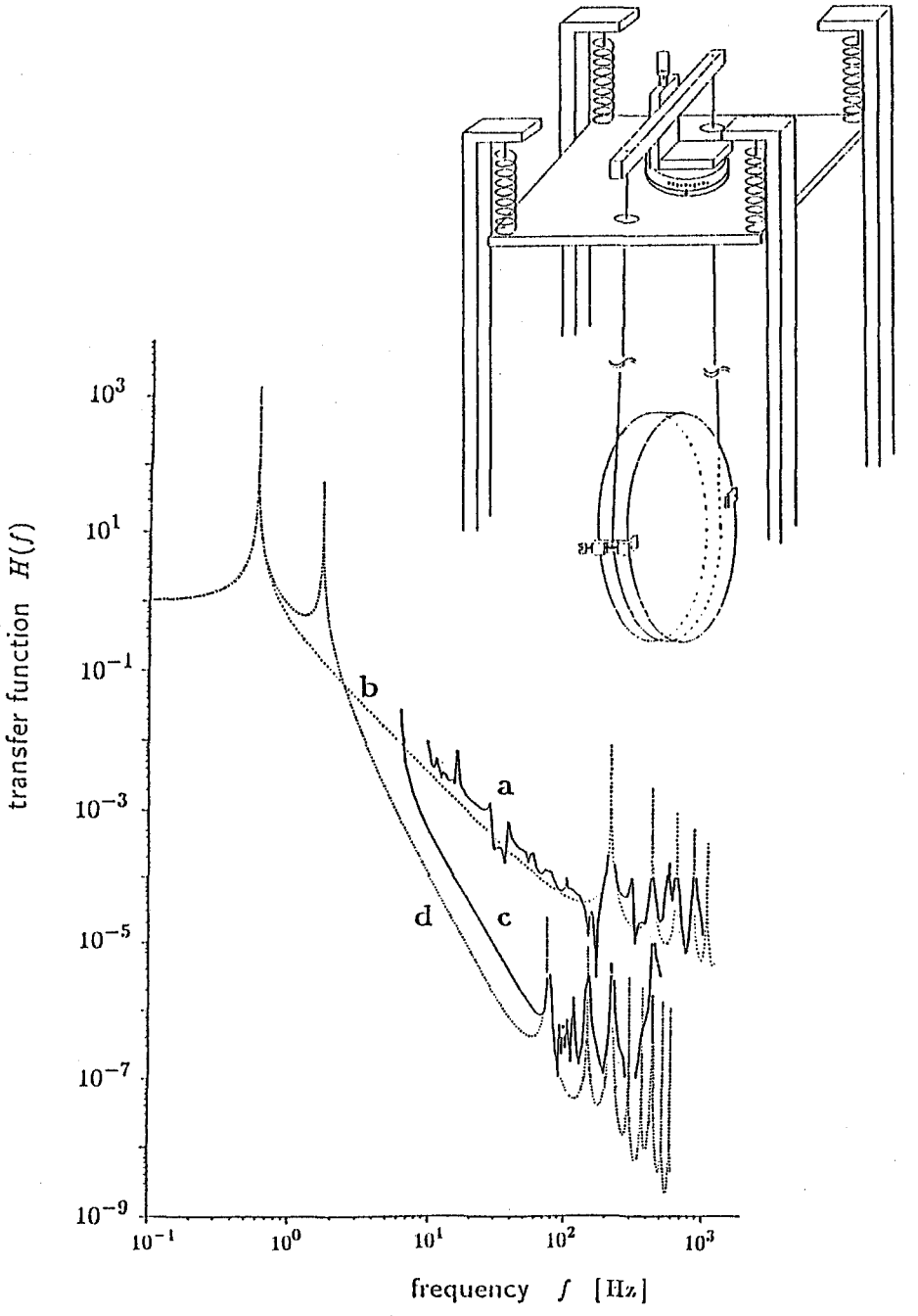

Fig. 4. Top: Schematic view of a (double) pendulum suspension, as in the Garching 30-m prototype of 1986. The upper stage, supported by four coil springs, carries translation and rotation stages for the suspension wire. The upper suspension is 0.10 m in length, the lower 0.72 m.

Bottom: The curves **a** and **b** are the measured and calculated transfer functions for a one-stage pendulum, **c** and **d** for two stages.

i.e. it rolls off only with $1/f$. With the above values and at, say, 550 Hz one finds an isolation by about 10^{-5}.

3.2.3 Local control

The high quality factor Q of the pendulum has the undesirable consequence of a very high resonant peak (see Figure 2). Using a pendulum with increased friction (and thus lower Q) would not be an ideal remedy. For the mirror mass, a low Q would be unacceptable from the thermal noise considerations of section 2.2.4. But also for an intermediate mass in a double pendulum system (see section 3.2.4), assuming simple velocity-proportional damping, the desirable feature of a roll-off with $(f_P/f)^2$ is valid only up to a frequency $f_Q \approx Q f_P$, from then on the further roll-off goes only with the inverse of f.

There is, however, a way to combine a shallow resonant peak with the desired f^{-2} roll-off. This is a frequency-selective damping incorporated in what we call 'local control' servos [18, 17]. This feed-back control measures the relative position of the mirror via a rather crude technique (a shadow sensor) and feeds back a low-passed control signal via coils that act on permanent magnets attached to the mirrors and/or the intermediate masses. The control currents are manipulated such that they damp the pendulum mode, but at higher frequencies leave the f^{-2} roll-off untouched.

These coil-and-magnet controls, now widely used in several prototypes, serve also other purposes, e.g. in very-low frequency control and for optical alignment.

3.2.4 Double pendulum

We demonstrated that a single pendulum cannot provide sufficient suppression even to satisfy our less ambitious goal of 10^{-21} at 1 kHz. An obvious approach is to use multiple pendulums, in the simplest case a two-stage pendulum. This is what was implemented in the Garching 30-m prototype. A schematic view of this suspension scheme is shown in Figure 4. The simple "wireless" model would give a straight $(f_P/f)^4$ roll-off. This would lead to a suppression by better than 10^{-13} at 1 kHz, but such values are not reached in a more realistic model.

The calculation of the transfer function including the wire resonances is a straightforward extension of that for the single pendulum. The comparison of the experimental and theoretical transmissibilities is given by the two lower curves in Figure 4. Again, rather good agreement is found, also in the kink in the roll-off, which (in the valleys between the resonances) is expected to go with f^{-2}.

The wire resonances limit the suppression that we can reach, and although the two-stage pendulum was found sufficient for the 30-m prototype, for the large projects a much better seismic isolation will be needed.

3.2.5 Vertical motion

One severe draw-back of wire pendulums is that their vertical isolation is inferior to the horizontal one. This is easily understood when we consider the resonant frequency for the tensile vertical motion,

$$f_{\mathrm{v}} = \frac{1}{2\pi}\sqrt{\frac{g}{\Delta l}}, \tag{23}$$

where Δl is the elongation of the wire length l_{P} due to the weight of the suspended mirror, and when we compare this with the horizontal pendulation frequency of Equation (18). The elastic, reversible regime for most wire materials will not go much beyond strains $\Delta l/l_{\mathrm{P}}$ of 1 %. So the vertical resonant frequency f_{v} is at least one power of 10 higher than the horizontal resonance. Above this resonance, a single pendulum will be about two powers of ten inferior in its vertical isolation, a double pendulum four powers of 10.

This shortcoming of the wire pendulum is, fortunately, not as dramatic as these numbers suggest, since the interferometer is in first approximation insensitive to mirror motions transverse to the optical axis.

There are, however, several mechanisms that may convert vertical motion into horizontal changes in mirror distance. A very fundamental one is the finite radius of the Earth. The vertical motions of the mirror spaced 3 km subtend an angle of 0.5 mrad, $0.5 \cdot 10^{-3}$, and it is this factor by which vertical motions (pointing to the center of the Earth) are converted into longitudinal distance variations.

There are many structural features that can also transform vertical, or tilting, motions into horizontal ones. In all cases, the conversion factor will be quite small, typically perhaps in the order 10^{-2}, but this just about compensates the inferior vertical isolation of a single pendulum stage. The vertical motions are not an unsolvable problem, but they must always be kept in mind in the design of the isolation system.

3.2.6 Multiple pendulums, VIRGO

A very ambitious scheme is being developed (and being tested) for the VIRGO project. The aim of this project is to be able to measure at even lower frequencies than 100 Hz. This makes achieving a good seismic isolation even more important (as well as more difficult).

The mirror masses are suspended by a chain of seven pendulums in series, with a total height of about 7 m, the total vacuum chamber towering 12 m. The individual stages are made from air springs, and they have vertical resonant frequencies that come quite close to the ones of the horizontal ('pendulation') modes. This is an important advantage over wire pendulums.

The roll-off with f^{-14} has been verified over a limited frequency range, and very impressive transmissibilities have been measured. Only a very sensitive interferometric measurement will be able to establish how good the transfer function is at higher frequencies. This 7-stage pendulum system is expected to come close to, but not yet quite reach, the very ambitious isolation specifications for VIRGO. The addition of an 'inverted pendulum' at the top of the 7-stage pendulum seems to be provide the required additional isolation.

3.3 Isolation via lead-and-rubber stacks

3.3.1 Stacks

As we have seen, the double pendulum cannot sufficiently isolate the mirrors, even for the relaxed goals of $h \approx 10^{-21}$ at 1 kHz. There is, however, the possibility additionally to isolate the suspension point via stacks made up of alternating layers of heavy (e.g. lead) bricks and a soft, elastic material (e.g. rubber or elastomers). Such stacks have been used very successfully in the seismic isolation of resonant bar gravitational wave antennas where very impressive seismic isolation values have been achieved. Such stacks have been used in interferometer prototypes at Glasgow and later at Caltech, and are now being implemented in the Garching 30-m prototype.

Stacks of such alternating layers have some features that are similar to those of the multiple pendulums. The more stages one uses, the steeper is the roll-off at frequencies sufficiently above the highest resonant mode of the stack. (This highest mode, incidentally, is the mode in which the heavy layers have alternating direction of motion).

3.3.2 Damping of stacks

Unlike the (multiple) pendulums, the stacks are typically systems of relatively high internal losses, *i.e.* of low Q. Values of Q between 1 and 10 are typical. An immediate consequence of this low Q is that the roll-off goes with f^{-n}, where n is the number of stages, and not with f^{-2n} as in the case of multiple pendulums.

Although the steeper roll-off with f^{-2n} would be more desirable, it would not outweigh the advantage of having the elastic layers made out of a very lossy material. One could easily achieve the high compliance (the 'softness') for instance with metallic coil springs, but these would, at higher frequencies, have their own internal resonances, entirely ruining the stack's isolation characteristics.

3.3.3 RAL stacks

A set of lead-and-rubber stacks was designed by Rutherford Appleton Laboratories, to be tested and used in an upgrade of the Garching 30-m prototype.

Boundary conditions in the design were the inner diameter (1000 mm) and the limited free headroom in the Garching vacuum tanks. A top plate (from which the double pendulum system is suspended) is supported by four stacks at the four corners. The stacks can have up to five stages and a total height of 86 mm.

Each lead brick has a mass of 4 kg. The rubber springs are cylindrical, 25 mm in diameter and 40 mm high (unloaded). There are four such rubber 'springs' in each layer between two lead bricks, at each of the four corner 'substacks'.

The bricks have cylindrical recesses of half the brick's thickness so that the bricks are supported in the plane of their center of mass. This is to avoid that tilting motions convert into horizontal displacements.

Measurements of the transfer function (the transmissibility) performed at RAL are shown as the solid curve of Figure 5. These measurements were made with vertical driving forces acting on the stack's bottom plate. The 'gravest' (*i.e.* the lowest) resonance, at about 6 Hz, was below the frequency range covered at RAL. The total number of peaks (4) is identical with the number of stages.

The roll-off at frequencies above, say, 40 Hz is quite steep, and a suppression of 10^{-4} was reached at about 70 Hz.

3.3.4 Transmissibility of stacks

Using a simple one-dimensional model of such stacks, assuming rigid metal bricks, with elastic, lossy rubber springs in between, one can derive a recursive scheme with which one can easily calculate the transmissibility for any given number of stages (n). The stages need not be identical, and some dependence of the rubber compliance on the total load can also be incorporated.

The dotted line in Figure 5 shows the computer calculation of the transfer function, taking the mass (4 kg) and the specified vertical stiffness (22 N/mm, or a compliance of 0.045 mm/N for a single rubber spring). Only the loss factor, or the Q, was fitted, such that the height of the resonant peaks was similar to the measured data. Quite good agreement is reached for the values of the resonant frequencies as well as for the roll-off.

Numerical tests were made with the two different damping laws already discussed in 2.2.5. There was some influence on the relative height of the gravest resonant peak as the number of stages was increased: they were of equal height for the 'imaginary spring constant' case, but dropped with rising n under the 'velocity-proportional' law.

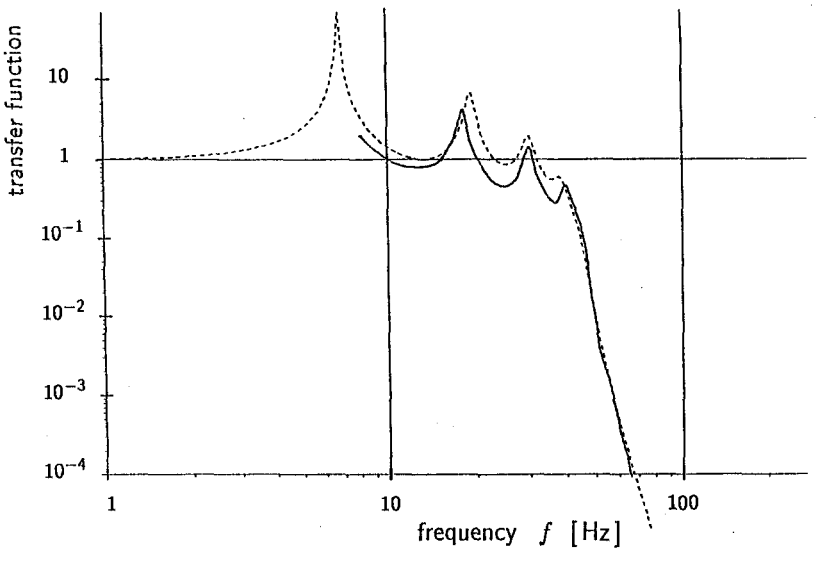

Fig. 5. Transmissibility (for vertical motion) of 4-stage RAL stack: RAL measurements (solid line) and MPQ model calculations (dashed line). Transmissibility is down to 10^{-4} at about 70 Hz.

The influence of the damping law on the roll-off was not substantial in the frequency range of Figure 5, the superior suppression of the 'imaginary spring constant' case would become manifest only at somewhat higher frequencies.

3.3.5 Garching stack measurements

A set of four RAL stacks was used in one Garching end tank to support the top frame used for pendulum suspension. The four bottom plates supporting the four stacks at each corner can be driven (with a swept sine) in horizontal or vertical direction by four low-voltage piezo vibrators (PI P-844.20).

Piezoelectric accelerometers measured the spectra of the horizontal motion at the top frame and at the piezo-driven bottom plate. The quotient of these values gives the transmissibility. Typical measurements are presented in Figures 6 and 7, in both cases for horizontal motion.

The seismic isolation even of a stack having only three stages, as shown in Figure 6, was found fully sufficient for the present Garching 30-m prototype. Therefore it was decided to use them in the current upgrade in all three tanks.

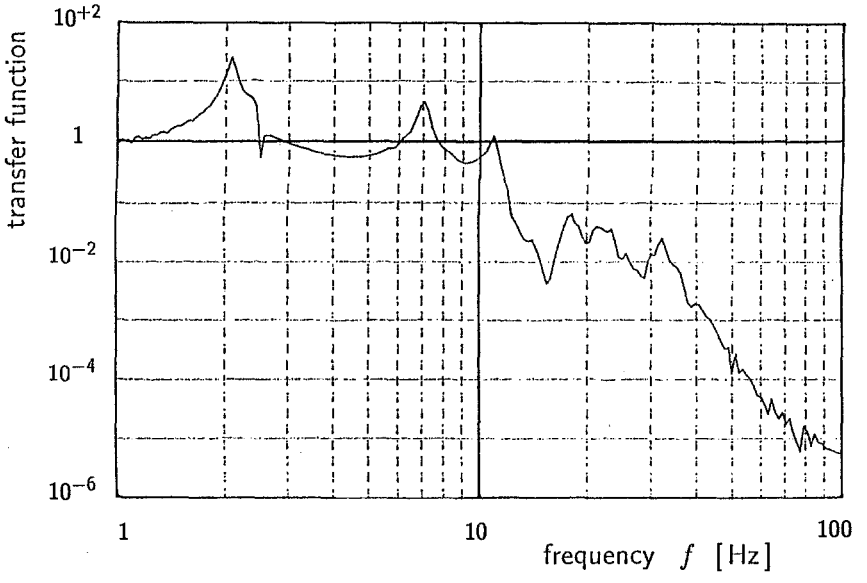

Fig. 6. Horizontal transmissibility of 3-stage RAL stack at Garching, in the frequency range from 1 to 100 Hz. The resonant frequencies are near 2, 8, and 11 Hz. The heavy horizontal line marks unity transfer function, the spacing of the horizontal lines is a factor of 10 each, the transmissibility is down to 10^{-5} at about 80 Hz.

3.3.6 Acoustic bypass

Figure 7 shows the frequency range from the lowest resonance (near 2 Hz) up to about 100 Hz. The observed data (resonant frequencies, roll-off) were in good agreement with the expected behaviour, and they nourished the hope that the roll-off would continue to higher frequencies. This was not the case, but rather a resurging curve (shown in Figure 7, upper curve) was observed, rising almost towards unity again at 1 kHz.

This behaviour is difficult to explain from the elastic properties of the rubber springs, and some observations pointed to acoustic coupling through the air. The measurements were repeated in vacuum. This took some experimental effort: preamplifiers for operation in vacuum and avoiding crosstalk between the feedthroughs and leads of the high-power piezo drives and the (low-voltage) accelerometer signals.

The results were convincing: with the tank evacuated, the rise at frequencies above 100 Hz disappeared. This is shown in the lower curve of Figure 7. Intuitively, one would not have expected the heavy lead bricks to be so strongly excited by acoustics in the air.

But still the results were not fully satisfactory, as the transfer function seemed to level off at something like 10^{-6} at best, rather than continuing the steep roll-off. This behaviour is not yet fully understood, but the stack investigations had to be broken off for the moment as all three vacuum tanks

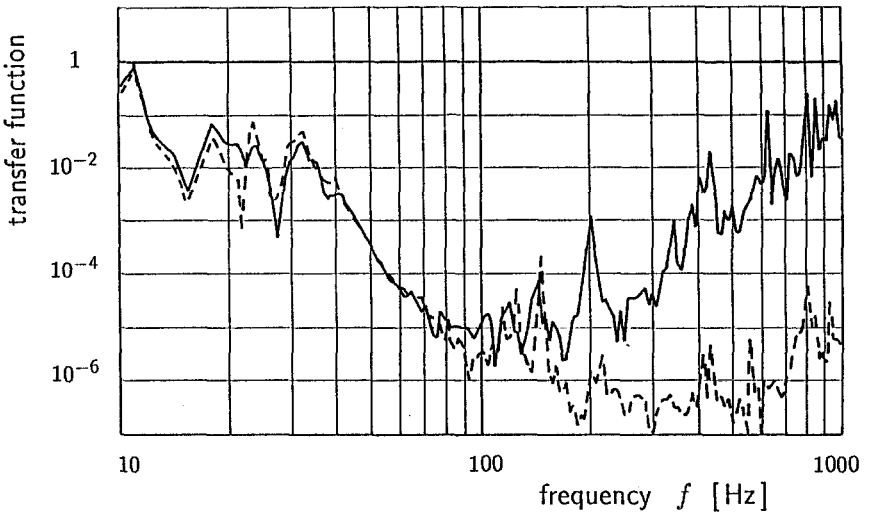

Fig. 7. Horizontal transmissibility of 3-stage RAL stack at Garching, in the range from 10 Hz to 1 kHz, measured in air (upper curve) and in vacuum (lower curve). Isolation is down to 10^{-4} at about 60 Hz, and in vacuum reaches 10^{-6} in a frequency range from 150 to 500 Hz.

are now used for installing an upgrade of the 30-m prototype. It is hoped that the stack tests will be resumed soon, in a dedicated test tank.

The results of the stack measurements, although incomplete, were quite encouraging. Suppression factors of 10^{-5} and even 10^{-6} are readily achieved even with a three-stage stack. Inside the frequency range up to 200 Hz, a further improvement with the number of stages can be predicted, the GEO specifications can no doubt be readily met.

One further important feature of stacks is that they can easily be made to have vertical isolation of similar quality as the horizontal one; or an even better one if one wants to make up for the shortcoming of the wire pendulum.

At higher frequencies it will yet have to be determined whether the measured levelling off in the transfer function is a physical effect of the stack structure, or whether it is an artefact in the very delicate measurement. But in any case, the specifications are not at risk, as the ground motion according to (17) rolls off with f^{-2}.

3.4 Ultra-low frequencies

3.4.1 Seismic noise at ultra-low frequencies

There are various causes for very slow drifts in the distance between points as far apart as our arm lengths (*i.e.* 3 km). Most prominent examples are the lunar and solar tides of the solid earth, but there are also seasonal and meteorological fluctuations.

Above frequencies $f_\ell = v_s/\ell$, determined by the velocity of sound v_s in the ground and the arm length ℓ, the motions at the two end points can be regarded as uncorrelated. This frequency f_ℓ is of the order of 1 Hz. For drifts at much lower frequencies, the relative motion $\delta\ell$ is given by the low-frequency strain in the ground multiplied by the arm length.

These slow drifts are not inside the frequency range of measurement, but their effects need to be suppressed nevertheless. This is so because their relatively large swings could drive the interferometer far out of its proper point of operation.

For an example, let us consider the diurnal tides of the solid earth. Although the strains are quite small, in the order of a few 'nanostrain', *i.e.* strains of a few 10^{-9}, this is, after all, twelve to thirteen powers of 10 larger than the gravity-wave signals we want to measure. The frequency, on the other hand, is only some six to eight powers of ten lower than our GW signal frequencies.

3.4.2 Dynamic range of control elements

Large low-frequency swings have grave consequences for the application of control signals. These have to do with the finite "dynamic range" of such control elements.

The first limitation (and the most fundamental one) is due to the electronic noise in the control amplifiers, say the current drivers for the coil-and-magnet control elements.

Typically, low-noise amplifiers have a noise current that is at best ten powers of 10 below the full current. A control element designed to compensate drifts up to 1 mm will thus introduce a displacement noise in the order of 10^{-13} m at low frequencies, and even including the reduction of this noise due to the inertial mass of the mirror, this will still be several powers of 10 above the noise allowed when we want to measure strains of $h = 10^{-21}$ or better.

An important consequence of this is that the control signals compensating the slow large drifts are not allowed to be applied to the mirrors themselves, but only to a stage higher up, say to the intermediate mass in a double pendulum system. A control system along these lines is just being investigated at Garching.

A second limitation arises because the coils of the coil-and-magnet systems are mounted to masses that in themselves are not totally quiet, that

in some cases will have the full seismic noise of the ground. The coil-and-magnet systems can be operated such that the force exerted on the magnet is *in first order* independent of the position of the coil. But in the case of large low-frequency swings, this optimal point of operation cannot be maintained. Recent investigations have led to configurations of the coils in which also the *second derivative* of force with position vanishes. Such coils allow much larger swings, and they will be incorporated into the current upgrade of the Garching prototype.

4 Conclusion

It has been demonstrated that mechanical noise of various kinds can impose serious limitations on the sensitivity of interferometric gravitational wave detectors. In particular, the discussion has shown that in order to cope with these the arm length has to be chosen on the order of a few kilometers.

But a large arm length alone is no guarantee for success. Each of the noise sources discussed will require special attention and will call for an optimal design of the relevant parts of the detector.

In experiments on various prototypes the world over, and in detailed design studies, the feasibility of a gravitational wave detector with strain sensitivities of 10^{-21} or even 10^{-22} has been demonstrated. This is why applications for funding of the large detectors can no longer be considered premature, they reflect the high state of the art already achieved.

References

1. G. Schäfer: *Gravity-Wave Astrophysics*, Springer Lecture Notes in Physics, *this issue.*
2. K. Danzmann *et al.*: *The GEO-Project: A Long-Baseline Laser Interferometer for the Detection of Gravitational Waves*, Springer Lecture Notes in Physics, *this issue.*
3. W. Winkler *et al.*: *The Optics of an Interferometric Gravitational-Wave Antenna*, Springer Lecture Notes in Physics, *this issue.*
4. R. Mönchmeyer, G. Schäfer, E. Müller, and R.E. Kates, Astron. Astrophys. **246** (1991) 417.
5. *Proposal for a Joint German-British Interferometric Gravitational Wave Detector*, J. Hough, B. J. Meers, G. P. Newton, N. A. Robertson, H. Ward, G. Leuchs, T. M. Niebauer, A. Rüdiger, R. Schilling, L. Schnupp, H. Walther, W. Winkler, B. F. Schutz, J. Ehlers, P. Kafka, G. Schäfer, M. W. Hamilton, I. Schütz, H. Welling, J. R. J. Bennett, I. F. Corbett, B. W. H. Edwards, R. J. S. Greenhalgh, and V. Kose, Max-Planck-Institut für Quantenoptik Report No. MPQ 147 (1989).
6. R. Vogt, R.W.P. Drever, F.J. Raab, K.S. Thorne, and R. Weiss: *Laser interferometer gravitational-wave observatory*, Proposal to the NSF, 1989.

7. A. Giazotto, A. Brillet, *et al.*, *The VIRGO Project*, Proposal to INFN, 1989.
8. R.J. Sandeman, D.G. Blair, and J. Collett, *Australian International Gravitational Research Centre*, proposal to the Australian Government, (Australian National University, 1991).
9. N. Kawashima, Proc. Sixth Marcel Grossmann Meeting, Kyoto, 1991.
10. B.J. Meers, Phys. Rev. D **38**, 2317 (1988); K.A. Strain and B.J. Meers, Phys. Rev. Lett. **66**, 1391 (1991).
11. C.M. Caves, Phys. Rev. D **23**, 1693 (1981).
12. W. Martin, Ph.D. Thesis, Glasgow (1978).
13. P.R. Saulson, Phys. Rev. D **42**, 2437 (1991).
14. N.A. Robertson: *Seismic Isolation*, in: *The Detection of Gravitational Radiation*, Ed. D. Blair, Cambridge University Press, 1991.
15. M. Steinwachs, Geol. Jahrbuch **E 3** (1974) 1–59.
16. A. Brüge and G. Lauer, PTB report PTB–MA–22 (1992).
17. D. Shoemaker, R. Schilling, L. Schnupp, W. Winkler, K. Maischberger, and A. Rüdiger, Phys. Rev. D **38**, 423 (1988).
18. H. Billing, K. Maischberger, A. Rüdiger, R. Schilling, L. Schnupp, and W. Winkler, J. Phys. E: Sci. Instrum. **12** (1979) 1043–1050.

List of Authors

A. Rüdiger, J. Chen, K. Danzmann, P.G. Nelson, T.M. Niebauer, R. Schilling, K.A. Strain, L. Schnupp, H. Walther, W. Winkler, *Max-Planck-Institut für Quantenoptik, D-8046 Garching, Germany*; J. Hough, A.M. Campbell, C.A. Cantley, J.E. Logan, B.J. Meers, E. Morrison, G.P. Newton, D.I. Robertson, N.A. Robertson, S. Rowan, K.D. Skeldon, P.J. Veitch, and H. Ward, *Department of Physics and Astronomy, University of Glasgow, Glasgow, UK*; H. Welling, P. Aufmuth, I. Kröpke and D. Ristau, *Laser-Zentrum and Institut für Quantenoptik, Universität Hannover, D-3000 Hannover, Germany*; J.E. Hall, J.R.J. Bennett, I.F. Corbett, B.W.H. Edwards, R.J. Elsey, and R.J.S. Greenhalgh, *Rutherford Appleton Laboratory, Chilton, Didcot, UK*; B.F. Schutz, D. Nicholson, and J.R. Shuttleworth, *Department of Physics, University of Wales, Cardiff, UK*; J. Ehlers, P. Kafka, and G. Schäfer, *Max-Planck-Institut für Astrophysik, D-8046 Garching, Germany*; H. Braun, *Bauabteilung der Max-Planck-Gesellschaft, D-8000 München, Germany*; V. Kose, *Physikalisch-Technische Bundesanstalt, D-3300 Braunschweig and D-1000 Berlin, Germany.*

Fermion and Boson Stars

Norbert Straumann
Institute for Theoretical Physics, University of Zurich

1 Introduction

In this first lecture I shall make a few scattered remarks on fermion and boson stars. We shall start with simple qualitative considerations based on the uncertainty and exclusion principles. These lead to rough estimates for the conditions under which heavy atoms, white dwarfs, neutron stars and mini-boson stars collapse. Some of these estimates can be turned into rigorous bounds when a semirelativistic Hamiltonian provides an appropriate description. It is then also possible to give a rigorous justification for a semiclassical treatment (e.g. à la Thomas-Fermi). This is good to know, because in situations when general relativistic effects become important (neutron stars, mini-boson stars) there is no other approximation scheme available.

As a preparation to the talks by Neugebauer and Herold I shall also give an introduction to the mean field approach for obtaining an equation of state for neutron star matter at very high densities.

Very recently we have constructed topological boson stars for the non-linear sigma model of Skyrme. I will briefly discuss their structure and the results of the stability analysis which we have just finished. (Black hole solutions for the same model will be described in my second lecture.) Finally I will make a few remarks about gauge boson stars and their instability.

Historically, our subject has a long tradition. It is very remarkable that the quantum statistics of identical particles found its first application in astrophysics. In the same year when Schrödinger discovered his equation Fowler realized that "the black-dwarf material is best likened to a single gigantic molecule in its lowest quantum state" and he developed the nonrelativistic theory of white dwarfs. It was afterwards recognized independently by several people [Frenkel (1928), Stoner(1930), Chandrasekhar(1931), Landau(1932)] that relativistic kinematics weakens the quantum mechanical kinematic energy (zero-point pressure) to the extent that there is a limiting mass for white dwarfs. (For a brief historical discussion and references I refer to [1].)

Many of the considerations and methods in the first part of my talk can already be nicely illustrated for atoms. Let me show this for the problem of stability (collapse condition) of a heavy neutral atom.

Consider first the non-relativistic Hamiltonian including all Coulomb inter-

actions

$$H^{NR} = \frac{1}{2m}\sum_{i=1}^{N} \mathbf{p}_i^2 - \sum_{i=1}^{N} \frac{Ze^2}{|\mathbf{x}_i|} + \sum_{i<j} \frac{e^2}{|\mathbf{x}_i - \mathbf{x}_j|}. \tag{1}$$

With the help of the variational principle one can find a rigorous upper bound for the ground state energy $E_0(N)$ of the form $(N = Z)$

$$E_0^{NR} \leq -CZ^{7/3}\alpha^2 mc^2, \tag{2}$$

with

$$C > 0.447. \tag{3}$$

(For details see the Appendix in Ref.[2]. I recall also that the Thomas-Fermi approximation gives $E_0^{NR} \approx -0.77Z^{7/3}\alpha^2 mc^2$.)

We make now the unjustified but interesting assumption that the dominant relativistic effects can be described by the following semi-relativistic Hamiltonian

$$H^R = \sum_{i=1}^{N} \sqrt{\mathbf{p}_i^2 c^2 + m^2 c^4} + V, \tag{4}$$

where the potential energy V is the same as in (1). The ground state energy E_0^R of H^R can be bounded in a useful manner by E_0^{NR}, as has been noted only quite recently [2]. Indeed, from $(a^{1/2} - b^{1/2})^2 \geq 0$ for $a = (\mathbf{p}_i^2 c^2 + \tilde{m}c^4)/2\tilde{m}c^2$ and $b = \tilde{m}c^2/2$ we obtain the operator inequality

$$H^R \leq \frac{1}{2}N\tilde{m}c^2 + \frac{1}{2}N\frac{m^2 c^2}{\tilde{m}} + H^{NR}(m \to \tilde{m}), \tag{5}$$

for any $\tilde{m} > 0$. From this and the bound (2) it follows that

$$E_0^R(N) \leq \tilde{m}[\frac{1}{2}Nc^2 - CN^{7/3}\alpha^2 c^2] + \frac{1}{2}N\frac{m^2 c^2}{\tilde{m}}. \tag{6}$$

Since $\tilde{m}$ can be arbitrarily large, collapse will occurs if $(N = Z)$:

$$Z^{4/3}\alpha^2 > \frac{1}{2C}, \quad \text{i.e. for } \alpha > 1.05Z^{-2/3}. \tag{7}$$

A similar method can be used for white dwarfs and boson stars (see Sect.2).

2 Semirelativistic fermion and boson stars

I begin with some remarks on white dwarfs (WD), which have in recent years also attracted the attention of mathematical physicists. It was shown in particular by Lieb and Yau [3] that the Chandrasekhar theory is the limit of quantum mechanics as the number of electrons $N \to \infty$ and the gravitational constant $G \to 0$ in such a way that $GN^{2/3}$ remains constant. A corresponding result has also been established rigorously for boson stars.

For WD's we can ignore general relativistic effects, except for stability considerations close to the critical mass for gravitational collapse. Matter in a WD is

completely ionized. The Coulomb forces establish local neutrality to a very high accuracy. For this reason the Coulomb interactions play energetically almost no role. (The corrections can be estimated and are on the few percent level.) The spatial distribution of nuclei and hence their momentum distribution is much the same as those of the electrons. Therefore, the ground state energy of a WD with N electrons and N_Z nuclei with charge Ze and mass m_Z can be estimated as follows

$$E_0(N) \approx \min\{N\sqrt{p^2+m^2} - \frac{1}{2}(\frac{N}{Z})^2 G m_Z^2 \frac{p}{N^{1/3}\hbar}\}, \tag{8}$$

where we have made use of the Pauli principle for the contribution of the gravitational energy. (The average momentum p of a particle satisfies $p \geq N^{1/3}\hbar/R$, where R is the radius of the star, because there can be at most one electron in a de Broglie cube $(\hbar/p)^3$.) The minimum in (8) exists only for $(m_Z = Am_N)$:

$$N < N_f := (\frac{Gm_N}{2\hbar c})^{-3/2}(\frac{Z}{A})^2. \tag{9}$$

For the ground state the momentum and energy are, respectively,

$$p_0 \approx mc(\frac{N}{N_f})^{2/3}[1-(\frac{N}{N_f})^{4/3}]^{-2}, \tag{10}$$

$$E_0(N) \approx Nmc^2[1-(\frac{N}{N_f})^{4/3}]^{1/2}. \tag{11}$$

For $N > N_f$ the expression in the curly bracket of (8) is not bounded from below, since in the extrem relativistic limit it becomes $N[1-(\frac{N}{N_f})^{4/3}]pc$. Therefore, the system collapses for $N > N_f$. The mass M_f corresponding to N_f gives the following estimate of the Chandrasekhar mass

$$M_{Ch} \approx \frac{N_f}{Z} m_Z \approx 2.8 \frac{M_{Pl}^3}{m_N^2}(\frac{Z}{A})^2, \tag{12}$$

where M_{Pl} is the Planck mass. In the Chandrasekhar theory the prefactor in the last expression is replaced by the number 3.1.

Before I say what has been shown rigorously about this limit, I want to make it clear that the Chandrasekhar theory is just the Thomas-Fermi theory for the WD, considered as one big "atom" with about 10^{57} electrons. Indeed, the Chandrasekhar equation can also be obtained as follows. Let $\varepsilon_F(r)$ be the local Fermi energy for the electrons. The potential energy per electron in the gravitational potential φ, satisfying

$$\triangle\varphi = 4\pi G\rho \tag{13}$$

(ρ=matter density), is equal to $\mu_e m_N \varphi$, where μ_e is the average value of A/Z. Therefore, the equilibrium condition is

$$\varepsilon_F(r) + \mu_e m_N \varphi(r) = \mu = \text{const}, \tag{14}$$

where μ is the chemical potential. If $p_F(r)$ denotes the local Fermi momentum of the electrons we have

$$\varepsilon_F(r) = m_e c^2 \sqrt{1+x^2}, \quad x := \frac{p_F}{m_e c}, \tag{15}$$

$$\rho = Bx^3, \quad B = \frac{8\pi m_e^3 c^3 m_N}{3h^3} \mu_e = 0.97 \times 10^6 \mu_e (g/cm^3). \tag{16}$$

If we solve (14) for φ and insert the result into the Poisson equation (13), we obtain with (15) and (16) for the quantity $z(r) := \sqrt{1+x(r)^2} = \varepsilon_F(r)/m_e c^2$ the following relativistic Thomas-Fermi equation for a sperically symmetric star:

$$\frac{1}{2}\frac{d}{dr}(r^2 \frac{dz}{dr}) = -4\pi \frac{GB}{c^2} \frac{m_N}{m_e} (z^2-1)^{3/2}. \tag{17}$$

This is identical to the well-known Chandrasekhar equation for WD's.

The Chandrasekhar theory can be derived as a limit of a more fundamental theory. Our qualitative discussion makes it plausible that we can chose as starting point the following semirelativistic Hamiltonian (for simplicity we consider N electrons and the same number of protons):

$$H_N = \sum_{i=1}^{N} [\sqrt{\mathrm{p}_i^2 + m^2} - m] - \sum_{i<j} \frac{Gm_N^2}{|\mathrm{x}_i - \mathrm{x}_j|}, \tag{18}$$

where x_i and p_i are canonically conjugate variables ($\hbar = c = 1$). It is now natural to compare the quantum energy

$$E^Q(N) := \inf \, spec \, H_N \tag{19}$$

with the semiclassical energy of the Thomas-Fermi theory:

$$E^{TF}(N) = \inf\{\mathcal{E}^{TF}(n) : \, n \geq 0, \int n d^3x = N, \, n \in L^{4/3}(\mathrm{R})\}, \tag{20}$$

where

$$E^Q(N) = \frac{1}{\pi^2} \int_0^{p_F(n)} [\sqrt{p^2+m^2} - m] p^2 dp - Gm_N^2 \frac{1}{2} \int \frac{n(\mathrm{x}) n(\mathrm{x}')}{|\mathrm{x}-\mathrm{x}'|} d^3x d^3x', \tag{21}$$

with

$$p_F(n) = (3\pi^2 n)^{1/3}. \tag{22}$$

One of the main results of Lieb and Yau [3] is the following

Theorem(fermions): Fix the quantity $\tau = Gm_N^2 N^{2/3}$ at some value below the critical value τ_c of the Chandrasekhar theory ($\tau \approx 3.1$). Then

$$\lim_{N\to\infty} E^Q(N)/E^{TF}(N) = 1. \tag{23}$$

If $\tau > \tau_c$, then

$$\lim_{N\to\infty} E^Q(N) = -\infty. \tag{24}$$

As a corallary one can show that for the critical numbers N_f^Q and N_f^{TF} for stability we have

$$\lim_{G\to 0} N_f^Q / N_f^{TF} = 1. \tag{25}$$

This demonstrates that we can study H_N by means of its semiclassical approximation. This is, of course, not really surprising. Indeed, corrections to the Thomas-Fermi approximation are of the order $N^{-1/3}$, i.e., of the order 10^{-19} for $N \approx 10^{57}$. (In contrast to this tiny number for WD's, corrections of the order $Z^{-1/3}$ for atoms are not negligible.)

* * *

We now turn to *boson stars* and begin again with a simple qualitative consideration. Using the uncertainty relation we obtain instead of (8) the following rough estimate of the ground state energy for *free* bosons ($m_Z \to m$)

$$E_0(N) \approx \min_p \{N\sqrt{p^2+m^2} - \frac{1}{2}Gm^2N^2p\}. \tag{26}$$

Obviously, the curly bracket is not bounded below if

$$N > N_b := 2(Gm^2)^{-1}. \tag{27}$$

In the opposite case $N < N_b$ the minimum is attained for the average momentum

$$p_0 \approx m[(\frac{N_b}{N})^2 - 1]^{-1/2} \tag{28}$$

and the ground state energy becomes

$$E_0(N) \approx Nm[1 - (\frac{N}{N_b})^2]^{1/2}. \tag{29}$$

The critical mass for boson stars is thus

$$M_{crit}^{(bosons)} \approx 2\frac{M_{Pl}^2}{m}. \tag{30}$$

In comparison to fermion stars we loose a factor M_{pl}/m. This may change drastically once we introduce interactions (see Sect. 4).

How do we now arrive at a semiclassical treatment of boson stars? Well, we just have to follow the Hartree procedure known from atomic physics. For the ground state energy of H_N for N bosons we use the variational principle with trial wave functions of the form

$$\Psi(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \prod_{i=1}^{N} \varphi(\mathbf{x}_i), \tag{31}$$

with a normalized real non-negative wave function $\varphi \in L^2(\mathbf{R}^3)$. We have (with $p^2 = -\Delta$)

$$\begin{aligned}(\Psi, H_N\Psi) &= N < \varphi|[\sqrt{p^2+m^2} - m]|\varphi > \\ &\quad -\frac{N(N-1)}{2}Gm^2 \int \frac{\varphi(\mathbf{x})^2\varphi(\mathbf{x}')^2}{|\mathbf{x}-\mathbf{x}'|}d^3x d^3x'.\end{aligned}$$

Let us introduce the "number density" $n(\mathbf{x}) = N\varphi(\mathbf{x})^2$. Neglecting the difference between N and $N-1$ we arrive at the following semiclassical (Hartree) functional

$$\mathcal{E}^H(n) = < n^{1/2}|[\sqrt{p^2+m^2} - m]|n^{1/2} > -\frac{1}{2}Gm^2 \int \frac{n(\mathbf{x})n(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} d^3x d^3x' \qquad (32)$$

and this functional has to be minimized with the subsidary condition

$$\int n(\mathbf{x}) d^3x = N. \qquad (33)$$

The corresponding Euler-Lagrange equation is the following Hartree-like equation for $\phi(\mathbf{x}) = n(\mathbf{x})^{1/2}$

$$[\sqrt{p^2+m^2} - m - Gm^2 U(\mathbf{x})]\phi(\mathbf{x}) = -\mu\phi(\mathbf{x}), \qquad (34)$$

where $\mu > 0$ is the chemical potential (Lagrange multiplier) and

$$U(\mathbf{x}) = \int \frac{n(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|} d^3x'. \qquad (35)$$

The critical number N_b^H for collapse has to be computed numerically. As far as I know this has not been done. One can, however, show [3] that

$$1.27(Gm^2)^{-1} < N_b^H < 2.7(Gm^2)^{-1}. \qquad (36)$$

Lieb and Yau have also proven that the ground state energy $E^H(N)$ for the Hartree functional and the critical number N_b^H agree asymptotically with the corresponding quantum values. More precisely, the following theorem holds:

Theorem (bosons): Let

$$E^H(N) = \inf\{\mathcal{E}^H(n) : n \geq 0, \int n = N,\ |p|^{1/2} n^{1/2} \in L^2(\mathbf{R})\} \qquad (37)$$

and fix $\omega := Gm^2 N$ below the critical value ω_c ($1.27 < \omega_c < 2.7$). Then

$$\lim_{N\to\infty} E^Q(N)/E^H/N) = 1. \qquad (38)$$

If $\omega > \omega_c$ then $\lim_{N\to\infty} E^Q(N) = -\infty$. Furthermore,

$$\lim_{G\to 0} \frac{N_b^Q}{N_b^H} = 1. \qquad (39)$$

* * *

I conclude this section in deriving rather stringent upper and lower bounds for the critical mass of boson stars described by the semirelativistic Hamiltonian (we include now the rest mass)

$$H_N = \sum_{i=1}^{N} \sqrt{\mathbf{p}_i^2 + m^2} - \sum_{i<j} \frac{Gm^2}{|\mathbf{x}_i - \mathbf{x}_j|}. \qquad (40)$$

In a first step we show that [4]

$$M_N := \inf \, spec\, H_N < N[\frac{m^2+\tilde{m}^2}{2\tilde{m}}] + \frac{\tilde{m}}{m} E_N^{NR}, \tag{41}$$

where $\tilde{m}$ is any mass and $E_N^{NR} = \inf \, spec\, H_N^{NR}$ for the *nonrelativistic* Hamiltonian

$$H_N^{NR} = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m} - \sum_{i<j} \frac{Gm^2}{|\mathbf{x}_i - \mathbf{x}_j|}. \tag{42}$$

Now we have already used in the introduction the operator inequality

$$H_N < N[\frac{m^2+\tilde{m}^2}{2\tilde{m}}] + (\sum_i \frac{\mathbf{p}_i^2}{2\tilde{m}} - \sum_{i<j} \frac{Gm^2}{|\mathbf{x}_i - \mathbf{x}_j|}). \tag{43}$$

Because of the different masses in (43) we need the κ-dependence of the ground state energy $E_0^N(\kappa)$ of the Hamiltonian

$$H_N(\kappa) = \sum \frac{1}{2}\mathbf{p}_i^2 - \kappa \sum_{i<j} \frac{1}{|\mathbf{x}_i - \mathbf{x}_j|}. \tag{44}$$

Applying the canonical transformation $\mathbf{x}_i' = \lambda \mathbf{x}_i$, $\mathbf{p}_i' = \frac{1}{\lambda}\mathbf{p}_i$ we find the scaling behavior

$$H_N(\kappa) = \frac{1}{\lambda^2} H_N(\lambda\kappa) \tag{45}$$

and thus $E_0^N(\lambda\kappa) = \lambda^2 E_0^N(\kappa)$. Hence we can conclude from (43) that the inequality (41) holds.

For large $\tilde{m}$ the right hand side of (41) is $\approx \tilde{m}[\frac{N}{2} + E_N^{NR}/m]$ and therefore M_N becomes $-\infty$ if

$$E_N^{NR} + \frac{Nm}{2} < 0. \tag{46}$$

In the opposite case ($E_N^{NR} + \frac{Nm}{2} > 0$) we find from (41) the useful upper bound

$$M_N < Nm[1 + \frac{2E_N^{NR}}{Nm}]^{1/2}. \tag{47}$$

This shows that we need now an upper bound for E_N^{NR}, which we derive with the help of the variational principle. As a trial wave function we take

$$\Psi(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \text{norm.} \prod_{i=1}^{N} \phi(\lambda \mathbf{x}_i), \tag{48}$$

where ϕ is a normalized one-particle wave function. We have

$$\frac{(\Psi, H\Psi)}{(\Psi, \Psi)} = \lambda^2 \frac{N\alpha}{2m} - \lambda \frac{1}{2} N(N-1) Gm^2 \beta, \tag{49}$$

with

$$\alpha = \int |\nabla \phi|^2 d^3x, \quad \beta = \int \frac{|\phi(\mathbf{x})|^2 |\phi(\mathbf{x}')|^2}{|\mathbf{x} - \mathbf{x}'|} d^3x \, d^3x'. \tag{50}$$

Minimizing with respect to λ gives

$$E_N^{NR} < -\frac{1}{8}\frac{\beta^2}{\alpha}N(N-1)^2G^2m^5. \tag{51}$$

With a wave function ϕ propotional to $e^{-\sigma r}$ one finds easily $\beta^2/\alpha = (5/8)^2$ and thus

$$E_N^{NR} < -0.049N(N-1)^2G^2m^5. \tag{52}$$

One can improve this bound a bit by chosing a two-parameter wave function

$$\phi \propto e^{-\sigma\sqrt{\mu^2r^2+1}}, \tag{53}$$

leading to [5]

$$E_N^{NR} < -0.0542N(N-1)^2G^2m^5. \tag{54}$$

The Hartree calculation gives [6]

$$E_N^{NR} < -0.05426N(N-1)^2G^2m^5. \tag{55}$$

Therefore, the criterion (46) shows that we have relativistic collapse when

$$N-1 > 3.04(\frac{M_{pl}}{m})^2. \tag{56}$$

Otherwise (47) leads to the upper bound

$$M_N < Nm[1-2\times 0.054(N-1)^2G^2m^4]^{1/2}. \tag{57}$$

Taking the maximum with respect to N gives the following upper bound for the critical mass

$$M_{crit}^{bosons} < 1.51\frac{M_{pl}^2}{m}. \tag{58}$$

Martin and collaborators have also derived a stringent lower bound for M_{crit} [4, 5]:

$$M_{crit}^{bosons} \geq 4[3\sqrt{3}Gm]^{-1} = 0.77(Gm)^{-1}. \tag{59}$$

Here we establish a somewhat less restrictive bound. To this end we make use of the inequality [7]

$$\frac{1}{|\mathbf{x}_i-\mathbf{x}_j|} \leq \frac{\pi}{2}\frac{|\mathbf{p}_i-\mathbf{p}_j|}{2} \tag{60}$$

(where the operators are considered as quadratic forms). Together with $|\mathbf{p}_i - \mathbf{p}_j| \leq |\mathbf{p}_i| + |\mathbf{p}_j|$ we then obtain

$$H_N \geq \sum_i[\sqrt{\mathbf{p}_i^2+m^2} - \frac{\pi}{4}Gm^2(N-1)|\mathbf{p}_i|]. \tag{61}$$

The square bracket on the right hand side is non-negative for

$$N-1 < \frac{4}{\pi}(Gm^2)^{-1}, \tag{62}$$

and this guarantees relativistic stability. If this condition holds we can minimize the square bracket in (61) with respect to each $|\mathbf{p}_i|$ and obtain the inequality

$$\begin{aligned} H_N &\geq Nm[1-(\frac{\pi}{4}Gm^2(N-1))^2]^{1/2} \\ &> Nm[1-(\frac{\pi}{4}Gm^2N)^2]^{1/2}. \end{aligned} \tag{63}$$

The last expression has a maximum for $N=(Gm^2)^{-1}2\sqrt{2}/\pi$, which satisfies the bound (62). Therefore, we obtain the bound

$$M_{crit}^{bosons)} \geq \frac{2}{\pi}(Gm)^{-1} = 0.64(Gm)^{-1}, \tag{64}$$

which is already very good.

It turns out that the improved lower bound (59) is above the maximum mass of a general relativistic boson star (see Sect. 4). This should, however, not be too surprising, since the semirelativistic Hamiltonian neglects, for instance, the fact that the kinetic energy of the bosons is also a source of gravitational fields.

3 Equation of state above nuclear densities

For neutron stars the semirelativistic Hamiltonian (40) provides at best a rough model. First of all, the degree of compactness is so great that we have to use general relativity for a quantitative description. In addition, nuclear forces become important and do not lead only to small corrections as the Coulomb interactions for WD's. From what has been said earlier, it should, however, be clear that we can savely use a semiclassical approach in which we use Einstein's field equations together with an equation of state (EOS) for neutron star matter.

Up to about nuclear densities the EOS is reasonably well known, but the central densities of neutron stars can be almost an order of magnitude higher. One is thus in a very difficult regime. (Below nuclear densities the nuclear gas is rather dilute and at much higher densities we would have asymptotic freedom of QCD. For neutron star matter we are just between these two limits where things become simple.)

Among the many approaches for arriving at an equation of state I discuss here only the mean field theory of Walecka and coworkers [8].

The main features of this approach can already be seen in a simple model. Assume that a neutral scalar meson field (ϕ) and a neutral vector field (V_μ) couple to the baryon current by interaction terms

$$g_s\bar{\psi}\psi\phi \quad \text{and} \quad g_v\bar{\psi}\gamma^\mu\psi V_\mu. \tag{65}$$

Thus the Lagrangian of this simple model is

$$\begin{aligned} \mathcal{L} &= \bar{\psi}[\gamma^\mu(i\partial_\mu - g_vV_\mu) - (M-g_s\phi)]\psi + \\ &\quad \frac{1}{2}(\partial_\mu\phi\partial^\mu\phi - m_s^2\phi^2) - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}m_v^2V_\mu V^\mu; \\ F_{\mu\nu} &= \partial_\mu V_\nu - \partial_\nu V_\mu. \end{aligned} \tag{66}$$

For very large densities one can replace the meson-field operators by their expectation values (mean field approx.):

$$< \phi > \equiv \phi_0, < V_\mu > \equiv \delta_{\mu 0} V_0. \tag{67}$$

We consider a static uniform system. Then the meson field equations give

$$\phi_0 = \frac{g_s}{m_s^2} < \bar{\psi}\psi > \equiv \frac{g_s}{m_s^2}\rho_s, \tag{68}$$

$$V_0 = \frac{g_v}{m_v^2} < \psi^\dagger \psi > \equiv \frac{g_v}{m_v^2}\rho_B. \tag{69}$$

In this mean field approximation the nucleon field operator satisfies a *linear* equation:

$$[-i\gamma^\mu \partial_\mu + g_v \gamma^0 V_0 + M^*]\psi = 0, \quad M^* = M - g_s\phi_0 \text{ (effective mass)}, \tag{70}$$

and the Lagrangian density takes the form

$$\mathcal{L}_{MFT} = \bar{\psi}[i\partial\!\!\!/ - g_v\gamma^0 V_0 - M^*]\psi - \frac{1}{2}m_s^2\phi_0^2 + \frac{1}{2}m_v^2 V_0^2. \tag{71}$$

The quantization is straightforward. We are mainly interested in the energy density

$$\varepsilon = < T_{00} > \tag{72}$$

and the pressure

$$p = \frac{1}{3} < T_{ii} > . \tag{73}$$

Note also

$$\rho_B = \frac{\gamma}{(2\pi)^3}\int_0^{k_F} d^3k = \frac{\gamma}{6\pi^2}k_F^3 \tag{74}$$

($\gamma = 4$ for symmetric nucleon matter, $\gamma = 2$ for neutron matter). The mean field ϕ_0 (or M^*) is determined selfconsistently:

$$M^* = M - g_s\phi_0 = M - \frac{g_s^2}{m_s^2}\rho_s = M - \frac{g_s^2}{m_s^2}\frac{\gamma}{(2\pi)^3}\int_{|k|\leq k_F} d^3k \frac{M^*}{E^*(k)}, \tag{75}$$

where $E^*(k) = \sqrt{k^2 + M^{*2}}$. This leads to a transcendental equation for a given $k_F(\rho_B)$.

Once $\varepsilon(\rho_B)$ is known, the pressure is determined also by

$$p = \rho_B^2 \frac{\partial}{\partial \rho_B}(\varepsilon/\rho_B). \tag{76}$$

It turns out that only the ratios g_s^2/m_s^2 and g_v^2/m_v^2 enter the equation of state. These parameters are fixed such that for nuclear matter we get the correct binding energy and saturation density:

$$(\frac{E - BM}{B})_0 = -15.75 MeV, \tag{77}$$

$$k_F^0 = 1.42\, fm^{-1} \ (\gamma = 4). \tag{78}$$

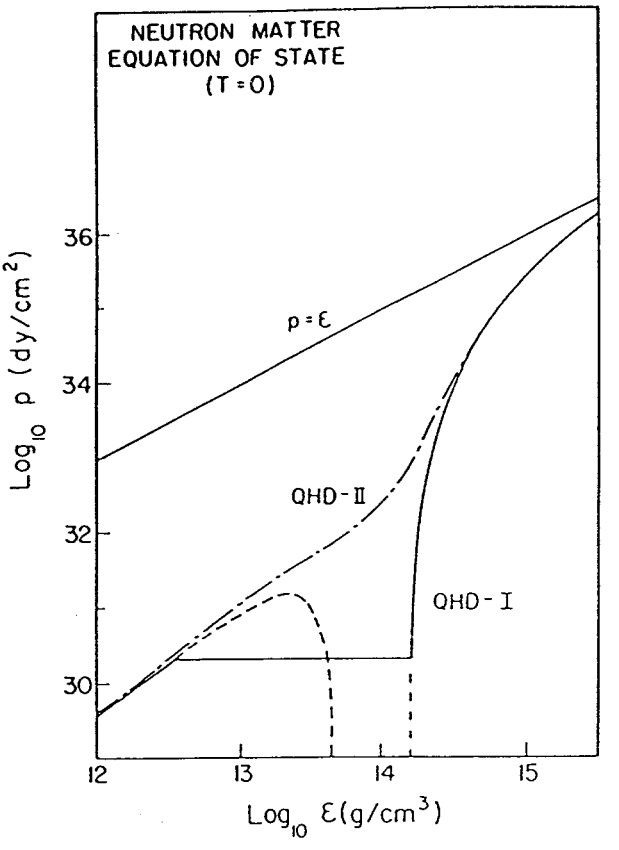

Fig. 1. Equation of state for neutron matter in the mean field theory with and without ρ mesons. The horizontal segment for QHD-I is the result of the Maxwell construction in the region of the phase transition. (From Serot and Walecka [8].)

This gives

$$C_s^2 \equiv g_s^2(\frac{M^2}{m_s^2}) = 267.1, \tag{79}$$

$$C_v^2 \equiv g_v^2(\frac{M^2}{m_v^2}) = 195.9. \tag{80}$$

With these values one can compute the equation of state for neutron matter. The result is shown in Fig.1. In this simple model there is a first order phase transition (similar to the liquid-gas transition in the van der Waals' eq. of state). At very high densities the velocity of sound approaches the velocity of light. a

One can make the model more realistic by including for instance also ρ-mesons (charged vector mesons). In one such model (QHD-II of Serot and Walecka) the ρ meson stiffens the equation of state at relatively low density and causes the gas-liquid phase transition to disappear (see Fig.1).

The neutron star mass is, however, changed only slightly, as is shown in

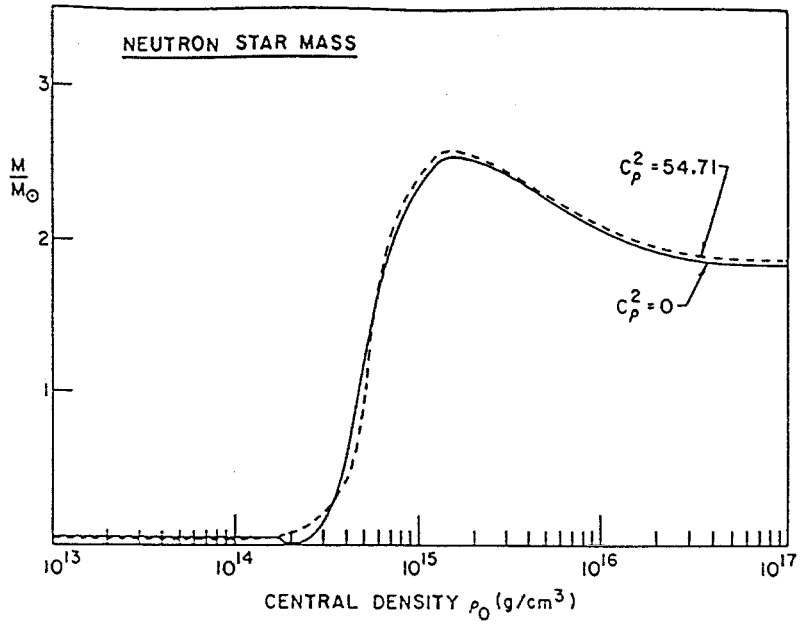

Fig. 2. Neutron star mass as a function of the central density for the equations of state in Fig.1. (From Serot and Walecka [8].)

Fig.2. In particular, the maximum mass changes only by about 1%:

$$M_{max} \approx 2.60 M_\odot \ \ (instead\ of\ 2.57 M_\odot). \tag{81}$$

These results should at least give some impression of the true equation of state. This problem is obviously one of the most difficult in physics, because we are not far from the QCD-phase transition and have thus a *strongly coupled dense many hadron system*, which we cannot treat on the basis of QCD. Lattice calculations cannot be done for non-vanishing chemical potentials. Basically this comes from the following circumstance. For $T = 0$ and $\mu \neq 0$ the energy density ε has the form

$$\varepsilon = c_2 \mu^4 + c_1 \mu^2 / a^2 + O(a^2) \ \ (\text{a=lattice spacing}). \tag{82}$$

The second term (which can be anticipated on dimensional grounds), creates a problem in the limit of vanishing lattice constant. So far, no satisfactory way out has been found.

The mean field approach can be criticized on several grounds. It gives too large incompressibilities for nuclear matter and the effective mass is strongly density dependent and surprisingly low ($M^*/M \approx 0.55$) for symmetric matter at saturation. Quantum fluctuation corrections turn out to be small at high densities. At nuclear densities they are, however, large. For instance, the effective mass is increased to $M^* \approx 0.85 M$.

4 General relativistic boson stars

In recent years a large variety of boson star configurations and their stability have been analysed. It is not my intention to give here a comprehensive survey (various review articles are in preparation [9]). I shall discuss briefly three types of boson stars for the following matter models: (i) complex self-coupled scalar field ϕ, (ii) non-linear sigma model, (iii) non-abelian gauge fields.

We use again a semiclassical approach and look only for spherically symmetric configurations. Let me first give the basic formulae which are in common to all models.

We can use Schwarzschild-like coordinates with the following form of the metric

$$g = e^{2a}dt^2 - [e^{2b}dr^2 + r^2(d\vartheta^2 + \sin^2\vartheta d\varphi^2)], \tag{83}$$

where a and b are only functions of r for the equilibrium configurations. (For radial pulsations a and b depend only on r and t.) On symmetry grounds, the energy-momentum tensor T^{μ}_{ν} has the general form

$$(T^{\mu}_{\nu}) = \begin{pmatrix} \rho & & O & \\ & -p_r & & \\ O & & -p_\perp & \\ & & & -p_\perp \end{pmatrix} \tag{84}$$

As a consequence of the Bianchi identity and $\nabla \cdot T = 0$ only the tt and the rr components of the Einstein field equations are independent. These read explicitly

$$\frac{1}{r^2} - e^{-2b}\left(\frac{1}{r^2} - \frac{2b'}{r}\right) = 8\pi G\rho, \tag{85}$$

$$\frac{1}{r^2} - e^{-2b}\left(\frac{1}{r^2} + \frac{2a'}{r}\right) = -8\pi G p_r, \tag{86}$$

where a prime denotes the derivative with respect to r. We note also the r-component of $\nabla \cdot T = 0$:

$$(\rho + p_r)a' = -p_r' + 2(p_\perp - p_r)/r. \tag{87}$$

Instead of the metric variable b we use also the mass function $m(r)$ defined by

$$e^{-2b} = 1 - \frac{2m(r)}{r}. \tag{88}$$

In terms of this the field equation (85) reads

$$m' = 4\pi G\rho r^2. \tag{89}$$

The Schwarzschild mass is equal to $m(\infty)$. Instead of (85) and (86) we have the independent field eqations (89) and

$$e^{-2b}(a' + b') = 4\pi G(\rho + p_r). \tag{90}$$

a) Complex scalar field

The Lagrangian of the matter model is taken to be

$$\mathcal{L} = \nabla_\mu \phi^* \nabla^\mu \phi - U(|\phi|^2), \tag{91}$$

with

$$U(|\phi|^2) = m^2|\phi|^2 + f^2|\phi|^4 + \cdots, \tag{92}$$

whose energy-momentum tensor is

$$T_{\mu\nu}(\phi) = \nabla_\mu \phi^* \nabla_\nu \phi + \nabla_\nu \phi^* \nabla_\mu \phi - g_{\mu\nu}\mathcal{L}(\phi). \tag{93}$$

This model has also a conserved particle number current, corresponding to the global $U(1)$ invariance:

$$J^\mu = i(\phi^* \nabla^\mu \phi - \nabla^\mu \phi^* \phi). \tag{94}$$

Even for a static star we must allow for a periodic time dependence of ϕ

$$\phi(r,t) = \frac{1}{\sqrt{2}}\sigma(r)e^{-i\omega t}. \tag{95}$$

That such a time dependence is indeed necessary is discussed in Ref. [10]. (ω plays the role of a Lagrange multiplier belonging to the conservation of the total particle number.)

The components ρ, p_r, $p_\perp$ are easily found to be

$$\rho = \frac{1}{2}\omega^2\sigma^2 e^{-2a} + \frac{1}{2}\sigma'^2 e^{-2b} + U(\sigma), \tag{96}$$

$$p_r = \rho - 2U(\sigma), \quad p_\perp = p_r - \sigma'^2 e^{-2b}. \tag{97}$$

The Lagrangian of the σ-field is

$$\mathcal{L} = \frac{1}{2}[e^{-2a}\omega^2\sigma^2 - e^{-2b}\sigma'^2 - U(\sigma)], \tag{98}$$

whose Euler-Lagrange equation reads (note that $\sqrt{-g} \propto r^2 e^{a+b}$):

$$\sigma'' + (a' - b' + \frac{2}{r})\sigma' = e^{2b}\frac{\partial U}{\partial \sigma} - \omega^2 e^{2b-2a}\sigma. \tag{99}$$

Many authors have integrated the coupled field equations (85), (86) and (99), with the explicit expressions (96), (97) for ρ , p_r and

$$U(\sigma) = \frac{1}{2}m^2\sigma^2 + \frac{1}{4}f^2\sigma^4 + \cdots. \tag{100}$$

In Fig.3 we show the result for the free case ($U(\sigma) = \frac{1}{2}m^2\sigma^2$). The Schwarzschild mass M is plotted as a function of $\sigma(0)$. The particle number N belonging to the current J^μ, i.e.,

$$N = \omega \int_0^\infty e^{b-a}\sigma^2 4\pi r^2 dr, \tag{101}$$

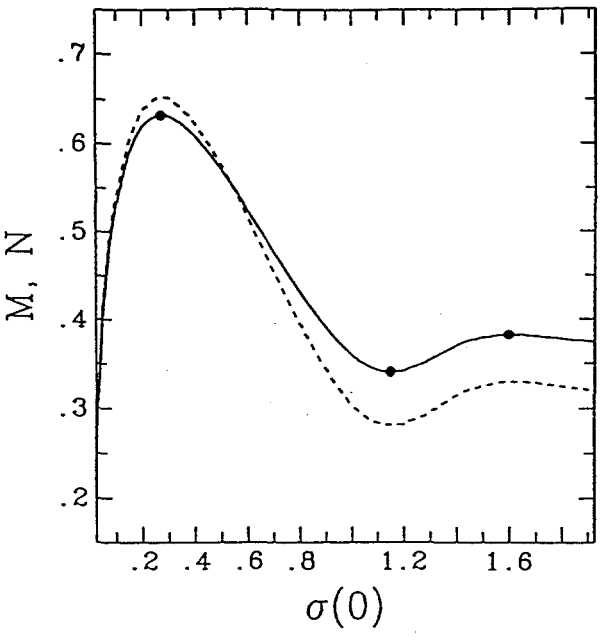

Fig. 3. Boson star mass in units of M_{Pl}^2/m (solid line) and particle number in units of M_{Pl}^2/m^2 (dashed line) as a function of the central density $\sigma(0)$ in units of $(4\pi G)^{1/2}$ for free bosons.

is also shown. As remarked earlier, the maximum mass (M_{max}) is *below* the lower bound (59) for a semirelativistic approach.

For non-vanishing selfcoupling ($f \neq 0$) the critical mass for boson stars can be much larger. In [11] it is shown that for $fM_{Pl}/m \gg 1$ one has the simple formula

$$M_{max} \approx 0.06 \frac{M_{Pl}^3}{(m/f^{1/2})^2} = f(\frac{0.1GeV}{m})^2 M_\odot. \qquad (102)$$

This shows that the maximum mass of a boson star is not very much smaller than the Chandrasekhar mass for a fermion mass $m/f^{1/2}$.

The interaction introduces (for $fM_{Pl}/m \gg 1$) a new scale into the structure of the boson star, given by $R \sim f\frac{M_{Pl}}{m}\frac{1}{m}$. Correspondingly, the typical magnitude of the bose field is drastically reduced and becomes $\approx m/f^{1/2}$.

One can demonstrate [12] that the critical points for M and N occur at the *same* values of $\sigma(0)$. This allows one to conclude that the rising branch, starting with small values of $\sigma(0)$, is *stable*. I do not go into the details of the stability analysis of this type of stars since that has been done in the contribution of Schunck [13]. (For a review, see also Ref [9].) I will, however, say something about the more difficult stability analysis of the next two types of boson stars.

b) Self-gravitating Skyrmions

In an attempt to find stable black holes with hair we have recently con-

structed self-gravitating sigma-model stars, which are examples of topological boson stars [14].

The basic field variable of the non-linear sigma model of Skyrme is an SU(2) valued function $U(x)$ on the space-time manifold (with metric g and covariant derivative ∇_μ). In terms of the quantities

$$A_\mu = U^\dagger \nabla_\mu U, \qquad F_{\mu\nu} = [A_\mu, A_\nu] \tag{103}$$

the SU(2) $\times$ SU(2) invariant Lagrangian density is given by [15]

$$\mathcal{L} = -\frac{f^2}{4}\mathrm{Tr}(A_\mu A^\mu) + \frac{1}{32g^2}\mathrm{Tr}(F_{\mu\nu} F^{\mu\nu}), \tag{104}$$

where f and g are two coupling constants. Note that the Lie algebra valued 1-form $A = A_\mu dx^\mu$ is just the pull-back of the Maurer-Cartan form Θ on SU(2) by the map $U : A = U^*(\Theta)$. Therefore, the Maurer-Cartan equation implies

$$dA + A \wedge A = 0 \tag{105}$$

and the 2-form F corresponding to $F_{\mu\nu}$ is also equal to $-dA$. It is then also obvious that

$$\omega = \frac{1}{24\pi^2}\mathrm{Tr}(A \wedge A \wedge A) \tag{106}$$

is a closed 3-form, whose Hodge dual is the well-known topological current of the Skyrme model. The normalization in (106) is chosen such that ω is – up to a sign – equal to the pull-back of the normalized invariant volume form on SU(2). If the asymptotics is such that we can consider U as map from the compactified three-dimensional space (for fixed time) into SU(2), then the integral of ω gives the degree (winding number) of this map and is thus an integer, which is a homotopy invariant. (In particle physics applications this topological integer is interpreted as the baryon number [15].)

Variation of the metric in the matter action leads to the following expression for the energy-momentum tensor of the Skyrme model

$$T_{\alpha\beta} = -\frac{f^2}{2}\mathrm{Tr}[A_\alpha A_\beta - \frac{1}{2}g_{\alpha\beta}A_\mu A^\mu] + \frac{1}{8g^2}\mathrm{Tr}[F_{\alpha\mu} F_{\beta\nu} g^{\mu\nu} - \frac{1}{4}g_{\alpha\beta}F_{\mu\nu} F^{\mu\nu}]. \tag{107}$$

We look now for static spherical symmetric solitons. For $U(x)$ we make the familiar hedgehog ansatz

$$U(x) = \cos\chi(r) + i\sin\chi(r)\,\tau\cdot\hat{\mathbf{x}}, \tag{108}$$

where τ_i denote the Pauli matrices. The corresponding Lie algebra valued 1-form A is then given by

$$-iA = \tau_r\chi' dr + [\sin\chi\cos\chi\,\tau_\vartheta + \sin^2\chi\,\tau_\varphi]d\vartheta + [-\sin^2\chi\,\tau_\vartheta + \sin\chi\cos\chi\,\tau_\varphi]\sin\vartheta d\varphi, \tag{109}$$

where $\tau_r, \tau_\vartheta, \tau_\varphi$ are the spherical projections of the Pauli matrices and a prime denotes the derivative with respect to r. This is of the general form of a spherically symmetric SU(2) connection.

For the relevant component of the energy-momentum tensor we find, after some calculations,

$$\rho = \frac{1}{2}f^2[e^{-2b}(\chi')^2 + \frac{2}{r^2}\sin^2\chi] + \frac{1}{g^2}\frac{\sin^2\chi}{r^2}[2e^{-2b}(\chi')^2 + \frac{\sin^2\chi}{r^2}], \quad (110)$$

$$p_r = \frac{1}{2}f^2[e^{-2b}(\chi')^2 - \frac{2}{r^2}\sin^2\chi] + \frac{1}{g^2}\frac{\sin^2\chi}{r^2}[2e^{-2b}(\chi')^2 - \frac{\sin^2\chi}{r^2}]. \quad (111)$$

The non-linear matter equation follows from the matter action which is

$$S_M = \int \mathcal{L}\sqrt{-g}d^4x = \int L_M(r)dr, \quad (112)$$

with

$$\frac{1}{4\pi}L_M = \frac{1}{2}f^2[r^2(\chi')^2e^{a-b} + 2\sin^2\chi e^{a+b}] + \frac{1}{2g^2}\sin^2\chi[2(\chi')^2e^{a-b} + \frac{\sin^2\chi}{r^2}e^{a+b}]. \quad (113)$$

The corresponding Euler-Lagrange equation is

$$f^2[(e^{a-b}r^2\chi')' - e^{a+b}\sin 2\chi] = \frac{1}{g^2}[-2(e^{a-b}\chi'\sin^2\chi)' + (\chi')^2e^{a-b}\sin 2\chi + e^{a+b}\frac{\sin^2\chi}{r^2}\sin 2\chi]. \quad (114)$$

We introduce the dimensionless radial variable

$$x = gfr \quad (115)$$

and the dimensionless coupling constant

$$\kappa = 4\pi(\frac{f}{M_{Pl}})^2. \quad (116)$$

(The coupling constant g is dimensionless.)

For the numerical integration we need the behavior of the functions a, b and χ near the origin. We first note that $a(0)$ will be adjusted such that $a(r)$ vanishes asymptotically. (This fixes the time coordinate t.) Furthermore, we must require that $\chi(0) = \pi$, in order that $U(x)$ in (108) is well-defined at $r = 0$. The expansion of the various quantities around the origin is then determined by $\gamma := \chi'(0)$. With the help of the field equations one finds :

$$\begin{aligned} a &= a(0) + \frac{\kappa}{2}\gamma^4x^2 + O(x^4), \\ b &= \frac{\kappa}{2}\gamma^2(1+\gamma^2)x^2 + O(x^4), \\ \chi &= \pi + \gamma x + \frac{\gamma^3}{1+2\gamma^2}\left(\frac{1}{10}\kappa(3+6\gamma^2+2\gamma^4) - \frac{1}{15}(2+\gamma^2)\right)x^3 + O(x^4). \end{aligned} \quad (117)$$

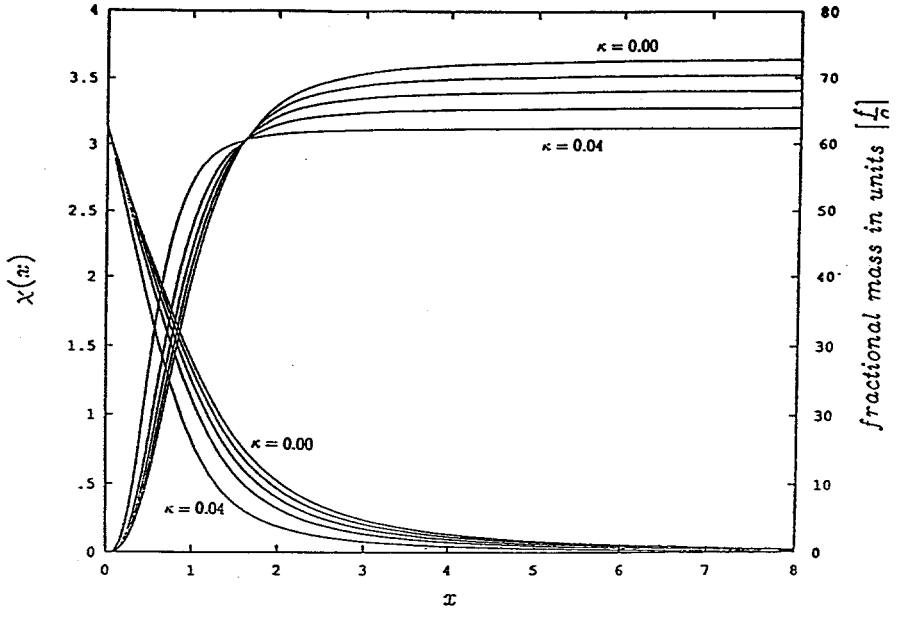

Fig. 4. Radial dependence of the matter field χ (left scale) and the metric function m (right scale) of self-gravitating Skyrmions for the following values of the coupling constant: $\kappa = 0$, 0.01, 0.02, 0.03, 0.04. χ and the asymptotic values of m decrease both monotonically with increasing κ. For values $\kappa > \kappa_0 \geq 0.0404$ there exist no particle-like solutions with winding number one.

The "shooting parameter" γ is chosen such that $\chi(\infty) = 0$. Since the topological 3-form (106) becomes

$$\omega = -\frac{1}{2\pi^2}\chi' \sin^2\chi \, dr \wedge d\vartheta \wedge d\varphi, \tag{118}$$

the winding number is then equal to one.

In Fig. 4 we show the results of the numerical integration for the radial dependence of the matter variable χ and of the mass fraction m for various coupling constants κ. (The other coupling constant g is completely absorbed in the dimensionless radial coordinate x.) Qualitatively the behavior of χ is like for the well-known Skyrmion (corresponding to $\kappa = 0$). The increasing gravitational binding energy of our topological boson stars as a function of κ can be read of from the asymptotic behavior of $m(r)$. For reasons of available space we do not show the metric variable $a(r)$; its qualitative behavior is as expected.

The curves shown in Fig. 4 cover almost the complete interval $[0, \kappa_0]$ of the coupling constant κ for which particle-like solutions with winding number one exist. Numerically we found $\kappa_0 = 0.0404$.

We have recently carried out a stability analysis of the self-gravitating Skyrmions [16]. At first sight one may think that this is superfluous, because these

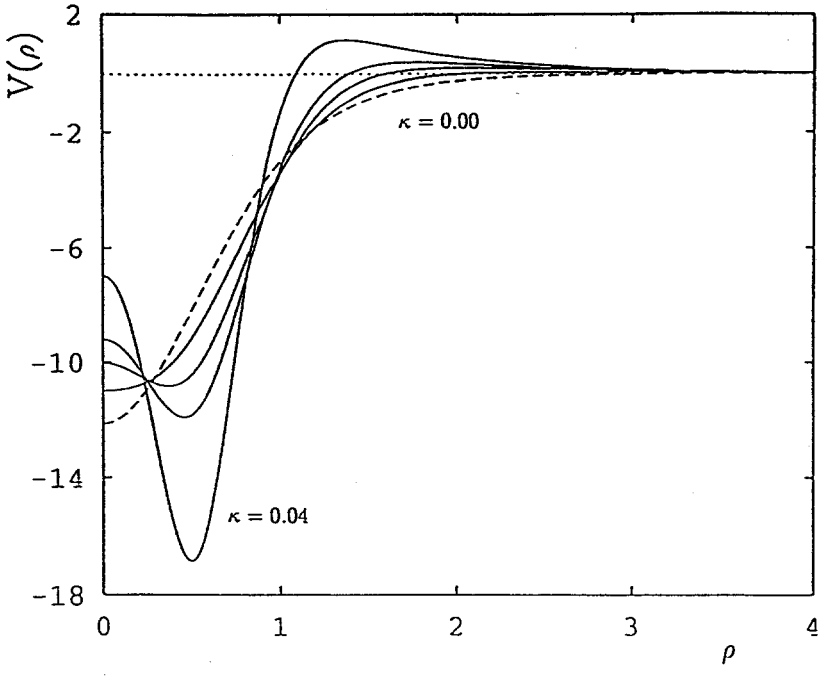

Fig. 5. The potential V as a function of the radial coordinate ρ. The dashed line shows the flat case, where V is nowhere positive. The fully drawn curves correspond to the potentials for self-gravitating solutions with $\kappa = 0.02$, 0.03, 0.035 and 0.04.

objects have an integer winding number. This implies a kind of global stability, which is, however, quite different from the more important notion of (local) Liapunov stability. Indeed, it has turned out that Liapunov stability is lost, once the coupling constant κ becomes larger than a critical value κ_c. On the other hand, the frequency spectrum of radial pulsations is real for $\kappa < \kappa_c$ and thus all radial modes are oscillatory.

The stability analysis is greatly simplified by the fact that the metric perturbation can be eliminated with the help of the gravitational field equations and that the determination of the frequence spectrum can thus be reduced to the solution of a one-dimensional eigenvalue problem. (I shall illustrate this technique in my second contribution to this meeting.) More precisely, we can reduce the determination of the frequency spectrum for radial perturbations of the Einstein-Skyrme equations to the problem of finding the energy spectrum of a p-wave Schrödinger equation with a bounded effective potential, which is determined by the equilibrium solution. Bound states of this Schhrödinger equation correspond to exponentially growing modes.

In Fig. 5 I show this effective potential as a function of a rescaled radial variable ρ, defined by

$$\frac{d\rho}{dx} = e^{(b_0-a_0)(x)-(b_0-a_0)(0)}, \; \rho(0) = 0, \tag{119}$$

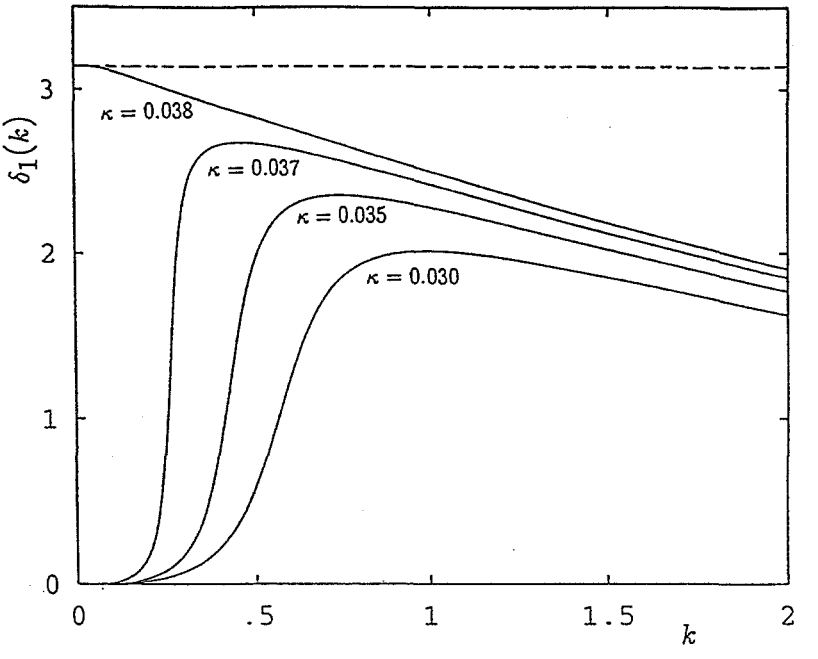

Fig. 6. The phase shift δ_1 as a function of the energy k for different coupling constants. δ_1 is evaluated at the zero of the potential $V(\rho)$, i.e. δ_1 is the asymptotic value of the phase function corresponding to the nowhere repulsive potential $V\theta(-V)$. The curves demonstrate that $V(\geq V\theta(-V))$ has no bound state for $\kappa < \bar{\kappa} \approx 0.038$.

for various values of the coupling constant κ. We have determined the possible bound states with the help the well-known Levinson theorem which states that the zero-energy value of the phase shift δ_l is π times the number of bound states with angular momentum l. As is obvious from Fig. 6, the number of bound states for our problem is zero for $\kappa < \kappa_c \approx 0.38$ and becomes to one for $\kappa > \kappa_c$. (We have shown that this is so in the whole interval $[\kappa_c, \kappa_0]$.) For the details of the stability analysis I refer to our paper [16]. (See also Ref. [30] where the nonlinear behavios is studied.)

c) Gauge boson stars

It came as a big surprise when Bartnik and McKinnon (BK) discovered a few years ago that the Einstein-Yang-Mills (EYM) system has particle-like solutions [17]. This was unexpected because there are a number of non-existence theorems[1] for related systems, for instance for the EYM equations in 2+1 dimensions [18]. The BK solutions are necessarily of a special type, because one can show that the EYM system does not admit spherically symmetric soliton so-

[1]We give some details on that in the Appendix, where static EYM fields are discussed in general.

lutions with nonvanishing YM charges [19]. In particular, there exist no regular monopoles and dyons.

I describe now some details of the BK solutions of the SU(2) EYM system. (Generalizations to other Lie groups have just been started [20].)

First, one has to parametrize spherically symmetric gauge fields. In geometric terms, one must describe $SO(3)$ invariant connections on principle bundles (over spacetime with the gauge group as the structure group) which admit an $SO(3)$ action by bundle automophisms, such that the induced action on the base manifold are the $SO(3)$ isometries of spacetime[2]. Without any loss of generality, the gauge potential can be represented in the following form [22]:

$$A = u\tau_3 dt + v\tau_3 dr + (w\tau_1 + \tilde{w}\tau_2)d\vartheta + (\cot\vartheta\tau_3 + w\tau_2 - \tilde{w}\tau_1)\sin\vartheta d\phi, \qquad (120)$$

where $\tau_j (j = 1, 2, 3)$ are the Pauli matrices and u, v, w and $\tilde{w}$ depend only on r (and t for radial pulsations). The form (120) remains invariant under a residual $U(1)$ gauge symmetry which can be used to set $v = 0$. One can also show [19, 24] that the electric part with amplitude u must vanish for a regular solution (see also the Appendix). Because the variables w and $\tilde{w}$ appear completely symmetrically in the EYM system the two amplitudes must be proportional for a particle-like solution and we can always set $\tilde{w} = 0$ (after a constant gauge transformation). In this way we arrive at the BK-ansatz:

$$A = w\tau_1 d\vartheta + (\cot\vartheta\tau_3 + w\tau_2)\sin\vartheta d\phi. \qquad (121)$$

The angular momentum part of the YM field strength $F = dA + A \wedge A$ is equal to $(w^2 - 1)\tau_3 d\vartheta \wedge \sin\vartheta d\phi$ and describes a monopole with charge distribution $w^2 - 1$. Working out the energy-momentum tensor gives

$$\rho = \frac{1}{4\pi}[B_T^2 + \frac{1}{2}B_L^2], \qquad (122)$$

$$p_r = \frac{1}{4\pi}[B_T^2 - \frac{1}{2}B_L^2], \qquad (123)$$

where

$$B_T^2 = \frac{e^{-2b}}{r^2}w'^2, \; B_L^2 = \frac{(1 - w^2)^2}{r^4}, \qquad (124)$$

The longitudiual magnetic part (B_L) acts as a repulsion.

The Schwarzschild solution is obtained for $w \equiv \pm 1$ and the Reissner-Nordstrøm solution corresponds to $w \equiv 0$.

Originally BK found their solution only numerically. In the meantime a mathematical existence proof has also been given [25]. In Fig. 7 we show the amplitude w for the node numbers $n = 1, 2, 3$. One can give a priori arguments that $|w| \leq 1$. The energy density is concentrated in a central region and decays rapidly. In an intermediate region, the YM fields are approximately

[2]The theory of these invariant connections has been described systematically by Wang; see, e.g., Ref. [21]. We have generalized the representation (121) to any compact semi-simple gauge group [23].

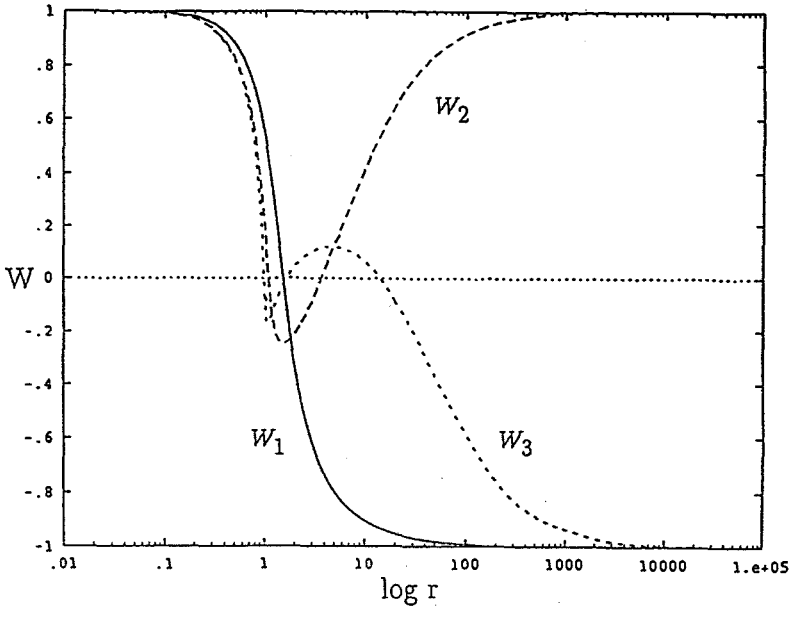

Fig. 7. Radial dependence of the matter field w (see eq. (121)) of the Bartnik and McKinnon solutions for the node number $n = 1, 2, 3$.

those of a unit charge Dirac monople and the gravitational field is close to the Reissner-Nordstrøm metric. In the far field region the solution approaches the Schwarzschild solution with discrete values of the ADM mass. Therefore, the total magnetic and electric charges vanish.

Zhi-hong Zhou and I have shown that the BK solutions are, unfortunately, unstable [26, 27, 28]. I do not go into the details of our work, since I will show in my second contribution how a similar analysis is done for the non-abelian black holes. It should, however, be mentioned that we have not only demonstrated the instability, but we have also followed numerically the non-linear evolution of small initial perturbations [27, 28]. (For another approach see Ref. [29].)

In this talk I have not discussed how boson stars might have been formed in the early universe. It is not totally inconceivable that a significant component of non-baryonic dark matter consists of boson stars.

Appendix

In this Appendix we discuss first in general terms static solutions of the EYM system and present the proofs of some non-existence theorems for particle-like solutions.

A static space-time (M, g) is of the form $M = \mathbf{R} \times N$,

$$g = -\alpha^2 dt^2 + h \tag{125}$$

where α is a non-negative function on N and (N,h) is a d-dimensional Riemannian space with (time-independent) metric h; t is the natural coordinate of $\mathbf{R}$ and ∂_t is a timelike Killing field. Here, we are not interested in black holes, and therefore assume that N is topologically an $\mathbf{R}^3$.

First, we split the source-free YM equations into space and time (d+1 splitting). (This can easily be generalized to stationary situations following the procedure in Ref.[31].) Let

$$\begin{aligned} A &= A_i dx^i + A_0 dt \\ &\equiv \mathbf{A} + \phi dt, \end{aligned} \tag{126}$$

be the decomposition of the YM gauge potential. Spatial objects are always denoted by boldface letters. For the YM field strength we have then the decomposion

$$\begin{aligned} F &= dA + A \wedge A \\ &= \mathbf{dA} + \mathbf{A} \wedge \mathbf{A} + \mathbf{d}\phi \wedge dt + [\mathbf{A}, \phi] \wedge dt \\ &= \mathbf{F} + \mathbf{D}\phi \wedge dt \\ &\equiv \mathbf{B} + \alpha \mathbf{E} \wedge dt, \end{aligned} \tag{127}$$

where

$$\mathbf{B} = \mathbf{F} = \mathbf{dA} + \mathbf{A} \wedge \mathbf{A}, \tag{128}$$

$$\alpha \mathbf{E} = \mathbf{D}\phi. \tag{129}$$

The homogeneous YM equation $DF = 0$ gives immediately

$$\begin{aligned} \mathbf{DB} &= 0, \\ \mathbf{D}(\alpha \mathbf{E}) + [\phi, \mathbf{B}] &= 0. \end{aligned} \tag{130}$$

For the Hodge-dual $*F$ we find the $d+1$ splitting

$$* F = \star \mathbf{E} - \alpha(\star \mathbf{B}) \wedge dt. \tag{131}$$

where $\star$ denotes the spatial Hodge-dual. Therefore, the YM equation $D * F = 0$ leads to the spatial equations

$$\begin{aligned} \mathbf{D} \star \mathbf{E} &= 0, \\ \mathbf{D}(\alpha \star \mathbf{B}) - [\phi, \star \mathbf{E}] &= 0. \end{aligned} \tag{132}$$

Note that in the non-abelian case the static field equations for $\mathbf{E}$ and $\mathbf{B}$ do *not decouple.*

Consider now a particlelike solution. We show first that the field $\mathbf{E}$ has to be vanish under mild fall off assumptions at infinity. Using the identity

$$Tr(\phi \mathbf{D} \star \mathbf{E}) = \mathbf{d} Tr(\phi \star \mathbf{E}) - Tr(\mathbf{D}\phi \wedge \star \mathbf{E}) \tag{133}$$

and (129) in the following consequence of (132)

$$\int_N Tr(\phi \mathbf{D} \star \mathbf{E}) = 0, \tag{134}$$

we obtain with Stokes theorem

$$\int_N \alpha Tr(\mathbf{E} \wedge \star \mathbf{E}) = 0, \tag{135}$$

and hence $\mathbf{E} = 0$. The field equations (130) and (132) then reduce to the 'magnetostatic equations':

$$\begin{aligned} \mathbf{DB} &= 0, \\ \mathbf{D}(\alpha \star \mathbf{B}) &= 0. \end{aligned} \tag{136}$$

A Einstein-Maxwell system

Let us first consider the abelian case. From $\mathbf{d}(\alpha \star \mathbf{B}) = 0$ we have globally $\alpha \star \mathbf{B} = \mathbf{d}\psi$. Hence, if we replace in the argument above E by $\star$B, we find also $\mathbf{B} = 0$. It is important to note that this reasoning can obviously not be generalized to the non-abelian theory. Therefore, the Einstein-Maxwell equations reduce to Einstein's vacuum equations, which split as follows:

$$\begin{aligned} \triangle \alpha &= 0, \\ Ric(h) &= \tfrac{1}{\alpha} Hess(\alpha). \end{aligned} \tag{137}$$

Using the maximum principle for harmonic functions and the asymptotic flatness condition

$$\alpha \to 1 \tag{138}$$

at spatial infinity, we obtain $\alpha \equiv 1$. Hence, $Ric(h) = 0$. In three dimensions this implies that the Riemann tensor vanishes and there remains only the trivial solution (Lichnerowicz [32]).

B EYM system in 2+1 dimensions

The argument in A cannot be generalized to the non-abelian case. As first remarked by Deser [18], we can, however, exclude easily non-trivial solutions in 2+1 dimensions. The crucial point is that $\omega := \alpha \star \mathbf{B}$ is a space-time scalar for $d = 2$. From $\mathbf{D}\omega = 0$, we obtain for the norm $|\omega|$ of ω (in group space) the condition

$$\mathbf{d}|\omega| = 0, \tag{139}$$

which implies for $d = 2$ that $|\omega| = const.$ This constant must vanish, otherwise there would be a $\mathbf{B}$ field which does not vanish asymptotically. The rest of the argument in A can be repeated.

C Non-existence of static YM solitons

We recall at this point the nice argument of Coleman [33] that there exist no particlelike solutions of the YM system in $d+1$ dimensions if $d \neq 4$.

From (127) we find

$$(F,F) = (\mathbf{F},\mathbf{F}) - (\mathbf{D}\phi,\mathbf{D}\phi). \tag{140}$$

The right hand side of this equation is proportional to the Yang-Mills-Higgs Lagrangian without self-couplings in d dimensions. Thus, we have

$$\begin{aligned} -\frac{1}{4}\int d^dx(F,F) &= S_{YMH}(\mathbf{A},\phi) \\ &\equiv S_1 + S_2, \end{aligned} \tag{141}$$

with

$$\begin{aligned} S_1 &= -\frac{1}{4}\int(\mathbf{F},\mathbf{F})d^dx, \\ S_2 &= \frac{1}{4}\int(\mathbf{D}\phi,\mathbf{D}\phi)d^dx. \end{aligned} \tag{142}$$

Since the energy of the soliton should be finite, both terms S_1 and S_2 have to be finite.

Suppose now that $(\mathbf{A}(\mathbf{x}),\phi(\mathbf{x}))$ is a critical point of the action S_{YMH}. Let us embed this field configuration into the two-parameter family of variations

$$\phi(\mathbf{x};\sigma,\lambda) = \sigma\lambda\phi(\lambda\mathbf{x}), \tag{143}$$

$$\mathbf{A}(\mathbf{x};\sigma,\lambda) = \lambda\mathbf{A}(\lambda\mathbf{x}). \tag{144}$$

The action has the following scaling behavior

$$S_{YMH}(\sigma,\lambda) = \sigma^2\lambda^{4-d}S_1 + \lambda^{4-d}S_2. \tag{145}$$

Since this function must be stationary for $\sigma = \lambda = 1$ we find for $d \neq 4$ that $S_1 = S_2 = 0$. Therefore, $\mathbf{F} = \mathbf{D}\phi = 0$, which implies $F = 0$ (for $d \neq 4$).

References

[1] W.E. Thirring: In *Rigorous Atomic and Molecular Physics*, ed. by G. Velo, A.S. Wightman (Plenum Press 1981).

[2] A. Martin, Phys. Lett. **B214**, 561 (1988).

[3] E. H. Lieb and H. -T. Yau, Commun. Math. Phys. **112**, 147 (1987); E. H. Lieb and H. -T. Yau, Astrophys. J. **323**, 140 (1987).

[4] A. Martin and S. M. Roy, Phys. Lett. **B233**, 407 (1989).

[5] J. L. Basdevant, A. Martin and J. M. Richard, Nucl. Phys. **B343**, 60 (1990).

[6] M. Membrano, A. F. Pacheco and J. Sanũdo, Phys. Rev. **A39**, 4207 (1989); R. Ruffini and S. Bonnazzola, Phys. Rev. **187**, 1767 (1969).

[7] I. Herbst, Commun. Math. Phys. **53**, 285 (1977); Errata ibid **55**, 316 (1977). R. Weder, J. Funct. Anal. **20**, 319 (1975). T. Kato: *Perturbation theory of linear operators.* Springer-Verlarg 1966. See remark 5.12, p.307.

[8] B. D. Serot, J. D. Walecka, Adv. Nucl. Phys. **16**, 1-321 (1985).
L. S. Celenza, C. M. Shakiu, *Relativistic Nuclear Physics: Theories of Structure and Scattering.* Singapore: World-Scientific (1985).
S. H. Kahana, Ann. Rev. Nucl. Part. Sci. **39**, 231 (1989).

[9] Ph. Jetzer, *Boson Stars*, Zürich University Preprint ZU-TH 25/91. This review paper contains an extensive list of references.

[10] R. Friedberg, T. D. Lee and Y. Pang, Phys. Rev. **D35**, 3640 (1987).

[11] M. Colpi, S. L. Shapiro and I. Wasserman, Phys. Rev. Lett. **57**, 2485 (1986).

[12] T. D. Lee and Y. Pang, Nucl. Phys. **B315**, 477 (1989).

[13] F. V. Kusmartsev, E. W. Mielke and F. E. Schunck, Phys. Rev. **D43**, 3895 (1991).
F. V. Kusmartsev, Phys. Rep. **183**, 1(1989).

[14] S. Droz, M Heusler and N. Straumann, Phys. Lett. **B268**, 371 (1991).

[15] T. H. R. Skyrme, Proc. R. Soc. **A260**, 127 (1960); J. Math. Phys. **12**, 1735 (1971);
N. Pak and H. C. Tze, Ann. Phys. **117**, 164 (1979);
G. S. Adkins, C. R. Nappi and E. Witten, Nucl. Phys. **B228**, 552 (1983).

[16] M. Heusler, S. Droz and N. Straumann, Phys. Lett. **B271**, 61 (1991).

[17] R. Bartnik and J. Mckinnon, Phys. Rev. Lett. **61**, 41 (1988).

[18] S. Deser, Class. Quantum Grav. **1**, L1 (1984).

[19] A. A. Ershov and D. V. Gal'tsov, Phys. Lett. **A150**, 159 (1990).

[20] H. P. Künzle, Alberta University preprint (1991); D. V. Gal'tsov and M. S. Volkov, Crete University preprint (1991).

[21] J. Harnad, S. Shnuider and L. Vinet, J. Math. Phys. **21**, 2719 (1980);
P. Forgács and N. S. Manton, Comm. Math. Phys. **72**, 15 (1980);
C. H. Gu and H. S. Hu, Comm. Math. Phys. **79**, 75 (1981);
S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, Vol.1, John Wiley & Sons, New York, 1963 (see section II.11).

[22] E. Witten, Phys. Rev. Lett. **38**, 121 (1977).

[23] O. Brodbeck and N. Straumann, to be published.

[24] P. Bizon and O. T. Popp, Universität Wien preprint, UWThPh-1991-20.

[25] J. A. Smoller, A. G. Wasserman, S.-T. Yau and J. B. McLeod, Commun. Math. Phys. **143**, 115 (1991).

[26] N. Straumann and Z.-H. Zhou, Phys. Lett. **B237**, 353 (1990).

[27] Z. -H. Zhou and N. Straumann, Nucl. Phys. **B360**, 180 (1991).

[28] Z. -H. Zhou, *Instability of Einstein-Yang-Mills Solitons and Non-Abelian Black Holes*, Thesis, University of Zürich, 1991.

[29] M. Heusler and N. Straumann, MPA preprint, MPA634; Zürich University preprint ZU-TH29/91.

[30] Z. -H. Zhou and N. Straumann, in preparation.

[31] R. Durrer and N. Straumann, Helv. Phys. Acta **61**, 1027 (1988).

[32] A. Lichnerowicz, in *Théories relativiste de la gravitation et de l'éléctromagnetisme* (Masson, Paris, 1975).

[33] S. Coleman, in *New phenomenon in subnuclear physics*, ed. A. Zichichi (Plenum, New York, 1975).

Black Holes with Hair

Norbert Straumann
Institute for Theoretical Physics, University of Zurich

1 Introduction

Until recently it was generally believed that the well-known uniqueness theorems for stationary black holes, established rigorously for the Einstein-Maxwell (EM) system[1], should have natural generalizations to other matter models, like Yang-Mills (YM) theories. It came, therefore, as a big surprise when the "colored" black hole solutions for the Einstein-Yang-Mills (EYM) system were discovered [2, 3, 4], because these non-abelian black holes represent counter examples to the "no hair conjecture", which states for this case that the structure of such black holes should be determined uniquely by the hole's mass, intrinsic angular momentum and the global YM charges defined at spatial infinity.

The colored black hole solutions are static , spherically symmetric and have vanishing YM charges. Asymptotically they approach the Schwarzschild solution, but they have "Yang-Mills hair". Near the horizon they are close to the Reissner-Nordstrøm solution with an effective charge that slowly decays in a transition zone. So far, their existence has only been established numerically. It should, however, be possible to extend the rigorous existence proof of the Bartnik-McKinnon ground state solution of the EYM equation [5] to the black hole case.

Zhi-hong Zhou and I have shown that the colored black holes are unstable [6, 7, 8]. Since there has been some controversy on this during the last year [9, 10, 11, 12] which is now settled, I shall discuss this point in some detail.

Because of their instability the colored black holes are not so interesting physically. This is the main reason why we have searched for black holes of other non-linear matter models which would provide counter example to the no hair conjecture and which might in addition be stable. In view of the structural similarities of the nonlinear sigma models and Yang-Mills theories, we looked for black hole solutions of the coupled Einstein-Skyrme (ES) system.

Recently we discovered numerically that the ES system has – for a certain range of coupling constants – indeed static black hole solutions with a regular event horizon and which approach asymptotically the Schwarzschild solution [13]. Outside the horizon the new solutions behave like self-gravitating Skyrmions, which we have also constructed numerically [13] (see also my first contribution to this meeting). A distant observer can, however, not distinguish these solutions with "Skyrme hair" from the Schwarzschild black hole. (The energy density of the matter fields decays rapidly.) Therefore, the new "Skyrme black holes"

provide counter examples to the no hair conjecture. Not everything which can be radiated away in the formation of a black hole will always be radiated away.

The question of (linear) stability has also been investigated by extending our stability analysis of the Skyrmions [14]. We found that there are only oscillatory radial modes, satisfying the necessary boundary conditions at the horizon and at infinity [15]. Therefore, the ES black holes are linearly stable. The nonlinear stability can, in practice, only be decided by detailed numerical studies. To our knowledge, this has not even been done for the Schwarzschild black hole.

2 Colored black holes

The basic formulae have already been discussed in my first lecture [16]. We chose the gauge potential again in the special form with only one independent amplitude

$$A = w(r)\tau_1 d\vartheta + (cot\vartheta\tau_3 + w(r)\tau_2)\sin\vartheta d\varphi, \tag{1}$$

because one can prove that there exist only (embedded) abelian solutions of the Reissner-Nordstrøm type if there is also an electric component [11, 18]. The matter variable $w(r)$ and the metric variables $a(r)$, $b(r)$ in

$$g = e^{2a}dt^2 - [e^{2b}dr^2 + r^2(d\vartheta^2 + \sin^2\vartheta d\varphi^2)] \tag{2}$$

must satisfy the following differential equations

$$m' = e^{-2b}w'^2 + \frac{(1-w^2)^2}{2r^2}, \tag{3}$$

$$a' = \frac{e^{2b}}{r}(e^{-2b}w'^2 - \frac{(1-w^2)^2}{2r^2} + \frac{m}{r}), \tag{4}$$

$$w'' = \frac{e^{2b}}{r^2}(\frac{(1-w^2)^2}{r} - 2m)w' - \frac{e^{2b}}{r^2}(1-w^2)w, \tag{5}$$

where we have also introduced the usual mass fraction variable $m(r)$:

$$e^{-2b(r)} = 1 - 2m(r)/r. \tag{6}$$

For black hole solutions the existence of a regular event horizon at $r = r_H$ requires that

$$2m(r_H) = r_H \quad \text{and} \quad a(r_H) + b(r_H) < \infty. \tag{7}$$

(Note that the determinant of the metric is proportional to e^{a+b}.) Outside the horizon the condition $r > 2m(r)$ must, of course, be fulfilled and in addition we have to impose the asymptotic flatness condition $a(r), b(r) \to 0, m(r) \to M < \infty$ as $r \to \infty$. (Actually we have to require for $a(r)$ only that it converges to a finite constant which we can choose to vanish by an appropriate choice of the time coordinate.)

Fig.1 shows the radial dependence for several quantities of the ground state solution ($n = 1$). The horizon is taken at $r_H = 1$. (There are black hole solutions for any value of r_H. This reflects a scaling property of the EYM system, as is

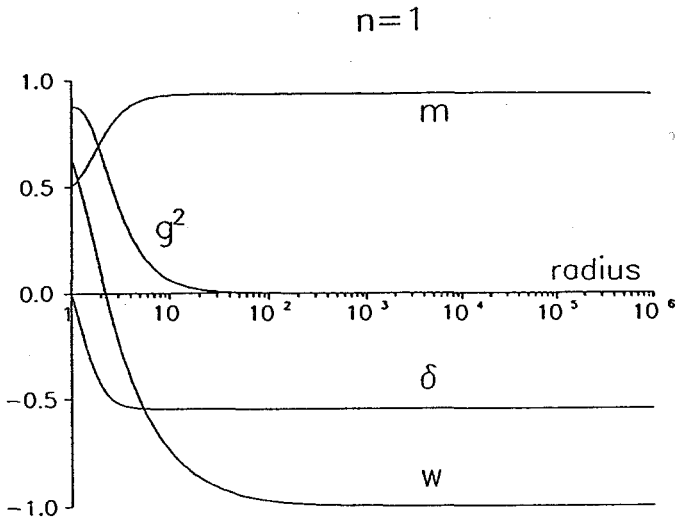

Fig. 1. Yang-Mills amplitude w, mass function and $\delta = -(a+b)$ for the lowest colored black hole solution with event horizon $r_H = 1$. (From [3].)

shown in Ref. [8].) The effective Reissner-Nordstrøm charge $g(r)$ is defined in the obvious way

$$g^2(r) := 2r(M - m(r)). \tag{8}$$

One sees that this effective magnetic charge slowly decays in a transition zone outside the horizon. It has to be emphasized that all YM charges vanish and that the solutions behave asymptotically like the Schwarzschild black hole. For a given r_H the gravitational mass M has a fixed value for each node number n. This family of non-abelian black hole solutions provides obviously counter examples to the originally expected uniqueness theorem for the EYM system. Many people had guessed that a generalized Isreal theorem should hold, namely that a static black hole solution with vanishing YM charges would have to be the Schwarzschild solution. More generally, it was expected that the only stationary black hole solutions would be essentially abelian (embedded Kerr-Newman family) [19].

3 Instability of colored black holes

It suffices to consider spherically symmetric perturbations, because we shall find an instability in this restricted class. Furthermore, we keep A in the form (1), but with an amplitude $w(r,t)$ which depends also on t. The same is, of course, true for the metric variables a and b in (2).

Decomposing a, b and w in the vicinity of the black hole solution a_0, b_0 and w_0

$$\begin{aligned} a(r,t) &= a_0(r) + a_1(r,t), \\ b(r,t) &= b_0(r) + b_1(r,t), \\ w(r,t) &= w_0(r) + w_1(r,t), \end{aligned} \tag{9}$$

and treating the perturbation as small quantities in the time-dependent field equations gives coupled linear equations for the deviations a_1, b_1 and w_1 (for details see [6, 7, 8, 21]). I summarize now the main points of our analysis of these equations.

The tr-component of Einsteins field equations leads to ($G = c = 1$)

$$\dot{b_1} = \frac{2}{r} w_0' \dot{w_1} \tag{10}$$

(with $\dot{\cdot} \equiv \partial_t$, $' \equiv \partial_r$). Considering also the tt-component leads to the conclusion, that b_1 must be the following special solution of (9)

$$b_1 = \frac{2}{r} w_0' w_1. \tag{11}$$

Beside b_1 we shall need also the combination $(a_1 - b_1)'$, which is already obtained from the sum of the linearized tt- and rr-equations, and (11):

$$b_1' - a_1' = 2(b_0' - a_0' - \frac{1}{r}) b_1 - \frac{4}{r^3} e^{2b_0} w_0 (1 - w_0^2) w_1. \tag{12}$$

In the derivation of (12) one must also repeatedly use the equilibrium equations (3)–(5).

Next, we linearize the YM equation $D{*}F = 0$ and find

$$\begin{aligned} e^{2(b_0 - a_0)} \ddot{w_1} \; &- \; w_1'' + (b_0' - a_0') w_1' + (b_1' - a_1') w_0' \\ &- \; \frac{1}{r^2} e^{2b_0} w_0 (1 - w_0^2) b_1 - \frac{1}{r^2} e^{2b_0} (1 - 3w_0^2) w_1 = 0. \end{aligned} \tag{13}$$

It is a very remarkable fact that the only combinations of the metric perturbations appearing in (13), i.e. $b_1' - a_1'$ and b_1 are already determined by the gravitational equations, without any further integration. This simplifies the analysis very much. Substituting (11) and (12) into (13), we obtain for the amplitude ξ in

$$w_1(r,t) = \xi(r) e^{i\sigma t} \tag{14}$$

the following eigenvalue equation

$$-\xi'' + \frac{1}{2} \alpha' \xi' + U\xi = \sigma^2 e^{\alpha} \xi. \tag{15}$$

Here, we have introduced the abbreviation $\alpha = 2(b_0 - a_0)$ and the effective potential

$$U = \frac{4}{r} (w_0')^2 (\frac{\alpha'}{2} - \frac{1}{r}) - \frac{8}{r^3} e^{2b_0} w_0 w_0' (1 - w_0^2) - \frac{1}{r^2} e^{2b_0} (1 - 3w_0^2). \tag{16}$$

For further investigations it is useful to bring (15) into the form of a one-dimensional Schrödinger equation. We introduce the new radial coordinate ρ by

$$\frac{d\rho}{dr} = e^{\alpha/2}, \quad \rho(r_H) = -\infty \tag{17}$$

and find

$$(-\frac{d^2}{d\rho^2} + U_{eff})\xi = \sigma^2\xi, \tag{18}$$

with

$$U_{eff} = e^{-\alpha}U. \tag{19}$$

One can show that $U_{eff} \to 0$ for $\rho \to \pm\infty$. It is clear that bound states of (18) are potentially unstable modes. Because of the presence of a horizon, we have to make sure that the corresponding perturbations of various quantities are all well behaved at the horizon. In order to do this, we introduce Kruskal-like coordinates u, v by

$$u = e^{\eta\rho}\cosh(\eta t), \quad v = e^{\eta\rho}\sinh(\eta t). \tag{20}$$

Here, η has to be chosen such that f^2 in the transformed metric,

$$g = f^2(dv^2 - du^2) - r^2(u,v)d\Omega^2, \tag{21}$$

does not vanish at the horizon. This leads to

$$\begin{aligned} \eta &= \frac{1}{2}\frac{d}{dr}e^{-\alpha/2}|_H \\ &= \frac{1}{2r_H}[1 - \frac{(1-w^2(r_H))^2}{r_H^2}]e^{a(r_H)+b(r_H)}. \end{aligned} \tag{22}$$

In arriving at the last expression we have used the equilibrium equations. For the ground state colored black hole solution one finds

$$\eta = 0.184 \quad (n = 1,\ r_H = 1). \tag{23}$$

The potential U (as a function of the original variable r) is shown in Fig.2. For this potential we found exactly one bound state with

$$\sigma = -0.0269. \tag{24}$$

The corresponding amplitude ξ is shown in Fig. 3.

The question now arises, whether this unstable mode is physically acceptable. If we transform to Kruskal coordinates it is obvious that the gauge potential (eq.(1) with the time dependent $w(r,t)$) is well-behaved at the horizon. The same can be shown for the perturbation of the metric (see [8]). If we look, however, at the field strength F its behavior seems, at first sight, to be untolerable [9]. We find easily

$$F = \dot{w}dt \wedge \Omega + w'dr \wedge \Omega - (1-w^2)\tau_3 d\vartheta \wedge \sin\vartheta d\varphi, \tag{25}$$

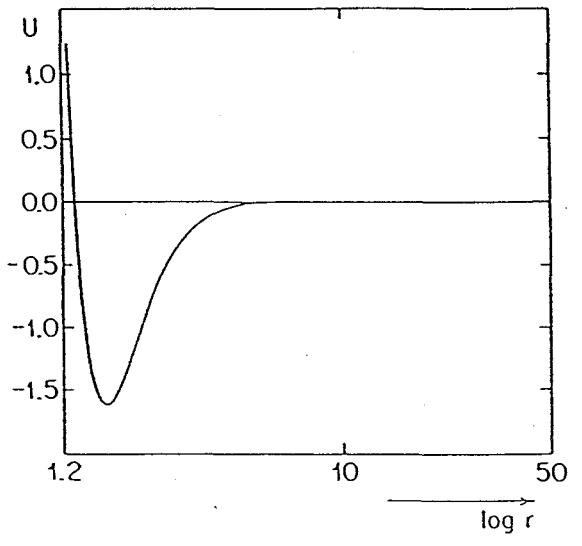

Fig. 2. Effective potential $U(r)$ in the eigenvalue equation (15).

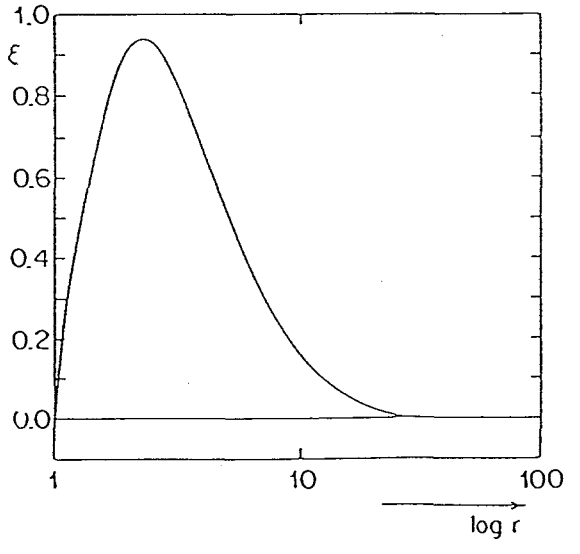

Fig. 3. Amplitude of the exponentially growing radial perturbation of the lowest colored black hole solution.

where

$$\Omega = \tau_1 d\vartheta + + \tau_2 \sin\vartheta d\varphi. \tag{26}$$

Inserting (14) gives for the perturbation of F at $t = 0$

$$\delta F|_{t=0} = i\sigma\xi dt \wedge \Omega + \xi' dr \wedge \Omega + w_0 \xi \tau_3 d\vartheta \wedge \sin\vartheta d\varphi. \tag{27}$$

Here the first two terms are dangerous because of the differentials dt and dr, which are related to du and dv for $t = 0$ by

$$dv = \eta e^{\eta\rho} dt, \quad du = \eta e^{\eta\rho} d\rho \quad (t = 0), \tag{28}$$

which shows that

$$\delta F|_{t=0} = \frac{1}{\eta} e^{-\eta\rho} (w\xi dv \wedge \Omega + \frac{d\xi}{d\rho} du \wedge \Omega) + w_0 \xi \tau_3 d\vartheta \wedge \sin\vartheta d\varphi. \tag{29}$$

Since ξ behaves near the horizon as $e^{\omega\rho}$, $\omega^2 = -\sigma^2$, we have in (29) a competition of two exponentials. From (23) and (24) (i.e. $\omega = 0.164$) we see that $\delta F|_{t=0}$ grows exponentially as we approach the horizon. Therefore, one might think that our exponentially growing mode should be excluded and that the colored black holes would be linearly stable [9]. This is, however, not correct, because one can choose a superposition of our unstable mode with stable modes in such a way, that the initial perturbation of F is regular. In the long run, the unstable mode wins, of course, and the colored black hole is unstable, as we have always claimed [10, 7, 12].

4 Black holes with Skyrme hair

I discuss now new black hole solutions with hair which we have recently found numerically for the non-linear sigma-model of Skyrme coupled to gravity [13]. These are also static and approach asymptotically the Schwarzschild solution.

First, I recall the Lagrangian of the Skyrme-model. This is an algebraic expression in the quantity

$$A_\mu = U^\dagger \nabla_\mu U, \tag{30}$$

where the basic field variable $U(x)$ is an $SU(2)$-valued function on space-time. The Lagrangian has two independent coupling constants f, g:

$$\mathcal{L}_S = \frac{f^2}{4} Tr(A_\mu A^\mu) + \frac{1}{32g^2} Tr(F_{\mu\nu} F^{\mu\nu}). \tag{31}$$

Here the 2-form F is just the commutator

$$F_{\mu\nu} = [A_\mu, A_\nu]. \tag{32}$$

If we insert in (31) the hedgehog ansatz

$$U(x) = \cos\chi(r) + i \sin\chi(r) \tau \cdot \mathbf{x}, \tag{33}$$

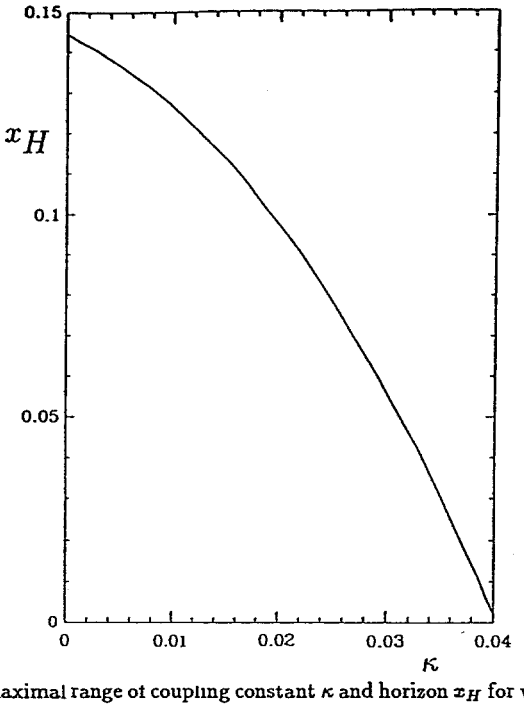

Fig. 4. The maximal range of coupling constant κ and horizon x_H for which black hole solutions exist lies below the curve of the figure.

we find for the Skyrme action

$$\int \mathcal{L}_S \sqrt{-g} d^4x = 4\pi \int_0^\infty L_S(r) dr, \tag{34}$$

with

$$\begin{aligned} L_S &= \frac{1}{2} f^2 [r^2 (\chi')^2 e^{a-b} + 2\sin^2\chi e^{a+b}] \\ &\quad + \frac{1}{2g^2} \sin^2\chi [2(\chi')^2 e^{a-b} + \frac{\sin^2\chi}{r^2} e^{a+b}]. \end{aligned} \tag{35}$$

To this we must add the gravitational Lagrangian

$$L_G = \frac{1}{2G} e^a [e^b - 2(1 + ra') + e^{-b}(1 + 2ra')]. \tag{36}$$

In [13] we have constructed black hole solutions for the coupled system described by (35) and (36). In doing this one has to derive first analytic expressions for the derivatives of the dynamical variables a, b, χ at the horizon in terms of the shooting parameter $\chi(r_H)$, by making use of the field equations. This shooting parameter is then chosen such that the boundary condition $\chi(\infty) = 0$ is fulfilled.

Our numerical studies of this non-linear boundary value problem revealed that for $\kappa := 4\pi(f/M_{Pl})^2$ in the interval $[0, \kappa_0]$, $\kappa_0 = 0.0404$, there exist always black hole solutions. For a given κ the position of the horizon can be chosen in a κ-dependent interval which is shown in Fig.4.

At $\kappa = \kappa_0$ this interval shrinks to zero and beyond κ_0 ther are no black hole solutions. It is a very curious fact that – within the numerical accuracy – the

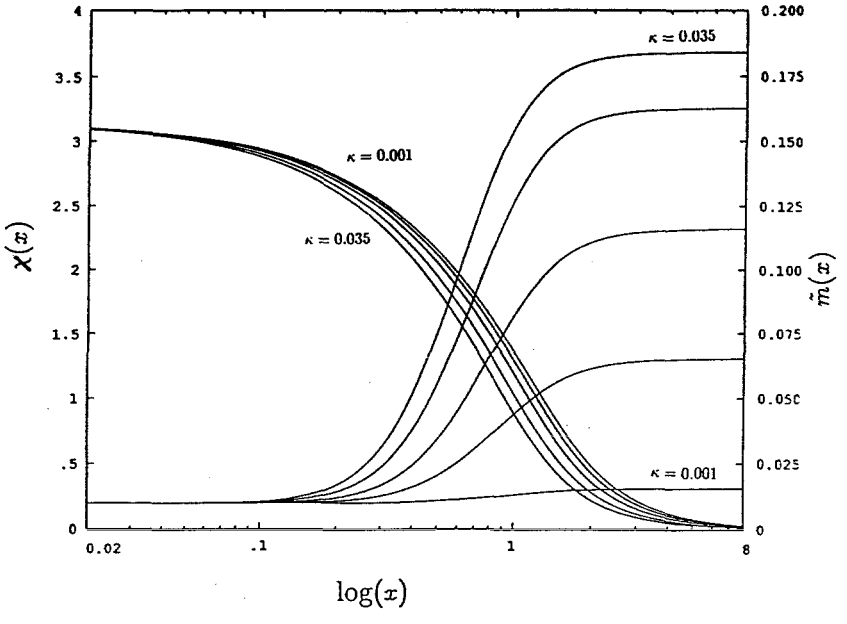

Fig. 5. Radial profiles of the matter field χ (left scale) and the mass function $\bar{m}$ (right scale) for Skyrme black holes with a regular event horizon at $x_H = 0.2$ for several values of the coupling constant κ. (Note that the horizon in physical units increases with decreasing $\sqrt{\kappa}$.)

interval $[0, \kappa_0]$ is the same as for the particle-like solutions discussed in my first lecture. Is there a deeper reason for this coincidence?

The structure of these new black hole solutions is illustrated in Figs.5,6. In both these figures we have chosen the position of the event horizon at $r_H = 0.02/fg$ and the radial profiles show the results for several values of the coupling constant κ. Outside the horizon the matter field behaves like the self-gravitating Skyrmions. Since the matter density falls off like r^{-4}, the mass fraction $m(r)$ rapidly approaches the total gravitational mass, which decreases with increasing κ. In Fig.6 we show the radial dependence of the lapse function $e^{2\alpha}$ and compare it with the Schwarzschild solution, for the same horizon. It is obvious that our sigma-model black holes are clearly separated from the Schwarzschild solutions, especially for relatively large κ. Asymptotically all our solutions approach, however, the Schwarzschild metric.

5 Linear stability of the Skyrme black holes

The linear stability of the black hole solutions of the ES system proceeds along similar lines as for the colored black holes. The frequency spectrum for radial oscillations is again determined by a one-dimensional Schrödinger equation on the whole real line. The effective potential, which is determined by the black hole solutions, is everywhere bounded and vanishes asymptotically. Its complicated form does not allow us to make use of general theorems concerning the number of bound states. We have, however, shown numerically [16] that there are no

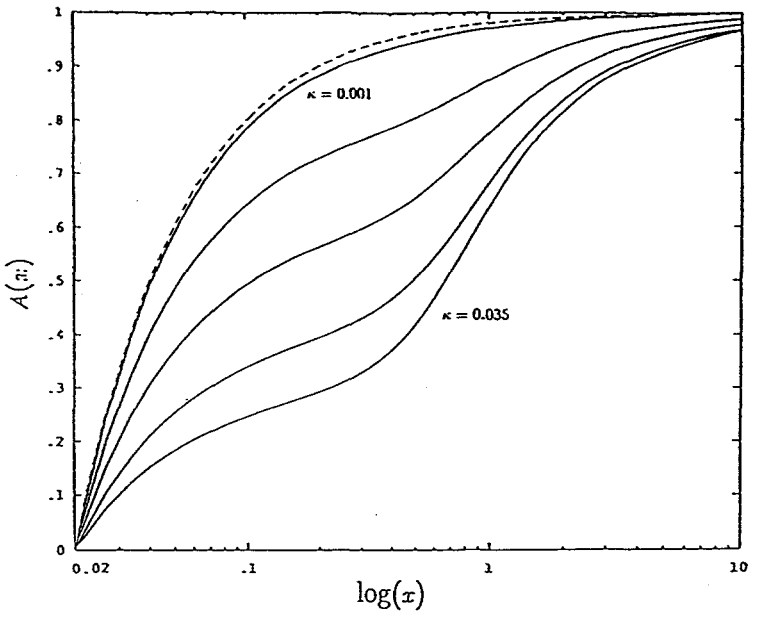

Fig. 6. Radial dependence of the lapse function $e^{2\alpha}$ for the same solutions as in Fig.5. For comparison we show also the lapse function of the Schwarzschild metric with the same horizon in the variable x.

bound states, i.e. unstable modes, for all pairs $(\kappa,\ x_H)$, except perhaps for the extreme cases with large κ and almost vanishing x_H. In this limiting region the effective potential varies rather strongly and numerical errors begin to develop. For details I refer to our recent paper [15].

The question of non-linear stability is outside of our abilities. In collaboration with Z.-H. Zhou we have, however, made some numerical investigations similar to the ones in Refs. [7, 8]. Strictly speaking, these apply so far only to the soliton solutions, for which we have solved numerically the full set of non-linear partial differential equations describing radial perturbations for a representative class of initial perturbations. For the time span which can be handled by our program no instabilities were found [8], in sharp contrast to our findings for the EYM system [7, 8]. Since some of the initial perturbations can also be considered as perturbations of the black hole solutions, we are quite confident that not only the self-gravitating Skyrmions, but also the black holes with "Skyrme hair" are stable even on the nonlinear level. For this reason these represent much more interesting counter examples to the no hair conjecture than the colored black holes.

For some further insight into the existence of such solutions, as well as for generalizations of no hair theorems to non-linear matter models, I refer also to a recent paper with M. Heusler [21], in which we have generalized scaling arguments to self-gravitating systems.

References

[1] W. Israel, Comm. Math. Phys. 8, 245 (1968);
D. Robinson, Gen. Rel. Grav. 4, 53 (1973);
G. Bartnik and A.K.M. Masoud-ul-Alam, Gen. Rel. Grav. 19, 147 (1987);
D. Robinson, Phys. Rev. Lett. 34, 905 (1975);
P. Mazur, J. Phys. A15, 3173 (1982);
B. Carter, in *Gravitation in Astrophysics*, Plenum Press, 63 (1987).

[2] H.P. Künzle and A.K.M. Masoud-ul-Alam, J. Math. Phys. 31, 928 (1990).

[3] P. Bizon, Phys. Rev. Lett. 64, 2844 (1990).

[4] M.S. Volkov and D.V. Galt'sov, Sov. J. Nucl. Phys. 51, 1171 (1990).

[5] J. A. Smoller, A. G. Wasserman, S. -T. Yau and J. B. McLeod, Commun. Math. Phys. 143, 115 (1991).

[6] N. Straumann and Z. -H. Zhou, Phys. Lett. B243, 33 (1990).

[7] Z.-H. Zhou and N. Straumann, Nucl. Phys. B360, 180 (1991).

[8] Z. -H. Zhou, *Instability of Einstein-Yang-Mills Solitons and non-Abelian Black Holes*, Thesis, University of Zürich, 1991.

[9] P. Bizon, Phys. Lett. B259, 53 (1991).

[10] P. Bizon and R. M. Wald, Phys. Lett B267, 173 (1991).

[11] D. V. Galt'sov and M. S. Volkov, to appear in Phys. Lett. A.

[12] R. M. Wald, Univ. of Chicago preprint, Chicago, 1991.

[13] S. Droz, M. Heusler and N. Straumann, Phys. Lett. 268, 371 (1991).

[14] M. Heusler,S. Droz and N. Straumann, Phys. Lett. B271, 61 (1991).

[15] M. Heusler, S. Droz and N. Straumann, Zürich University preprint, ZU-TH 92/1.

[16] N. Straumann: *Fermion and Boson Stars*, this volume.

[17] D. V. Galt'sov and A. A. Ershov, Phys. Lett. A138, 160 (1989).

[18] P. Bizon and O. T. Popp, Universität Wien preprint, UWThPh-1991-20.

[19] P. B. Yasskin, Phys. Rev. D12, 2212 (1975).

[20] N. Straumann and Z.-H. Zhou, Phys. Lett. B237, 353 (1990).

[21] M. Heusler and N. Straumann, MPI preprint, MPA 634 (1991); Zürich University preprint ZU-TH29/91.

[22] Z. -H. Zhou and N. Straumann, in preparation.

Gravitational Fields of Rapidly Rotating Neutron Stars: Theoretical Foundation

Gernot Neugebauer [1], Heinz Herold [2]

[1]MPG–Arbeitsgruppe „Gravitationstheorie" an der
Friedrich-Schiller-Universität Jena, O-6900 Jena, Germany
[2]Lehr- und Forschungsbereich Theoretische Astrophysik,
Universität Tübingen, W-7400 Tübingen, Germany

Abstract: The gravitational fields of rotating bodies describe 'minimal' surfaces in a four-dimensional (Pseudo-)Riemannian Potential Space with the line element

$$dS^2 = -2d\alpha\, dW + 2W\, dU^2 - \frac{e^{4U}}{2W}\, dA^2 - 2\kappa_0 W e^{2\alpha - 2U} p(V)\, dW^2 \ .$$

This paper *I* is meant to prepare the numerical application of the minimal surface formalism to rapidly rotating neutron stars in a following paper *II*.

1 Introduction

In order to solve the problem of the rotating body in General Relativity more effort must be made to analyze the field equations for the interior of the body. Particularly, for numerical calculations of rotating star models a simplifying reformulation of the straightforwardly specialized field equations would be desirable.

In the first part of the present bipartite contribution we undertake such an attempt for axisymmetric fluid bodies uniformly rotating around their axis of symmetry, which assumption implies a twofold space-time symmetry (axisymmetry and stationarity). We develop a rigorously geometrical approach considering the gravitational fields as 'minimal' surfaces in a (Pseudo-)Riemannian space (see (41)). In this language their regularity conditions on the axis of symmetry, the reflection symmetry with respect to the equatorial plane and the behavior of the gravitational field at infinity define the boundary of the wanted 'minimal' surface. The relationship to strings in (Pseudo-)Riemannian spaces is obvious. The roots of the procedure were the

stationary Einstein equations (Neugebauer and Kramer 1969) and a first goal was to extend the powerful generation techniques for vacuum fields as, e.g., Bäcklund transformations (Maison 1978, Harrison 1978, Neugebauer 1979, Hoenselaers et al. 1979) or the inverse (scattering) method (Belinski and Zakharov 1978, Hauser and Ernst 1979) to the gravitational fields inside the body. Unfortunately, the search did not prove successful till now. That is due to the break of the exterior SU(1,1) symmetry by a matter term in the metric of the Potential space (cf. 41).

Nevertheless, the existence of a variational principle (Plateau problem) for both the exterior and interior region is a good base for numerical calculations of the gravitational fields of relativistic star models. In this connection we focus our attention on rapidly rotating neutron stars. The activities in this field will be represented in a second contribution (cited as paper *II*).

The present contribution (paper *I*) is organized as follows: The fundamental equations and symmetries are compiled in Section 2. A variational principle yielding the field equations and the relations among the exterior parameters (surface red shift, angular velocity) presented in Section 3 is reformulated in a minimal surface language in Section 4. In the following sections, parametrizations of the minimal surfaces and boundary conditions are discussed.

2 Field Equations and Symmetry

In the subsequent section we turn to the mathematical description of rotating stars in General Relativity. We model the star matter by an one-component perfect fluid. The variables describing its state may be chosen to be the energy density $\varepsilon(x)$, the pressure $p(x)$, the four-velocity $u_i(x)$, and the gravitational field $g_{ik}(x)$ $(i,k = 1,2,3,4)$. These fields are solutions to the Einstein equations

$$R_{ik} - \frac{1}{2} R\, g_{ik} = -\kappa_0 \left((\varepsilon + p)\, u_i u_k + p g_{ik} \right) , \tag{1}$$

κ_0 being Einstein's gravitational constant. Here the four velocity u_i has the norm -1 (we take $c = 1$):

$$u_i u^i = -1 . \tag{2}$$

We may assume that the matter distribution and the gravitational field of rotating stars are stationary *and* axially symmetric, i.e. the metric admits a 2-dimensional Abelian group of motions G_2,

$$\begin{aligned} \xi_{i;k} + \xi_{k;i} = 0 , \qquad & \eta_{i;k} + \eta_{k;i} = 0 \\ \xi^i{}_{,k}\eta^k - \eta^i{}_{,k}\xi^k = 0 , \qquad & \xi_i \xi^i < 0 , \quad \eta^i \eta_i > 0 \end{aligned} \tag{3}$$

with Killing vector fields ξ^i and η^i. (A comma denotes the partial derivative and a semicolon the covariant derivative.)

From these equations we conclude that we can choose a coordinate system so that the Killing vectors take the form $\xi^i = \delta^i_4$ and $\eta^i = \delta^i_3$ and the metric tensor $g_{ik}(x)$ is independent of the timelike coordinate $x^4 = t$ and the spacelike (azimuthal) coordinate $x^3 = \varphi$.

The space-like Killing vector η^i generating axial symmetry has closed (compact) trajectories and vanishes on the rotation axis.

To describe a rotational motion, the four velocity must be a linear combination of the group generators ξ^i and η^i,

$$u^i = e^{-V}(\xi^i + \Omega\eta^i)\,, \tag{4a}$$

where, in general, the coefficients V and Ω depend on the coordinates. Obviously,

$$e^{2V} = -\left(\xi_i + \Omega\eta_i\right)\left(\xi^i + \Omega\eta^i\right)\,. \tag{4b}$$

The conditions (3) and (4a) together with the field equations (1) imply (Kundt and Trümper 1966)

$$\varepsilon_{iklm}\eta^i\xi^k\xi^{l;m} = 0 = \varepsilon_{iklm}\xi^i\eta^k\eta^{l;m}\,, \tag{5}$$

i.e. the space-time of rotating fluid bodies admits 2-spaces orthogonal to the 2-dimensional group orbits formed by the Killing trajectories (ε_{iklm} is the Levi-Civita tensor). Then, in the adapted coordinate system, the space-time line element can be written in the form (Lewis 1932, Papapetrou 1966)

$$\begin{aligned} &ds^2 = g_{AB}dx^A dx^B + g_{\mu\nu}dx^\mu dx^\nu\,, \\ &\xi^i = \delta^i_4\,, \quad \eta^i = \delta^i_3\,, \end{aligned} \tag{6}$$

where the non-vanishing components of the metric tensor g_{AB} $(A, B = 1, 2)$ and $g_{\mu\nu}$ $(\mu, \nu = 3, 4)$ depend only on the coordinates (x^1, x^2) which label the points on the 2-spaces orthogonal to the Killing trajectories.

Stationarity and *Axisymmetry* of the star matter distribution means invariance of the interior variables $\varepsilon(x), p(x)$ and $u^i(x)$ under the symmetry group

$$\varepsilon_{,i}\xi^i = 0 = \varepsilon_{,i}\eta^i\,, \quad p_{,i}\xi^i = 0 = p_{,i}\eta^i \tag{7}$$

$$u^i{}_{,k}\xi^k - \xi^i{}_{,k}u^k = 0 = u^i{}_{,k}\eta^k - \eta^i{}_{,k}u^k \tag{8}$$

From (3), (4a) and (8) one obtains

$$V_{,i}\xi^i = 0 = V_{,i}\eta^i\,, \quad \Omega_{,i}\xi^i = 0 = \Omega_{,i}\eta^i \tag{9}$$

which means that V, Ω in the adapted coordinates (6) depend on x^1 and x^2 alone. The integrability conditions of the Einstein equations (1), i.e. the Euler equations, reduce to

$$-p_{,i} = (\varepsilon + p)\left(V_{,i} + e^{-V}\eta_k u^k \Omega_{,i}\right) \tag{10}$$

or

$$dp = -(\varepsilon + p)(dV + e^{-V}\eta_k u^k \, d\Omega) \, . \tag{11}$$

Hence, ε and p must be functions of V and Ω. For *rigidly* rotating bodies, the angular velocity Ω is a constant,

$$\Omega = \text{constant} \, . \tag{12}$$

In this case, ε must be a function of p,

$$\varepsilon = \varepsilon(p) \, ,$$

and this equation of state determines the function $p = p(V)$ (and $\varepsilon = \varepsilon(V)$) via the differential equation

$$\varepsilon(p) + p = -\frac{dp}{dV} \, . \tag{13}$$

The equation of state must be chosen such that the resulting pressure function has a zero V_0,

$$p(V_0) = 0 \, . \tag{14}$$

Then

$$V = V_0 \tag{15}$$

may describe the star surface at which the pressure coincides with the vanishing pressure of the vacuum region. V_0 determines the relative red shift

$$z = e^{-V_0} - 1 \tag{16}$$

of photons from the surface received by an observer at infinity.

As an illustration consider an incompressible material,

$$\varepsilon = \varepsilon_0 = \text{constant} \, .$$

Here we get from (13)

$$p(V) = \begin{cases} \varepsilon_0(e^{V_0 - V} - 1) & \text{inside the star} \\ 0 & \text{outside the star} \end{cases} \tag{17}$$

This pressure function is schematically represented on Figure 1. Other materials could behave as sketched in Figure 2.

It should be emphasized that p is always continuous at $V = V_0$ but need not be differentiable there. In our mathematical analysis we consider p as a well-defined function of V from which $\varepsilon(V)$ can be derived via (13).

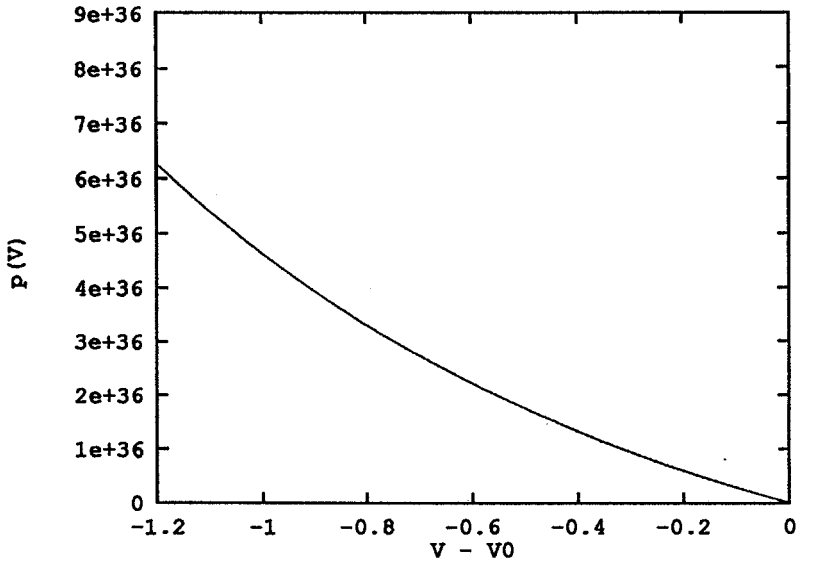

Fig. 1. Pressure function (17) for typical values of neutron stars (nevertheless, the function is not realistic)

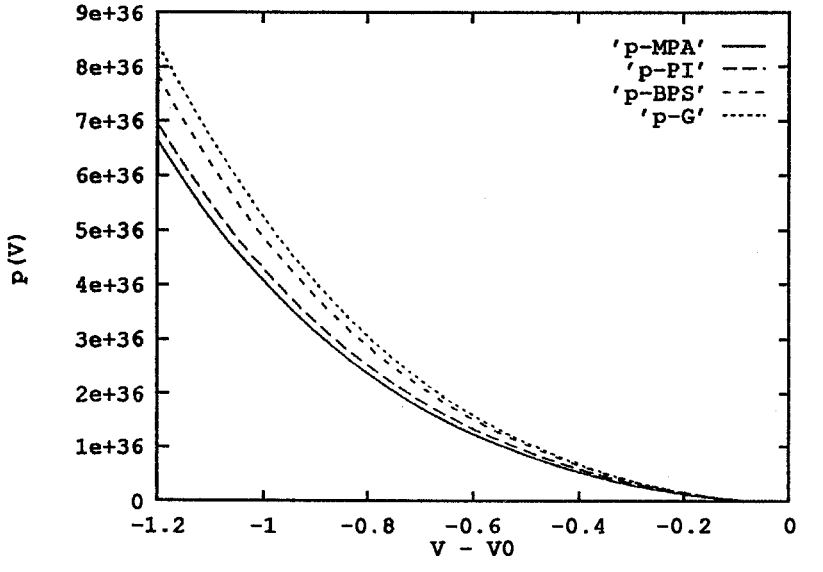

Fig. 2. Pressure functions for several "realistic" neutron star models. The denotation is explained in the context of Fig. 3 in paper *II* (Herold and Neugebauer 1992)

3 A Variational Principle

In the adapted coordinate system (6) $V(x)$ can be expressed in terms of the metric coefficients $g_{\mu\nu}$ $(\mu,\nu = 3,4)$. This follows from (4b). As a consequence, the state variables $p(x), \varepsilon(x)$, and $u_i(x)$ become functionals of the metric tensor, so that the Einstein equations (1) reduce to a nonlinear system of partial differential equations for the components $g_{AB}(x)$ and $g_{\mu\nu}(x)$ of the metric tensor.

It is the aim of the subsequent considerations to reformulate these differential equations in the (coordinate-free) language of the minimal surface calculus. For this purpose we consider a stationary and axisymmetric spacetime (e.g. in the form (6)) and define the action integral (Neugebauer 1970)

$$L = \int\limits_{t=t_0} d^3x\, \sqrt{-g}\left(\frac{R}{2\kappa_0} - p\right) \tag{18}$$

over a space-like hypersurface $t = t_0 =$ constant. (g is the fundamental determinant $g = \det g_{ik}$).

To get an explicit expression for p, one has to choose an equation of state and to integrate the Euler equations (11). Since we want to confine ourselves to rigidly rotating bodies we may fix a connection $\varepsilon = \varepsilon(p)$ and integrate (13). The resulting p depends, via V, on the metric g_{ik} (cf. (4b)) and, besides of other constant parameters, on the constants V_0 (cf. (17)) and Ω (cf. (4b)). Thus, L is a functional of the metric g_{ik} and a function of the red shift parameter V_0 and the angular velocity Ω.

$$L = L\{g_{ik}, V_0, \Omega\} \tag{19}$$

To derive a (global) Gibbs relation for L we compare two infinitesimally neighbouring states $\{g_{ik}, V_0, \Omega\}$ and $\{g_{ik} + \delta g_{ik}, V_0 + \delta V_0, \Omega + \delta\Omega\}$. Then the variation of the matter part in (18) yields (cf. (13))

$$\begin{aligned}\delta(p\sqrt{-g}) = {} & \frac{1}{2}\sqrt{-g}\left((\varepsilon + p)\, u^i u^k + p g^{ik}\right)\delta g_{ik} \\ & + \sqrt{-g}\left((\varepsilon + p)\,\eta_i u^i\, e^{-V}\delta\Omega + (\varepsilon + p)\,\delta V_0\right) \ .\end{aligned} \tag{20}$$

Finally, for variations δg_{ik} vanishing at the boundary of the domain of integration we obtain (Neugebauer 1988)

$$\begin{aligned}\delta L = {} & -\mathcal{J}\,\delta\Omega - \mathcal{N}\,\delta V_0 - \frac{1}{2}\,\delta M \\ & - \frac{1}{2\kappa_0}\int\limits_{t=t_0} d^3x\, \sqrt{-g}\,\delta g_{ik}\left(R^{ik} - \frac{1}{2}R g^{ik} + \kappa_0\, T_{ik}\right) \ ,\end{aligned} \tag{21}$$

where

$$T^{ik} := (\varepsilon + p)\, u^i u^k + p g^{ik} \,, \tag{22}$$

$$\mathcal{J} := \int\limits_{t=t_0} d^3x \, \sqrt{-g} \left((\varepsilon + p)\, e^{-V} \eta_i u^i \right) , \tag{23}$$

$$\mathcal{N} := \int\limits_{t=t_0} d^3x \, \sqrt{-g} (\varepsilon + p) \,. \tag{24}$$

T^{ik} denotes the energy momentum of the perfect fluid of the rotating body. M is the total mass (energy) of the source ($1/2 \cdot \delta M$ comes from a surface term) and $\mathcal{J}$ is its angular momentum. From (21) we may conclude

$$(i) \quad \delta \left(L + \frac{M}{2} \right) \Big|_{V_0, \Omega} = 0 \iff R^{ik} - \frac{1}{2} R g^{ik} = -\kappa_0 T^{ik} \tag{25}$$

$$(ii) \quad \text{If} \quad R^{ik} - \frac{1}{2} R g^{ik} = -\kappa_0 T^{ik} \quad \text{then}$$

$$\delta \left(L + \frac{M}{2} \right) = -\mathcal{J}\, \delta\Omega - \mathcal{N}\, \delta V_0 \,. \tag{26}$$

The first statement implies that the Einstein equations of rigidly rotating bodies can be derived as Euler-Lagrange equations from a variational principle. This theorem remains valid for nonrigid (differential) rotation too (in this case one has to replace (13) by (11) (Kramer 1988)). Furthermore, the relations (i) and (ii) give rise to a thermodynamic interpretation: The rotating body (matter and field) is a thermodynamic system characterized by the thermodynamic potential

$$\mathcal{L} = L + \frac{M}{2} \,, \tag{27}$$

which depends on the order parameter $g_{ik}(x)$ and the control parameters Ω and V_0. The equations in *(i)* define the equilibrium state of the system, i.e. the Einstein equations are the equilibrium conditions. For equilibrium states (solutions $g_{ik}(x)$ of the Einstein equations), $\mathcal{L} = L + M/2$ is an Gibbs potential satisfying the Gibbs relation

$$\delta \mathcal{L} = -\mathcal{J}\, \delta\Omega - \mathcal{N}\, \delta V_0 \tag{28}$$

and yielding the 'global' equations of state

$$M = M(V_0, \Omega) \,, \qquad \mathcal{J} = \mathcal{J}(V_0, \Omega) \tag{29}$$

which connect the exterior (far field) parameters mass M and angular momentum $\mathcal{J}$ with the interior parameters V_0 and Ω. To derive the constitutive equations (29) from $\mathcal{L}$ the following equilibrium relations are useful:

$$\mathcal{L} = \frac{1}{2} \int\limits_{t=t_0} d^3x \, \sqrt{-g}(p - \varepsilon) + \frac{M}{2} \tag{30}$$

$$M = 2\,\Omega\mathcal{J} + \int\limits_{t=t_0} d^3x \, \sqrt{-g}(\varepsilon + 3p) \tag{31}$$

$$\mathcal{N} = M - \Omega\mathcal{J} - \mathcal{L} = \int\limits_{t=t_0} d^3x \, \sqrt{-g}(\varepsilon + p) \, , \tag{32}$$

where we have employed the Einstein equations (25) and the Killing equations (3). By means of (32), the Gibbs relation (28) takes the form

$$\delta\mathcal{L} = -\mathcal{J}\,\delta\Omega - (M - \Omega\mathcal{J} - \mathcal{L})\,\delta V_0 \tag{33}$$

or

$$\delta\mathcal{S} = e^{-V_0}(\delta M - \Omega\,\delta\mathcal{J}) \tag{34}$$

after a Legendre transformation

$$\mathcal{S} = e^{-V_0}(M - \mathcal{L} - \Omega\mathcal{J}) \, . \tag{35}$$

For a solution $g_{ik}(x)$, $\mathcal{L} = \mathcal{L}(V_0, \Omega)$ can be calculated from $\varepsilon = \varepsilon(V)$ and $p = p(V)$ via (30), (31) and (23), in principle. (A more convenient algorithm will be given in the next section, cf. (46)).

Now the parameter relations (29) are a simple consequence of (33),

$$M = \Omega\mathcal{J} + \mathcal{L} - \left.\frac{\partial\mathcal{L}}{\partial V_0}\right|_{\Omega} \, , \quad \mathcal{J} = -\left.\frac{\partial L}{\partial\Omega}\right|_{V_0} \, . \tag{36}$$

4 Minimal Surface Formulation

We define in an invariant way four gravitational potentials, the Newtonian potential U, the gravitomagnetic potential A, the axis potential W and the superpotential α:

$$\begin{aligned} e^{2U} &= -\,\xi_i\xi^i \, , & A &= -e^{-2U}\eta_i\xi^i \\ W^2 &= (\eta_i\xi^i)^2 - \eta_k\eta^k\,\xi_i\xi^i \, , & e^{-2\alpha} &= e^{-2U}W_{,i}W^{,i} \end{aligned} \tag{37}$$

In terms of these potentials the line element (6) reads (Neugebauer and Herlt 1984)

$$\begin{aligned} ds^2 = {} & e^{-2U}\left(e^{2\alpha}W_{,C}W_{,D}\,h^{CD}h_{AB}\,dx^A dx^B + W^2 d\varphi^2\right) \\ & - e^{2U}(dt + A\,d\varphi)^2 \, , \end{aligned} \tag{38}$$

where U, A, W, α only depend on $x^A (A = 1, 2)$. U is the generalization of the Newtonian gravitational potential for weak fields. A represents the

rotation of the source and plays the same role as the azimuthal component of the electromagnetic vector potential does for rotating charges. For rotating stars, its order of magnitude compared with the centrifugal potential, $2\Omega A/\Omega^2 R^2$(R : radius of the star), varies between 0.2 (rapidly rotating neutron stars) and 10^{-7} (sun). In a Minkowski space, $W = \varrho$ (ϱ, z, φ: cylindrical coordinates) measures the distance from the symmetry axis (z-axis) of a rotational body. The denotation 'super potential' comes from chiral field theory.

Inserting the metric (38) in the Lagrangian (18) and taking advantage of the axisymmetry one obtains after a straightforward calculation

$$\mathcal{L} = L + \frac{M}{2} = \frac{2\pi}{\kappa_0} \int\limits_{\Sigma} d^2x \, \sqrt{\gamma} \, , \tag{39}$$

where γ is the determinant of the coefficients γ_{AB} $(A, B = 1, 2)$ of the first fundamental form (metric) of the 2-surface Σ ,

$$\begin{aligned} U &= U(x^1, x^2) \, , \qquad A = A(x^1, x^2) \, , \\ W &= W(x^1, x^2) \, , \qquad \alpha = \alpha(x^1, x^2) \, , \end{aligned} \tag{40}$$

embedded in a (Pseudo-)Riemannian Potential Space (RPS) with the line element

$$dS^2 = -2\, d\alpha\, dW + 2W\, dU^2 - \frac{e^{4U}}{2W}\, dA^2 - 2\kappa_0 W\, e^{2\alpha - 2U} p(V)\, dW^2, \tag{41}$$

where p is a given function of the co-rotating Newtonian potential V,

$$\begin{aligned} V &= \frac{1}{2} \ln(-1) \left(\xi^i + \Omega\eta^i\right) \left(\xi_i + \Omega\eta_i\right) \\ &= U + \frac{1}{2} \ln\left([1 + \Omega A]^2 - \Omega^2 W^2\, e^{-4U}\right) \, . \end{aligned} \tag{42}$$

By the same straightforward calculation it turns out, that the surface metric

$$\begin{aligned} \gamma_{AB} = &- \left(\alpha_{,A} W_{,B} + \alpha_{,B}\, W_{,A}\right) + 2W\, U_{,A} U_{,B} - \frac{e^{4U}}{2W}\, A_{,A} A_{,B} \\ &- 2\kappa_0 W\, e^{2\alpha - 2U} p(V)\, W_{,A} W_{,B} \end{aligned} \tag{43}$$

is conformally equivalent to the space-time metric h_{AB} ,

$$\gamma_{AB} = \lambda\, h_{AB} \, , \qquad (\lambda \quad \text{a conformal factor}) \tag{44}$$

so that h_{AB} in (38) can be replaced by γ_{AB} ,

$$\begin{aligned} ds^2 = &\, e^{-2U} \left(e^{2\alpha} W_{,C} W_{,D} \gamma^{CD} \gamma_{AB}\, dx^A dx^B + W^2 d\varphi^2\right) \\ &- e^{2U} (dt + A\, d\varphi)^2 . \end{aligned} \tag{45}$$

Obviously, (39) provides a geometrical interpretation of the Lagrangian $\mathcal{L}$: $\mathcal{L}$ is (apart from the factor $2\pi/\kappa_0$) the area $\mathcal{A}$ of the surface Σ ,

$$\mathcal{L} = \frac{2\pi}{\kappa_0}\mathcal{A} \,. \tag{46}$$

Now we can geometrize the relations (21), (25), and (26). Eq. (25) tells us that the Einstein equations for axisymmetric stationary gravitational fields of rigidly rotating bodies are equivalent to the statement

$$\delta \int_{\Sigma} d^2x \, \sqrt{\gamma} = 0 \,, \tag{47}$$

i.e. the surfaces (40) are "minimal" surfaces (Neugebauer and Herlt 1984, Neugebauer 1979, 1985). We choose this term for the sake of simplicity, merely. (Correctly, we should say that $\mathcal{A}$ has a stationary value for the surfaces considered). Hence, *the axisymmetric and stationary gravitational fields of rigidly rotating bodies are 'minimal' surfaces in the (Pseudo-)Riemannian Potential Space* (41). Moreover, the thermodynamic potential $\mathcal{L}$ is essentially the area of the 'minimal' surface, cf. (46).

Now the gravitational field equations are the Euler-Lagrange equations of the geometrical variational principle (47). Having solved those second order equations, i.e. having calculated the four fields (40) one obtains γ_{AB} in (43) by differentiation and finally the space-time metric (45) without any further integration.

5 Parametrizations of the Minimal Surfaces

Without loss of generality we may choose surface coordinates $x^A (A = 1, 2)$ such that γ_{AB} is conformally euclidian,

$$\gamma_{AB} = \lambda_1 \, \delta_{AB} \qquad (\lambda_1 \text{ a conformal factor}) \,. \tag{48}$$

Introducing the more compact notation

$$\{\varphi^i\} = \{U, A, W, \alpha\} \,, \qquad dS^2 = G_{ik} \, d\varphi^i d\varphi^k \tag{49}$$

for (41) and denoting the 2-dimensional gradient by ∇ , the Lagrangian (39) can be written as

$$\mathcal{L} = \frac{\pi}{\kappa_0} \int_{\Sigma} d^2x \; G_{ik} \, \nabla\varphi^i \cdot \nabla\varphi^k \,. \tag{50}$$

The Euler-Lagrange equations corresponding to (47)

$$\Delta\varphi^i + \Gamma^i_{kl} \, \nabla\varphi^k \cdot \nabla\varphi^l = 0 \tag{51}$$

are a system of semilinear partial differential equations of second order (Δ is the 2-dimensional Laplacian and Γ^i_{kl} the Christoffel symbol of the RPS metric G_{ik}). Additionally, the auxiliary conditions (43) with (44) have to be taken into consideration.

We will use this *Weyl* parametrization (see (45) for static fields ($A = 0$)) for our numerical calculation in part *II*, cf. (3) and (6) there.

If one chooses the parametrization

$$W = W(U, A) , \qquad \alpha = \alpha(U, A) , \tag{52}$$

the Euler-Lagrange equations reduce to two second order partial differential equations ('minimal' parametrization).

6 Physical Minimal Surfaces

The interior parameters V_0 and Ω are well-suited to classify rotating body solutions. No exact rotating body solution is known in the general case ($V_0 \neq 0$, $\Omega \neq 0$). The generation techniques (Bäcklund transformations, Inverse (scattering) method) for asymptotically flat vacuum solutions are not applicable to the interior of the body. Approximative solutions for discs (rotating dust) were discussed by Bardeen and Wagoner (1969, 1971). In this case the analysis reduces to a boundary value problem for the vacuum equations (e.g. the Ernst equations). Recently, Meinel and Neugebauer (1992) have found the exact solution of this problem by solving the corresponding Riemann-Hilbert problem. In the following three cases explicit solutions are known:

(*i*)	$0 \leq \lvert V_0 \rvert < \infty$,	$\Omega = 0$:	'Static stars'
(*ii*)	$\lvert V_0 \rvert \ll 1$,	$\Omega \neq 0$:	'Newtonian stars'
(*iii*)	$\lvert V_0 \rvert \gg 1$,	$\Omega \neq 0$:	'Black Holes'

We will end up with some remarks on these three cases:

(*i*) All *static* spherically symmetric minimal surfaces (interior and exterior fields) can be written in the form

$$e^{-2\alpha} = 1 + W^2 f(U) , \tag{53}$$

where $f(U)$ has to satisfy an ordinary second order differential equation. To incorporate the boundary conditions (Section 7) the following parametrization is more convenient:

$$\begin{aligned} e^{-2\alpha} &= 1 + B \sin^2 \vartheta \\ W &= A \sin \vartheta , \end{aligned} \tag{54}$$

where A and B are functions of U alone. The boundary (regularity) condition on the symmetry axis ($\vartheta = 0, \pi$) is then automatically satisfied. To ensure $W = 0$ in the center ($U = U_C$) and $\alpha = 0$ at infinity ($U = 0$) we have to postulate

$$A(U_C) = 0 \,, \qquad B(0) = 0 \,. \tag{55}$$

The minimal surface equation takes the form

$$A'(U) = A\sqrt{\frac{1+B}{B-\kappa_0 PA^2}} \,, \tag{56}$$

$$B'(U) = (4\kappa_0 PA^2)\sqrt{\frac{1+B}{B-\kappa_0 PA^2}} \,, \tag{57}$$

$$P := p(U)\, e^{-2U} \,. \tag{58}$$

An algorithm for a numerical solution of this system under the conditions (55) was given in (Neugebauer 1990).

(*ii*) The 'Newtonian' Potential Space

$$dS^2 = -2\, dW d\alpha + 2W\, dU^2 - 2\kappa_0 W p(V)\, dW^2 \tag{59}$$

follows from (41) for $A \equiv 0$, $\alpha \ll 1$, $U \ll 1$. In this approximation, V takes the form

$$V = U - \frac{1}{2}\,\Omega^2 W^2 \,. \tag{60}$$

The first incompressible fluid solution found by Maclaurin describes the field of a rotating 2-axial oblate ellipsoid bounded by the surface

$$V \equiv U - \frac{1}{2}\,\Omega^2 W^2 = V_0 \,. \tag{61}$$

Here, the Gibbs potential $\mathcal{L}$ takes the form

$$\mathcal{L} = \varrho^2 \left(\frac{-V_0}{\varrho}\right)^{\frac{5}{2}} l\left(\frac{\Omega}{\sqrt{\varrho}}\right) \,, \tag{62}$$

where ϱ is the constant mass density (l is a well-defined function of $\Omega/\sqrt{\varrho}$).

A survey of other Newtonian solutions of the rotating body problem may be found in (Chandrasekhar 1969). They can be used to test numerical solution techniques for 'minimal' surface equations.

(*iii*) Black Hole thermodynamics is closely related to the parameter thermodynamics (34) (Neugebauer 1985).

7 Boundary Conditions

The 'natural' border of the 'minimal' surface is the symmetry axis (z-axis) of the body, its equatorial plan (if it exists) and a circle at infinity connecting axis and equatorial plane (Fig. 3). (In what follows, we will postulate reflection symmetry and therefore the existence of an equator.) The conditions defining the axis, the equator and infinity can be taken from Fig. 3:

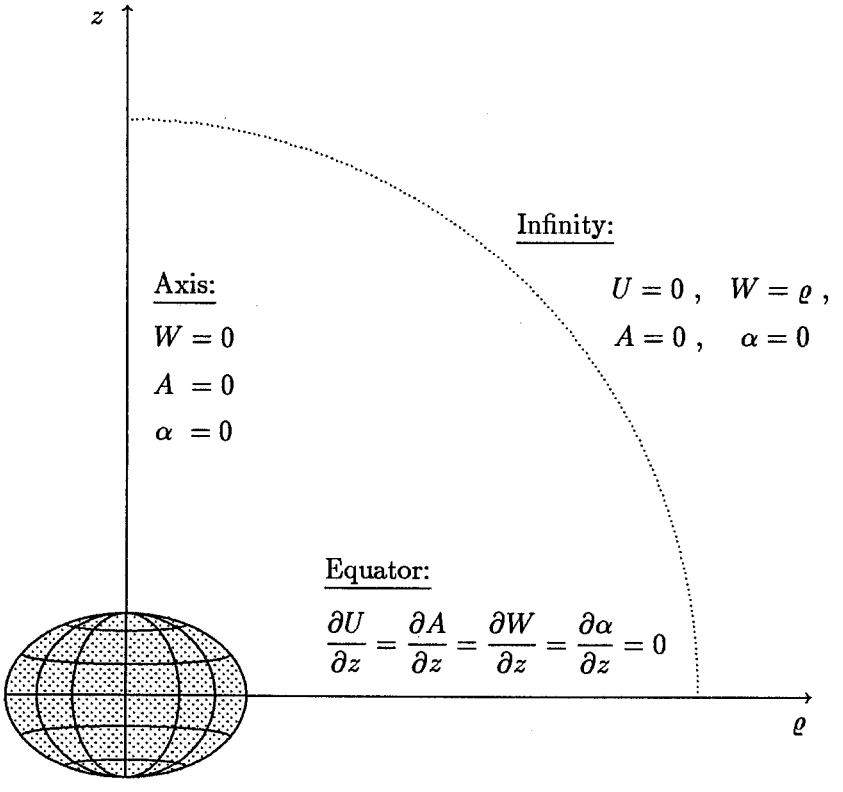

Fig. 3. The boundary conditions on the border of the 'minimal' surface (symmetry axis, equatorial plane and infinity)

The axis conditions express microeuclidicity ($\alpha = 0$, cp. (Kramer et al. 1980) and symmetry around the axis ($|\eta^i| \to 0$, cp. (37)). Since the gravitational potentials are symmetric with respect to the equator, their normal derivatives have to vanish there. At infinity, the space-time (45) is minkowskian (in cylindrical coordinates). It should be mentioned that the border is a null curve ($dS^2 = 0$), whereas the metric of the 'minimal' surface inside the border is definite.

References

Bardeen, J.M., Wagoner, R.W. (1969): Ap. J. **158** L65

Bardeen, J.M., Wagoner, R.W. (1971): Ap. J. **167** 359

Belinsky, V.A., Zakharov, V.E. (1978): Zh. Eksp. Teor. Fiz. **75** 1953

Chandrasekhar, S. (1969): "Ellipsoidal Figures of Equilibrium", Yale Univ. Press, New Haven and London

Harrison, B.K. (1978): Phys. Rev. Lett. **41** 119

Hauser, I., Ernst, F.J. (1980): J. Math. Phys. **21** 1126

Hoenselaers, C., Kinnersley, W., Xanthopoulos, B. (1979): J. Math. Phys. **20** 2530

Kramer, D. (1988): Astron. Nachr. **309** 267

Kramer, D., Stephani, H., MacCallum, M., Herlt, E. (1980): "Exact Solutions of Einstein's Field Equations", Dt. Verlag d. Wiss., Berlin

Kundt, W., Trümper, M. (1966): Z. Phys. **192** 419

Lewis, T. (1932): Proc. Roy. Soc. Lond. **A 136** 176

Maison, D. (1978): Phys. Rev. Lett. **41** 521

Meinel, R., Neugebauer, G. (1992) to be published

Neugebauer, G. (1970): Habilitationsschrift (Jena University)

Neugebauer, G. (1979): J. Phys. A: Math. Gen. **12** L67

Neugebauer, G. (1985): in "Proceedings of the Balatonszéplak Relativity Workshop", ed. by Z. Perjes (KFKI Publ., Central Research Institute for Physics, Budapest) p. 103

Neugebauer, G. (1988): in "Relativity Today, Proceedings of the 2nd Hungarian Workshop", ed. by Z. Perjes (World Scientific, Singapore) p. 134

Neugebauer, G. (1990): Ann. Phys. (Leipzig) **47** 177

Neugebauer, G., Herlt, E. (1984): Class. Quantum Grav. **1** 695

Neugebauer, G., Kramer, D. (1969): Ann. Phys. (Leipzig) **24** 62

Papapetrou, A. (1966): Ann. Inst. H. Poincaré **A 4** 83

Gravitational Fields of Rapidly Rotating Neutron Stars: Numerical Results

Heinz Herold [1], Gernot Neugebauer [2]

[1]Lehr- und Forschungsbereich Theoretische Astrophysik,
Universität Tübingen, W-7400 Tübingen, Germany
[2]Theoretisch-Physikalisches Institut der
Friedrich-Schiller-Universität Jena, O-6900 Jena, Germany

1 Introduction

The first observation of millisecond pulsars in 1982 (Backer et al. 1982) has stimulated the research on rapidly rotating neutron stars, particularly as in the meantime the number of observed sources in this category has steadily increased. Since neutron stars are extremely compact objects with strong gravity (for a typical neutron star the ratio between radius and Schwarzschild radius is approximately 2), general relativity must be employed in modelling such stars in order to be able to draw definite conclusions from the observations.

General relativistic calculations of *rotating* neutron stars are relatively rare. The case of slow rotation has been treated for the first time by Hartle and Thorne (see Hartle 1967; Hartle and Thorne 1968) and later by other authors (see Datta 1988, e.g.). The construction of realistic models for *rapidly* rotating neutron stars has been performed by Friedman et al. (1986) (see also Friedman et al. 1989) derived from a method developed by Butterworth and Ipser (1976). The numerical method employed by these authors is based on the idea to discretize directly Einstein's field equations for a chosen parametrization of the metric (to be more precise, they use the linearized equations obtained by Newton's method). To the authors' knowledge, there are no other groups which have developed another independent approach for the problem of rapidly rotating neutron stars.

In this contribution we will show that the minimal surface formalism, as described in Neugebauer and Herold (1992) (in the following cited as paper I), is well suited for the numerical calculation of the gravitational fields and the structure of rapidly rotating neutron stars. Especially the minimal surface extremum principle facilitates the procedure of discretizing very much,

since it is not necessary to consider the field equations explicitly which are rather complicated indeed. After a short description of the basic formulation of our approach in Sect. 2, the numerical procedure is explained in some detail in Sect. 3. Typical results of the numerical calculation are presented in Sect. 4, where the gravitational fields as well as global properties of fast rotating neutron stars are discussed. Furthermore, the solutions are visualized by embedding diagrams and 4D ray-tracing pictures.

2 Basic Formulation

In this section the basic equations are summarized which are used for the description of the structure of rapidly rotating neutron stars and their gravitational fields (for more details, see paper I). The space-time generated by an isolated rotating star admits two commuting Killing vectors: the (at least asymptotically) time-like vector $(\xi^i) = \partial/\partial t$ and the space-like (azimuthal) vector $(\eta^i) = \partial/\partial\varphi$. Additionally, it is assumed that the matter is a perfect fluid with the energy-momentum tensor

$$T_{ik} = (\varepsilon + p)\, u_i u_k + p\, g_{ik} \tag{1}$$

and rotates rigidly with the angular velocity Ω. Thus, the four velocity is given by

$$u^i = e^{-V}(\xi^i + \Omega\eta^i) \quad . \tag{2}$$

The minimal surface formulation for the stationary and axially symmetric field equations requires that the metric is parametrized in the following form:

$$ds^2 = -e^{2U}\,(dt + A\,d\varphi)^2 + e^{-2U}\left(e^{2\alpha}\,(\nabla W)^2\,(dr^2 + r^2 d\theta^2) + W^2 d\varphi^2\right) \tag{3}$$

Therein, we have specialized the – in principle arbitrary – "meridional" coordinates x^1, x^2 (which are needed additionally to the Killing coordinates t, φ) to spherical coordinates r, θ, which are related to the quasi-Euclidean cylindrical coordinates ρ, z by $\rho = r\sin\theta$, $z = r\cos\theta$. The gradient operator is used here in the normal Euclidean meaning, i.e. $(\nabla W)^2 = W_{,r}^2 + 1/r^2\, W_{,\theta}^2$. It should be noted that the potentials $U(r,\theta)$, $W(r,\theta)$, $A(r,\theta)$, $\alpha(r,\theta)$ are scalar quantities which can invariantly be defined through the Killing vectors (see paper I).

The normalization of the four vector (2) of the matter yields an expression for the quantity V, which represents in the non-relativistic limit the sum of the Newtonian and the centrifugal potential:

$$V = U + \frac{1}{2}\ln\left((1 + \Omega A)^2 - \Omega^2 W^2 e^{-4U}\right) \tag{4}$$

A consequence of energy-momentum conservation is that the energy density ε as well as the pressure p only depend on V and are related by

$$\varepsilon(p) + p = -\frac{dp}{dV} \quad . \tag{5}$$

For a given equation of state (EOS) $\varepsilon = \varepsilon(p)$, the differential equation (5) can be integrated to obtain the function $p = p(V)$. The zero point of V is fixed by the prescription of vanishing pressure on the surface, $p(V_0) = 0$.

Thus, the parameters which characterize a rigidly rotating neutron star are the angular velocity Ω and the surface gravity V_0, which describes the compactness of the system.

It can be shown (Neugebauer and Herlt 1984; Neugebauer 1985, 1988) that, for the situation described, Einstein's field equation are equivalent to the minimal surface equations in an abstract Riemannian potential space with a well-defined (indefinite) metric, whose coordinates are the potentials U, W, A, α. Without describing this geometric interpretation in detail (see paper I), one can formulate the problem with the help of an Lagrangian $\mathcal{L}$, whose variation must vanish, $\delta\mathcal{L} = 0$. The Lagrangian is given by

$$\mathcal{L} = \frac{1}{2G}\int_0^{\frac{\pi}{2}} d\theta \int_0^{\infty} dr\, r \left[-\nabla\alpha\nabla W + W(\nabla U)^2 - \frac{e^{4U}}{4W}(\nabla A)^2 \right.$$
$$\left. - 8\pi G\, W e^{2\alpha - 2U} p(V)(\nabla W)^2 \right] \tag{6}$$

and should be considered as a functional of the 4 potentials $\alpha(r,\theta)$, $W(r,\theta)$, $U(r,\theta)$, $A(r,\theta)$. Here, we have additionally assumed symmetry with respect to the equatorial plane ($z = 0$ or $\theta = \pi/2$). The admissible functions α, W, U, A must fulfill the boundary conditions

on the rotation axis ($\rho = r\sin\theta = 0$):

$$\alpha = 0\,, \quad W = 0\,, \quad A = 0\,, \tag{7a}$$

at infinity ($r \to \infty$):

$$\alpha = 0\,, \quad U = 0\,, \quad A = 0\,, \quad W - r\sin\theta = 0\,. \tag{7b}$$

The potential U must be regular on the rotation axis. Furthermore, the unconstrained variation of (6) yields the natural boundary conditions on the equatorial plane

$$\frac{\partial\alpha}{\partial\theta} = 0\,, \quad \frac{\partial U}{\partial\theta} = 0\,, \quad \frac{\partial A}{\partial\theta} = 0\,, \quad \frac{\partial W}{\partial\theta} = 0\,, \tag{7c}$$

which express the reflection symmetry.

Note that it is an essential feature of our formulation that there is no distinction between the interior and the exterior of the star; both regions are treated simultaneously. The outside is just characterized by $p(V) = 0$, i.e. the last term in (6) vanishes. The position and shape of the surface come out automatically from a self-consistent solution.

3 Numerical Procedure

In order to calculate the structure of rapidly rotating stars, a convenient procedure is to start from a non-rotating star and to increase the angular velocity gradually. Therefore, in the following section we want to describe non-rotating neutron stars as our starting point.

3.1 Starting Point: Non-Rotating Neutron Stars

The usual parametrization for a static, spherically symmetric metric reads (see Weinberg 1972, e.g.):

$$ds^2 = -e^{2\nu}\,dt^2 + e^{2\lambda}\,d\bar{r}^2 + \bar{r}^2(d\theta^2 + \sin^2\theta\,d\varphi^2) \tag{8}$$

with the potentials $\nu = \nu(\bar{r})$ and $\lambda = \lambda(\bar{r})$ being only dependent on the radial coordinate $\bar{r}$. (We use here a modified notation for this coordinate to distinguish it clearly from the radial coordinate r in the metric (3) which will be used for the non-rotating as well as the rotating case.)

Einstein's field equations for a perfect fluid with the energy-momentum tensor (1) lead to the following equations (known as Tolman-Oppenheimer-Volkoff (TOV) equations)

$$m(\bar{r}) = 4\pi \int_0^{\bar{r}} \varepsilon(\bar{r}')\,\bar{r}'^2\,d\bar{r}' \tag{9a}$$

$$\frac{d\nu}{d\bar{r}} = \frac{G(m + 4\pi\bar{r}^3 p)}{\bar{r}^2(1 - 2Gm/\bar{r})} \tag{9b}$$

$$\frac{dp}{d\bar{r}} = -(\varepsilon + p)\frac{d\nu}{d\bar{r}} \tag{9c}$$

for the potential ν and the pressure p (which is related to the energy density by the equation of state $\varepsilon = \varepsilon(p)$ or $p = p(\varepsilon)$, respectively), while the other potential λ is given by the algebraic relation

$$e^{-2\lambda} = 1 - \frac{2Gm(\bar{r})}{\bar{r}} \quad . \tag{9d}$$

To obtain a solution of (9a–d), one integrates from the center (i.e. $\bar{r} = 0$) starting from a given central density $\varepsilon = \varepsilon_c$ (corresponding to a central pressure $p = p_c$) up to the radius $\bar{r} = R$ where the pressure p vanishes (assuming as boundary condition vacuum outside the star). This yields then the total mass $M = m(R)$ and by variation of the central density the mass-radius relation for the considered equation of state. (Typical results for various EOSs can be found in Shapiro and Teukolsky (1983), e.g.)

Since we want to use the non-rotating solutions as the starting point for the calculation of rapidly rotating neutron stars, it is appropriate to transform the TOV equations to the coordinates which appear in our general

metric (3). Specializing (3) to the case $\Omega = 0$ (no rotation) is equivalent to setting $A = 0$. Thus, the comparison with the "Schwarzschild coordinate" form (8) reveals that a radial coordinate transformation is sufficient, while the angle coordinates θ and φ can remain unchanged. The explicit transformation formulae read:

$$U = \nu \tag{10a}$$

$$W = e^{\nu}\,\bar{r}\,\sin\theta =: W_0(r)\,\sin\theta \tag{10b}$$

$$\frac{d\bar{r}}{dr} = \frac{\bar{r}}{r}\,e^{-\lambda} \tag{10c}$$

$$e^{-2\alpha} = 1 + \left(\frac{r^2 W_0'^2}{W_0^2} - 1\right)\sin^2\theta \tag{10d}$$

so that in the non-rotating case the potentials U and $W_0 = W/\sin\theta$ only depend on r, and for α we obtain the relation $\exp(-2\alpha) = 1 + f(U)\,W^2$ with a well defined function $f(U)$. (In special cases, e.g. for the inner and outer Schwarzschild solution this function has a simple analytic form (Neugebauer and Herlt 1984).)

Thus, combining (9) and (10) yields the following ordinary differential equations

$$\frac{d\bar{r}}{dr} = \frac{1}{r}\,[\bar{r}(\bar{r} - 2Gm)]^{\frac{1}{2}} \tag{11a}$$

$$\frac{dm}{dr} = 4\pi\frac{\bar{r}^2}{r}\,[\bar{r}(\bar{r} - 2Gm)]^{\frac{1}{2}}\,\varepsilon(p) \tag{11b}$$

$$\frac{dU}{dr} = \frac{G(m + 4\pi\bar{r}^3 p)}{r\,[\bar{r}(\bar{r} - 2Gm)]^{\frac{1}{2}}} \quad . \tag{11c}$$

Since for $\Omega = 0$ the quantity V is identical to the potential U, the pressure function $p = p(V)$, which characterizes the matter, is actually a function of U, and $\varepsilon(p)$ in (11b) must be determined from $\varepsilon = -p - dp/dU$.

The integration of (11a–c) is performed as usual: Starting at $r = 0$ with appropriate initial conditions (the essential parameter is here the difference $U_c - U_0$ between the potential U at the center and the potential U on the surface) one arrives at the surface when $p = 0$. There, matching to the outer Schwarzschild solution, which can be written in our coordinates as

$$e^{U} = \frac{r - \frac{1}{2}GM}{r + \frac{1}{2}GM} \tag{12a}$$

$$W = \left(r - (GM)^2/4r\right)\sin\theta \tag{12b}$$

$$e^{-2\alpha} = 1 + \frac{(GM)^2}{\left(r - (GM)^2/4r\right)^2}\sin^2\theta \quad , \tag{12c}$$

where the Schwarzschild radial coordinate $\bar{r}$ is

$$\bar{r} = \frac{(r + \frac{1}{2}GM)^2}{r} \quad , \tag{12d}$$

yields the mass and the radius of the non-rotating star.

In Fig. 1 we show typical results (equation of state MPA, see Sect. 4.1) for the density profiles $\varepsilon = \varepsilon(r)$ of non-rotating neutron stars for three values of the surface gravity parameter V_0, corresponding to gravitational masses of $M = 0.476 M_\odot$, $M = 1.384 M_\odot$, and $M = 1.559 M_\odot$, respectively. When the central density is increased the radius of the star decreases. Note that the radii which can be read up from this figure are smaller than the radii measured in the "normal" Schwarzschild coordinates (see Tables 1–3). This, of course, originates from our choice of coordinates and demonstrates once more that for such strongly gravitating objects one has to be careful in giving observable quantities.

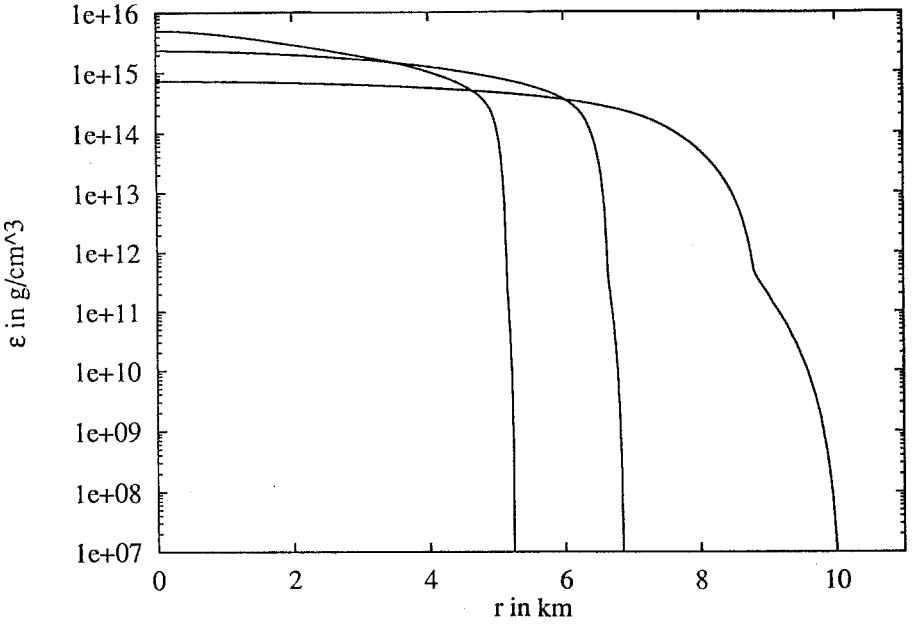

Fig. 1. The density profiles of three non-rotating neutron stars with different masses. The values of the surface gravity parameter V_0 are -0.07, -0.3, and -0.4455 in the order of increasing central density. These values yield the neutron star masses $M = 0.476 M_\odot$, $M = 1.384 M_\odot$, and $M = 1.559 M_\odot$, respectively.

3.2 Solution of the Variational Principle

For rapidly rotating stars, we must solve the variational principle $\delta\mathcal{L} = 0$, i.e. we have to determine those metric potentials α, W, U, A for which the variation of the integral (6) vanishes.

The domain $0 \leq r < \infty$, $0 \leq \theta \leq \pi/2$ of the coordinates (r, θ) in (6) is unbounded and thus not very suitable for the numerical treatment. Therefore, we transform the coordinate r to a new coordinate $\tilde{r}$, which has a finite domain, e.g. $0 \leq \tilde{r} \leq 1$, by the definition $r = S(\tilde{r})$ with a monotonic function S, which should satisfy $S(0) = 0$ and $S(1) = \infty$. There are various possibilities, but a choice which was flexible enough in our actual calculations is

$$r = S(\tilde{r}) = c_0 \frac{\tilde{r}}{1 - \tilde{r}} \qquad \text{or} \qquad \tilde{r} = \frac{r}{r + c_0} \quad , \tag{13}$$

where the constant c_0 can be adapted approximately to the radius of the star (this means that the surface of the star is in the middle of the $\tilde{r}$ domain). Additionally, instead of the angle θ we use $\mu = \cos\theta$ as the second independent coordinate (this avoids the numerically somewhat more expensive trigonometric functions) such that the domain of the variational integral (6) is the unit square in $(\tilde{r}, \mu)$.

It turned out during the numerical calculations that it is the best procedure to implement the boundary condition (7b) for W in the following form. Since $W - r\sin\theta = O(1/r)$ for $r \to \infty$, a new function $\widetilde{W}$ is introduced by

$$W = r\sin\theta + (1 - \tilde{r})\,\widetilde{W} \quad . \tag{14}$$

Then, because of (13) the modified potential $\widetilde{W}$ takes finite values at infinity, which in general are not zero. As the potential α is strongly coupled to W in the functional (6), we had to use instead of α also a new function $\widetilde{\alpha}$ defined by

$$\alpha = (1 - \tilde{r})^2\,\widetilde{\alpha} \quad . \tag{15}$$

Using the behaviour $\alpha = O(1/r^2)$, which can be deduced from the field equations, it follows that also $\widetilde{\alpha}$ takes non-vanishing values at infinity.

In summary, the actual functional which is used in the numerical calculations can be written in the form

$$\mathcal{L} = \int_0^1 d\tilde{r} \int_0^1 d\mu\; I\big(\tilde{r}, \mu, \widetilde{\alpha}, \widetilde{W}, U, A, \widetilde{\alpha}_{,\tilde{r}}, \widetilde{\alpha}_{,\mu}, \widetilde{W}_{,\tilde{r}}, \widetilde{W}_{,\mu}, U_{,\tilde{r}}, U_{,\mu}, A_{,\tilde{r}}, A_{,\mu}\big) \,. \tag{16}$$

To determine the 4 functions $\widetilde{\alpha} = \widetilde{\alpha}(\tilde{r}, \mu)$, $\widetilde{W} = \widetilde{W}(\tilde{r}, \mu)$, $U = U(\tilde{r}, \mu)$, $A = A(\tilde{r}, \mu)$, one has to discretize the integral (16). A natural way is to apply the *Finite Element* approach (see Zienkiewicz 1977, e.g.). The domain is divided into N_e elements with a number of node points (usually at the corners and edges of the elements) which simultaneously belong to neigbouring

elements. On each element an unknown function is approximated by a low order polynomial interpolation through its values at the node points. Thus, e.g. the potential U is given by

$$U(\tilde{r},\mu) = \sum_n U_n f_n(\tilde{r},\mu) \tag{17}$$

where U_n are the node point values of U and $f_n(\tilde{r},\mu)$ are the polynomial shape functions. In total, one has N_n node points and approximately $4N_\mathrm{n}$ unknown function values. (The number of unknowns is somewhat smaller than $4N_\mathrm{n}$ after the boundary conditions (7a,b) have been taken into account.) Since the functional (6) is non-linear in the potentials, in the integration over each element one has to employ a numerical integration procedure, usually a Gauß integration formula.

Since here the domain is a unit square, we have used a very simple finite element discretization, namely rectangular 4-node bilinear finite elements with Gauß-Legendre integration formula. As the results with these ansatz were satisfactory, it was not necessary to turn to more complicated finite elements.

If we denote the set of unknown potential values by X_i $(i = 1, N)$, then the discretized Lagrangian (16) is a (non-linear) function of these variables,

$$\mathcal{L} = \mathcal{L}(X_1, \ldots X_N) \quad . \tag{18}$$

The discretized field equations are equivalent to

$$F_i(X) \equiv \frac{\partial \mathcal{L}}{\partial X_i}(X) = 0 \qquad (i = 1, \ldots N) \quad . \tag{19}$$

This non-linear system of algebraic equations is then solved by the Newton-Raphson method, i.e. by the iteration

$$\frac{\partial F_i}{\partial X_j}(X^{(k)}) \left(X_j^{(k+1)} - X_j^{(k)} \right) = -F_i(X^{(k)}) \tag{20}$$

where the superscript $^{(k)}$ denotes the stage of the iteration. The Newton matrix

$$\frac{\partial F_i}{\partial X_j} = \frac{\partial^2 \mathcal{L}}{\partial X_i \partial X_j} \tag{21}$$

is a symmetrical matrix and can be calculated analytically. (Unfortunately, it is not positive definite, otherwise the solution of the linear system (20) would be easier.) At each Newton step we use a direct sparse matrix solver as linear equation solver.

After convergence we have got a solution represented by the node point values of the 4 potentials α, W, U, A. The actual procedure is to start from a non-rotating solution (see Sect. 3.1), to calculate the potentials on the finite element grid, to determine a converged solution for $\Omega = 0$ by the

finite element iteration (the comparison with the solution from the modified TOV equations gives a good estimate of the discretization error), and then to change the input parameters Ω and/or V_0 slightly to obtain a new solution with the old solution as the starting point of the iteration. The procedure which turned out to be most efficient was to fix the surface gravity parameter V_0 (this determines the mass, at least to a great extent) and to increase the angular velocity Ω.

4 Results

As has been already described in Sect. 2, an essential ingredient for the calculation of *realistic* neutron star models is the equation of state, i.e. the relation $p = p(\varepsilon)$ between pressure p and energy (or mass) density ε. Therefore, in the next section we make some remarks on this topic.

4.1 Equation of State

Up to densities of roughly $10^{14}\,\mathrm{g/cm^3}$ the equation of state $p = p(\varepsilon)$ of cold catalyzed neutron star matter is well known and the physics up to this region is well understood (see Shapiro and Teukolsky 1983, e.g.). But around the nuclear saturation density ($3\ 10^{14}\,\mathrm{g/cm^3}$) and especially above, the situation concerning the EOS is more complicated, on the one hand due to the uncertainties of the parameters of strong interaction, and on the other hand due to the complexity of the nuclear many-body problem. Therefore, a lot of different EOSs for neutron stars exist in the literature (for a review see Arnett and Bowers 1977; Glendenning 1988, e.g.). Since, in this contribution, the main point is to discuss objects with strong gravity (here: neutron stars) and their behaviour in rapid rotation, we do not want to discuss the problem of EOSs in more detail. Just to give the reader a rough impression of how different reasonable EOSs vary in the "region of uncertainty" we present in Fig. 2 the overall behaviour of that EOS (MPA, see Wu et al. 1991) which has been used in producing the results given in the next section, while in Fig. 3, a comparison between this EOS and three other ones, namely EOS BPS (Baym et al. 1971), EOS G (Canuto and Chitre 1974; notation as in Arnett and Bowers 1977) and the pion-condensed EOS π (Weise and Brown 1975) is given.

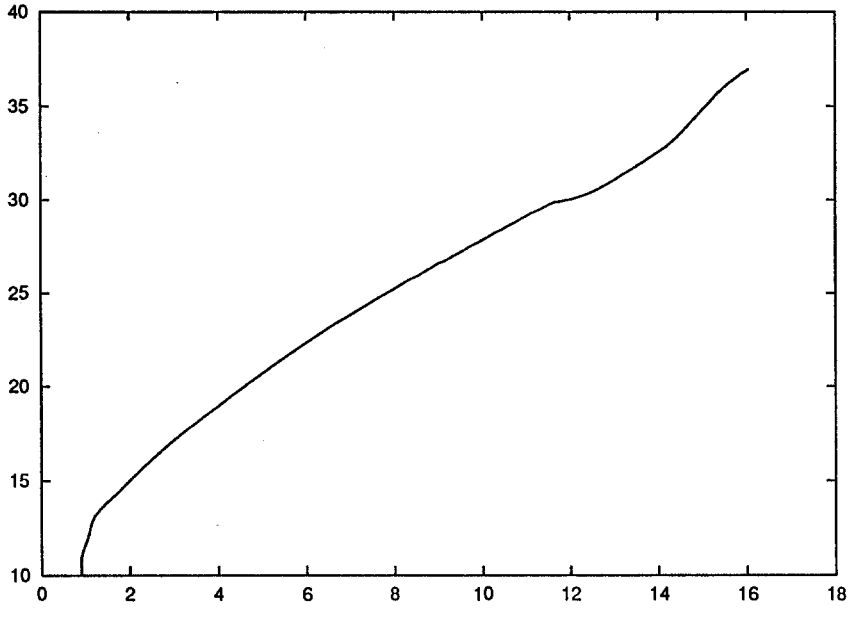

Fig. 2. The equation of state MPA, $p = p(\varepsilon)$, for the whole density range. The picture shows $\log p$ in dyn/cm^2 versus $\log \varepsilon$ in g/cm^3.

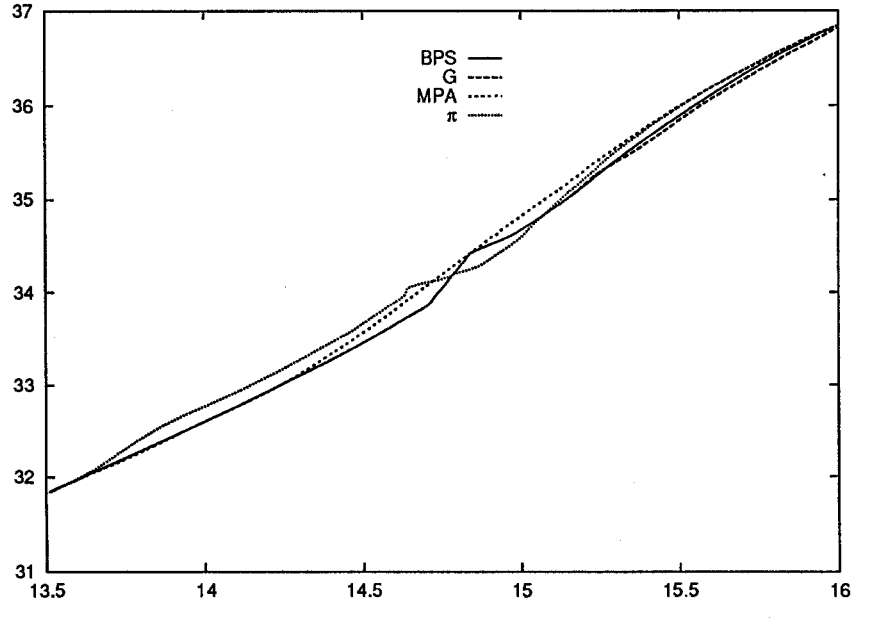

Fig. 3. Comparison of the equations of state BPS, G, MPA, and π around nuclear density and above. The units are the same as in Fig. 3.

4.2 Metric Potentials

For various EOSs we have calculated solutions for many combinations of the basic parameter V_0 and Ω, but to discuss the behaviour of the metric potentials U, W, A, and α, it is sufficient to consider only one EOS. We have chosen the equation of state MPA. Furthermore, we restrict ourselves to one parameter value of V_0, namely $V_0 = -0.3$. The gravitational masses of those neutron stars vary approximately between $1.38\,M_\odot$ and $1.40\,M_\odot$ for all possible values of the angular velocity Ω (see Table 2). (This example shows, as claimed before, that also for rapid rotation the essential parameter which determines the mass is the surface gravity.) Additionally, it turns out that the metric potentials are very similar in their qualitative behaviour for different rotation rates. Thus, we consider here only the model with the maximum angular velocity, i.e. for $V_0 = -0.3$ the solution with $\Omega \approx 8.95\ 10^3\ \mathrm{s}^{-1}$. (see Sect. 4.3) For this neutron star, in Figs. 4–7 the metric potentials U, W, α, and A are shown in their dependence on the coordinate r along radial rays from the center to infinity with the angles $\theta = 0°$, $15°$, $30°$, $45°$, $60°$, $75°$, $90°$.

The line which is crossing the different curves in each picture connects points on the surface of the star so that one can distinguish the inside and the outside. Comparing the polar radius and the equatorial radius reveals a rather large deformation due to the rapid rotation. Nevertheless, the splitting of the "Newtonian" potential U (see Fig. 4) in different directions is relatively small. An essential effect of the rotation is that the depth of this potential is decreased when the angular velocity is increased.

As the potential W represents a sort of distance from the rotation axis, we have plotted in Fig. 5 the difference $W - r\sin\theta$ (scale of W is the dimensionless length unit). For $\Omega = 0$, this quantity (as well as W itself) is proportional to $\sin\theta$, but this property is lost when the rotation is taken into account.

The potential α is similar to $W - \rho$ as can be seen in Fig. 6. A difference is its $1/r^2$ behaviour at large r compared to the $1/r$ dependence of $W - \rho$.

In Fig. 7 the gravitomagnetic potential A is presented, which vanishes in the non-rotating case. For small angular velocities it is proportional to $\sin^2\theta$ (this can be deduced from Hartle (1967), e.g.), but this is no longer true for rapid rotation, as can be recognized from Fig. 7.

4.3 Global Properties

In this section we will discuss some global properties of typical solutions. First, there is the gravitational mass (total mass) M which characterizes a star. This quantity can be measured asymptotically through its gravitational action. Quantitatively, this means that near infinity the potential U can be approximated by $U \approx -GM/r$, which leads to the mass formula

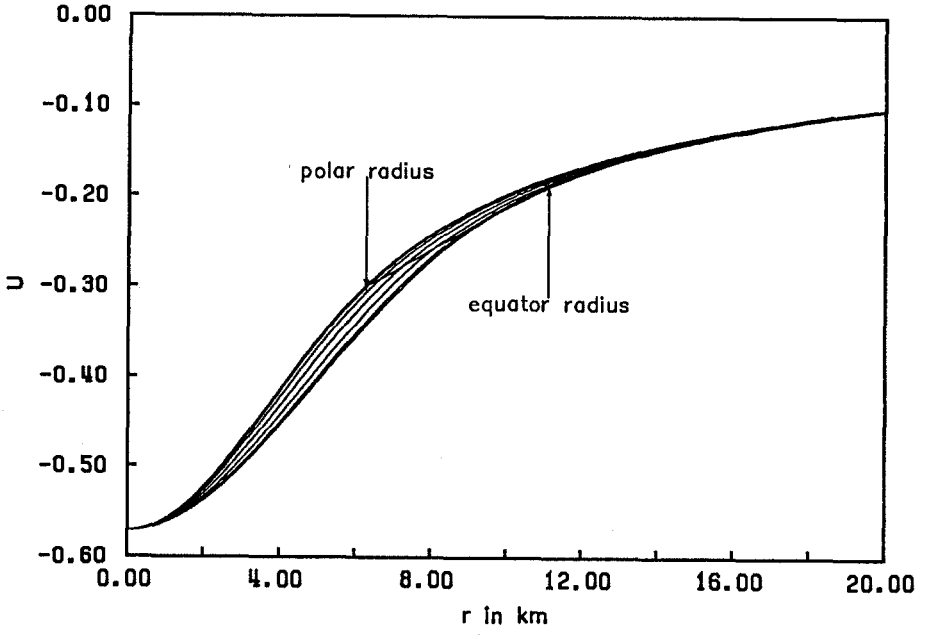

Fig. 4. The potential U of the fastest rotating star of Table 2 as function of the radial coordinate r for different values of the angle θ. The values $\theta = 0°, 15°, 30°, 45°, 60°, 75°, 90°$ belong to the curves from left to right.

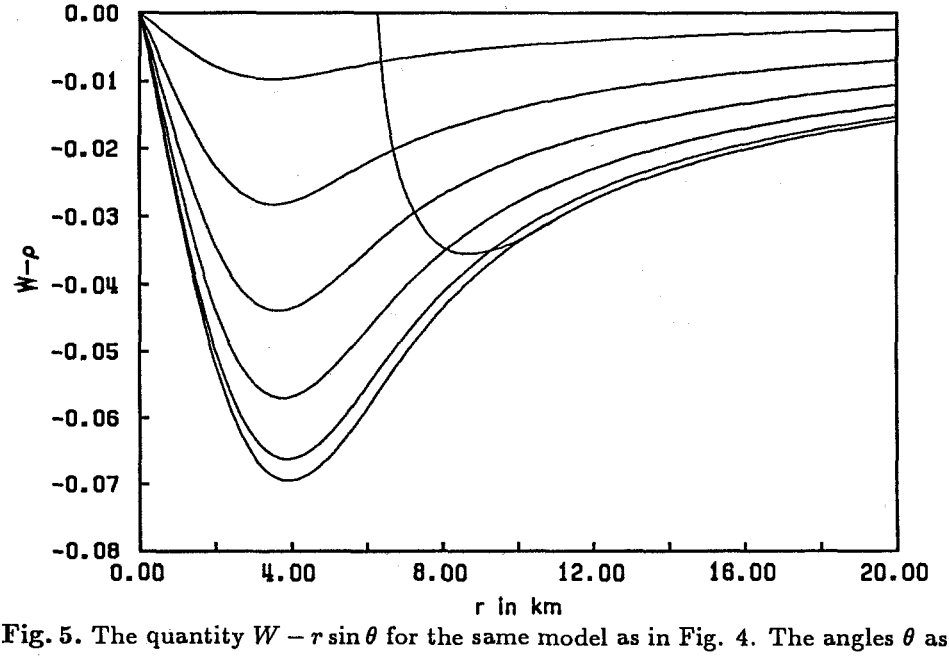

Fig. 5. The quantity $W - r\sin\theta$ for the same model as in Fig. 4. The angles θ as parameters of the different curves are also the same as in Fig. 4, i.e. the function for $\theta = 0°$ vanishes, while the lowest curve corresponds to $\theta = 90°$. The surface of the star is indicated by the crossing line.

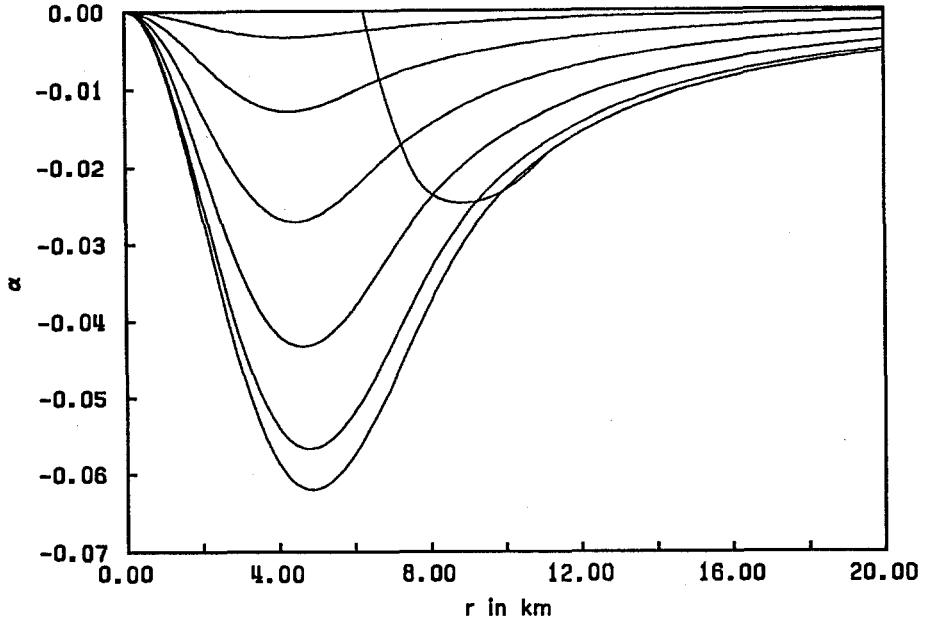

Fig. 6. The potential α for the same model as in Fig. 4. All other remarks in the caption of Fig. 5 apply here, too.

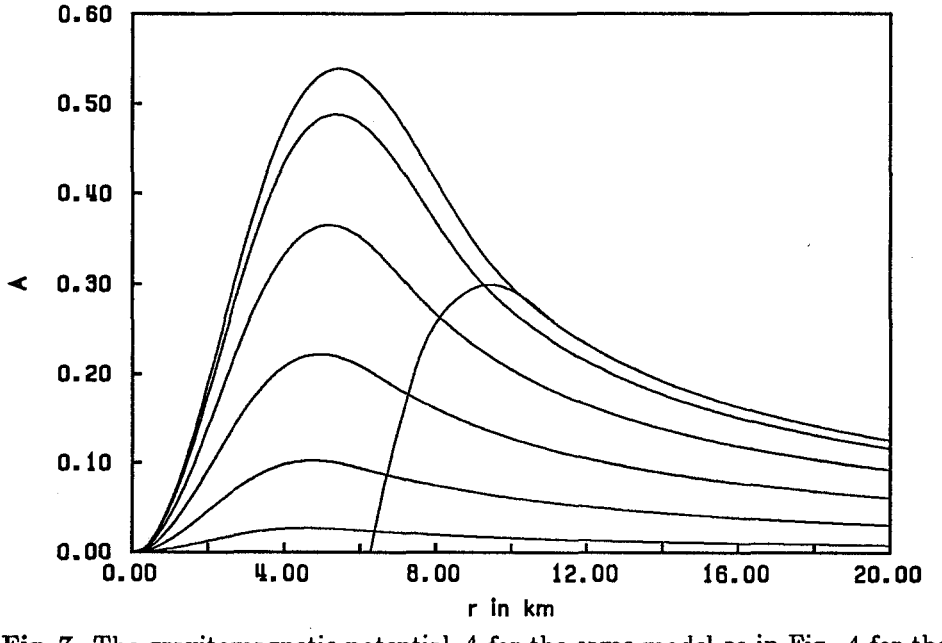

Fig. 7. The gravitomagnetic potential A for the same model as in Fig. 4 for the same angles. Here, the uppermost curve belongs to $\theta = 90°$, while for $\theta = 0°$ the potential A vanishes. Again, the crossing line connects the surface points.

$$G\,M = \lim_{r\to\infty}(-r\,U) \quad . \tag{22a}$$

Alternatively, the total mass has been calculated by an integral over the matter distribution (see paper I). Similarly, the angular momentum J is an asymptotically measurable quantity (by Lense-Thirring precession, e.g.). Therefore, we have an analogous formula:

$$G\,J = \lim_{r\to\infty}\left(\frac{r}{2\sin^2\theta}A\right) \quad . \tag{22b}$$

Relation (22b) can be evaluated for various values of θ, which should always give the same result and thus is a test of the numerical accuracy. Additionally, we have calculated the angular momentum of our solutions by the appropriate integral over the matter (see paper I). All these different methods for the determination of M and J are in good agreement (the deviations are of the order of 0.1 percent or smaller).

The baryon mass (or equivalently the total number of baryons) cannot be determined from the asymptotic fields, rather one has to integrate over the interior of the star (for the formula, see paper I).

In Tables 1–3 we present a series of solutions which are typical for our results and have been calculated with the equation of state MPA. Beside the total mass M and the baryon mass M_0, the central density ϵ_c, and the equatorial radius R (measured by the circumference) are shown. From the angular momentum J the moment of inertia I, defined by the (Newtonian) relation $I = J/\Omega$, is calculated. Additionally, the value of the Lagrangian $\mathcal{L}$ (in units of $M_\odot$) is given.

Table 1. Results for the EOS MPA for $V_0 = -0.07$. The angular velocity Ω is given in s^{-1}, the central density ϵ_c in $10^{15}\,g/cm^3$, the total mass M and the baryon mass M_0 in units of the solar mass $M_\odot$, the equatorial radius R in km, the moment of inertia I in $10^{45}\,g\,cm^2$, and the Lagrangian $\mathcal{L}$ in units of $M_\odot$.

Ω	ϵ_c	M	M_0	R	I	$\mathcal{L}$
0	0.74	0.476	0.494	10.77	0.245	0.0195
1255	0.73	0.474	0.492	10.90	0.248	0.0194
2509	0.71	0.470	0.487	11.47	0.256	0.0191
3137	0.70	0.467	0.484	12.04	0.263	0.0188
3764	0.67	0.463	0.479	13.40	0.275	0.0185
3952	0.67	0.462	0.477	14.41	0.279	0.0184
3990	0.66	0.461	0.477	15.65	0.280	0.0183

Table 2. Results for the EOS MPA for $V_0 = -0.3$. The different quantities are given in the same units as in Table 1.

Ω	ϵ_c	M	M_0	R	I	$\mathcal{L}$
0	2.39	1.384	1.578	9.06	0.866	0.223
5019	2.21	1.382	1.567	9.49	0.918	0.217
6274	2.11	1.381	1.562	9.81	0.956	0.213
7528	1.98	1.383	1.556	10.38	1.022	0.208
8156	1.89	1.386	1.554	10.88	1.077	0.205
8783	1.78	1.393	1.555	11.88	1.168	0.202
8909	1.75	1.396	1.556	12.52	1.196	0.201
8946	1.74	1.397	1.557	12.89	1.206	0.201
8953	1.74	1.398	1.557	13.35	1.207	0.201

Table 3. Results for the EOS MPA for $V_0 = -0.4455$. The different quantities are given in the same units as in Table 1.

Ω	ϵ_c	M	M_0	R	I	$\mathcal{L}$
0	5.09	1.559	1.834	7.81	0.847	0.396
3137	4.90	1.565	1.839	7.90	0.863	0.394
6274	4.30	1.590	1.862	8.24	0.933	0.386
8783	3.65	1.624	1.890	8.82	1.048	0.376
10038	3.25	1.650	1.910	9.37	1.152	0.369
10872	2.92	1.677	1.932	10.06	1.271	0.363
11293	2.72	1.699	1.950	10.78	1.373	0.359
11456	2.63	1.711	1.961	11.74	1.434	0.358

In each table the angular velocity increases from zero to the maximum possible angular velocity, which corresponds to that solution in which the star rotates at the Kepler frequency $\Omega_{\rm K}$, i.e. the angular frequency of a particle in circular orbit at the equator (cf. Friedman et al. 1986). For the metric (3),

$$\Omega_{\rm K} = \frac{1}{2} e^{4U} \frac{A' + 2AU' + \left(A'^2 + 4We^{-4U}U'(W' - WU')\right)^{\frac{1}{2}}}{W(W' - WU') - Ae^{4U}(A' + AU')} \quad , \qquad (23)$$

where primes denote partial derivatives with respect to r and all quantities are evaluated at the equator ($r = r_{\rm eq}$, $\theta = \pi/2$).

No rigidly rotating star can have $\Omega > \Omega_{\rm K}$ and at this point mass shedding at the equator sets in. Certainly, this is the endpoint of such a series of increasing angular velocity. But that does not mean that the solutions

up to this point are stable. There are instabilities due to non-axisymmetric modes which produce gravitational radiation, but those modes are damped by shear viscosity inside the star, which is believed to strongly depend on the temperature. These effects have been understood for non-relativistic stars, but a quantitative treatment in the relativistic regime is lacking (see Weber et al. 1991 and references therein).

As general trends in Tables 1–3 one recognizes that with increasing angular velocity the central density diminishes, the equatorial radius is going up as well as the moment of inertia. The last effects are clearly an indication of rotational flattening of the star. In the next section we will see this more clearly.

4.4 Visualization by Embedding and 4D Ray-Tracing

The gravitional potentials shown in Sect. 4.2 depend on the meridional coordinates r and θ and, thus, their visualization as functions of these coordinates is not invariant against coordinate transformations in the meridional plane. There are certainly many possibilities to characterize the gravitational fields and the structure of the star in a coordinate-independent manner. We have chosen two methods:

First, we have calculated embedding diagrams which visualize the intrinsic geometry of the surface of the neutron star and of internal surfaces of constant pressure (or density). The procedure is as follows: We consider in a (t = const) slice of our stationary space-time a (p = const) surface which can be described by $r = r_s(\theta)$, φ arbitrary. The metric on this two-dimensional axially symmetric surface, which is induced by the four-dimensional metric (3) is (all quantities taken with $r = r_s(\theta)$)

$$ds^2 = e^{2\alpha-2U}(\nabla W)^2 \left[(r_s'^2(\theta) + r_s^2(\theta)\right] d\theta^2 + \left(e^{-2U}W^2 - e^{2U}A^2\right) d\varphi^2 \ . \quad (24)$$

Comparing this with the metric of an axially symmetric 2-surface in Euclidean 3-space, described by $\rho = \rho_e(\theta)$, $z = z_e(\theta)$, φ arbitrary, namely

$$ds^2 = \left[\rho_e'^2(\theta) + z_e'^2(\theta)\right]^2 d\theta^2 + \rho_e^2(\theta)\, d\varphi^2 \ , \quad (25)$$

yields

$$\rho_e(\theta) = \left(e^{-2U}W^2 - e^{2U}A^2\right)^{\frac{1}{2}} \quad (26a)$$

$$z_e'(\theta) = -\left(e^{2\alpha-2U}(\nabla W)^2 \left[(r_s'^2(\theta) + r_s^2(\theta)\right] + \rho_e'^2(\theta)\right)^{\frac{1}{2}} \ . \quad (26b)$$

To integrate (26b) in the interval $0 \leq \theta \leq \pi/2$ one uses the boundary condition $z_e(\pi/2) = 0$. In such a way the embedding functions $\rho_e(\theta)$, $z_e(\theta)$ are obtained. Typical embedding diagrams are presented in Figs. 8–13.

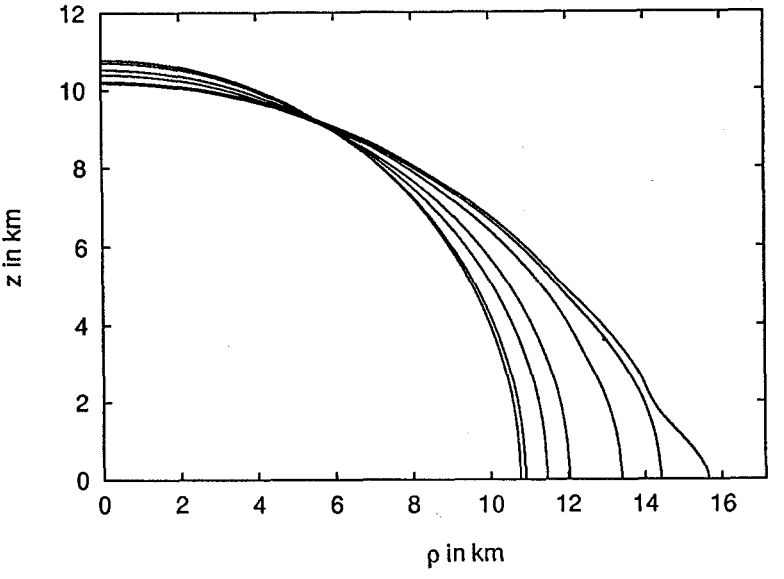

Fig. 8. Embedding diagrams of the surfaces of the stars of Table 1. The equatorial radius increases with angular frequency.

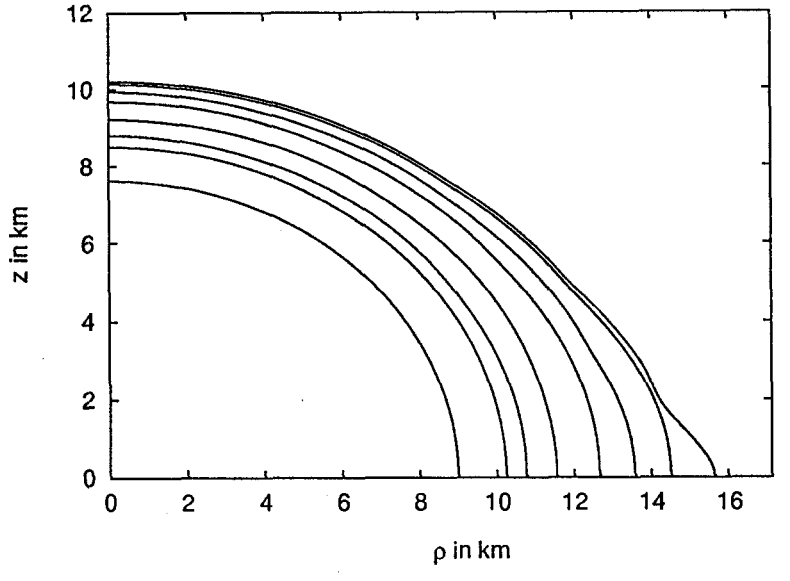

Fig. 9. Embedding diagrams of some internal constant-density surfaces of the fastest star of Fig. 8, which rotates just at the mass shedding limit. The outermost curve represents the surface of the star itself, while the other ones belong to the density values 10^7, 10^9, 10^{10}, 10^{11}, 10^{12}, 10^{13}, 10^{14} g/cm^3.

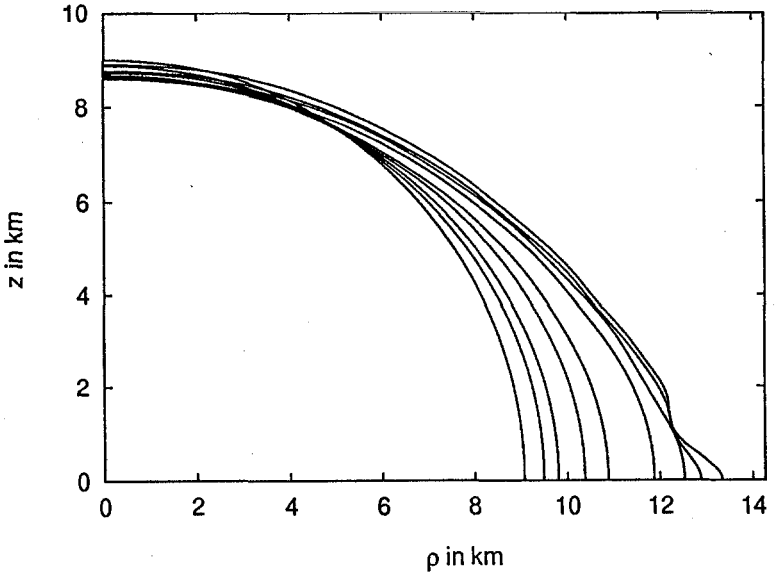

Fig. 10. Embedding diagrams of the surfaces of the stars of Table 2.

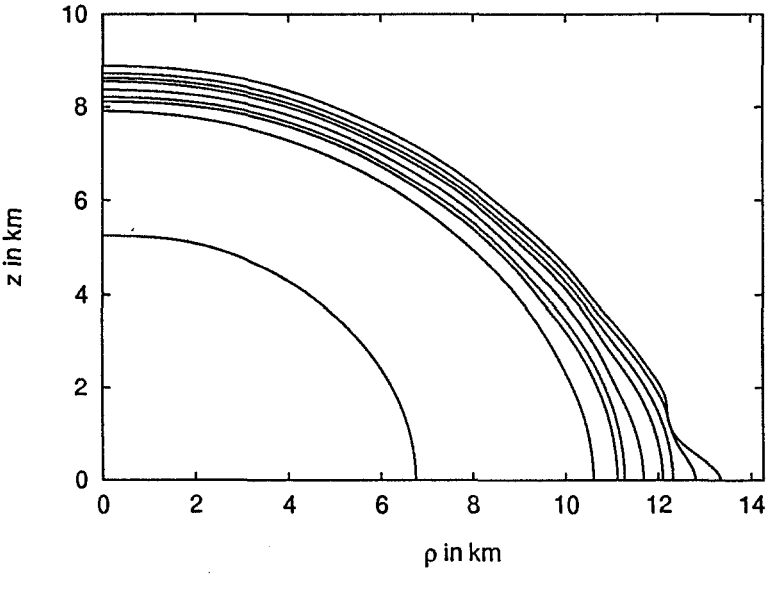

Fig. 11. Embedding diagrams of some internal constant-density surfaces of the fastest star of Fig. 10. The innermost surface belongs to the density 10^{15} g/cm^3, the other ones are the same as in Fig. 9.

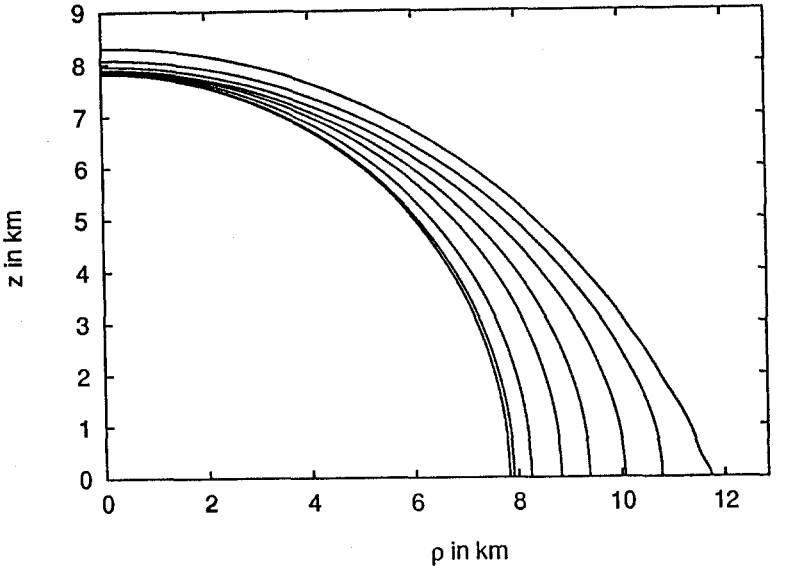

Fig. 12. Embedding diagrams of the surfaces of the stars of Table 3.

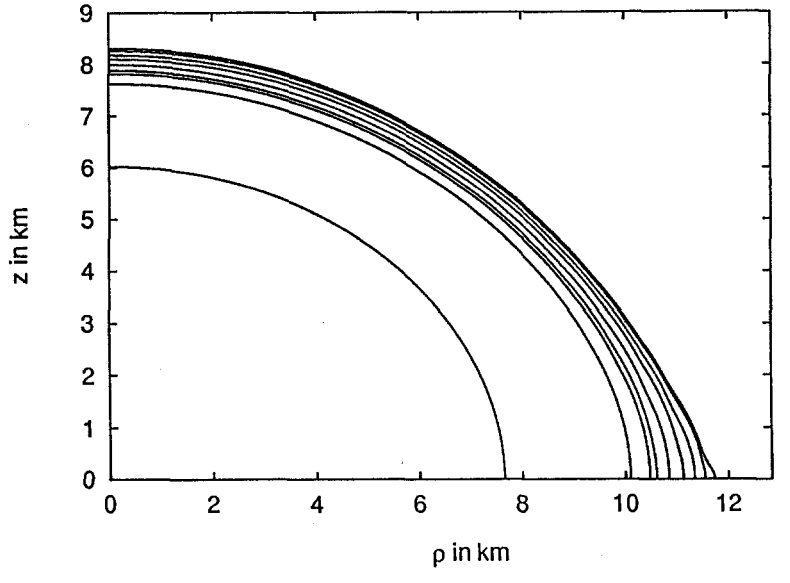

Fig. 13. Embedding diagrams of some internal constant-density surfaces of the fastest star of Fig. 12. The density values are the same as in Fig. 11.

Figure 8 shows the surfaces of the solutions of Table 1, i.e. relatively light neutron star models. Additionally, in Fig. 9 the internal structure of the fastest star is depicted by surfaces of constant density. It can be recognized that the mass shedding at the equator begins with a bump there caused by the outermost layers only. Analogous pictures are shown in Figs. 10 and 11 for the series of Table 2 (mass $M \approx 1.4 M_\odot$) and in Figs. 12 and 13 for the models of Table 3 (highest masses). The behaviour is qualitatively similar, apart from the feature that the density gradient becomes steeper for higher masses. This is of course a general feature of neutron star structure (see Fig. 1) and has nothing to do with rapid rotation.

As second visualization method of the space-time structure of rapidly rotating neutron stars we show some pictures of "how a rotating neutron star does look like". The idea is to assume that from the surface of the considered neutron star photons are emitted which are moving through the curved space-time and eventually reach the observer which is located far away (near infinity, in the asymptotically flat region). Obviously, this visualization method may be considered to be complementary to the embedding pictures of the interior of the star, as the photons propagate in the region outside the star. Practically, we use a ray-tracing approach: from the observer's position the paths of photons are followed back in different directions by integrating the null geodesic equations, using the Christoffel symbols of the numerically determined metric (3), until each photon does (or does not) hit the surface of the rotating star. We call this procedure *4D ray-tracing*, since apart from three-dimensional space also the time plays a role (the time when a photon hits the star determines the position on the surface).

In Figs. 14 and 15 two examples are presented. Both are models from Table 3 with $V_0 = -0.4455$; the first is a relatively slow neutron with angular velocity $\Omega \approx 0.63\ 10^4\,\mathrm{s}^{-1}$ (corresponding Kepler frequency $\Omega_\mathrm{K} \approx 1.94\ 10^4\,\mathrm{s}^{-1}$), the second is the fastest one with $\Omega \approx 1.15\ 10^4\,\mathrm{s}^{-1}$ (near the mass shedding limit: $\Omega_\mathrm{K} \approx 1.19\ 10^4\,\mathrm{s}^{-1}$). For the sake of visibility the surface of the star is painted in a checkerboard pattern (with 30° by 30° patches in the angles θ and φ).

Note that these are not two pictures of the same star with different rotation rates: the number of baryons differs; one star has $M_0 = 1.86 M_\odot$, the other $M_0 = 1.96 M_\odot$. For this reason also the optical images have different sizes; the heavier one ($M = 1.71 M_\odot$) appears bigger than the lighter one ($M = 1.59 M_\odot$) due to the different light deflection. Relativistic light deflection is also responsible for the fact that one can see both poles simultaneously and more than the front hemisphere in equatorial regions (for non-rotating stars, see also Nollert et al. 1989). The effects of rotation are only weakly visible in Fig. 14 (a slight bending of meridional lines), while in Fig. 15 the asymmetric appearance is more spectacular. A closer inspection

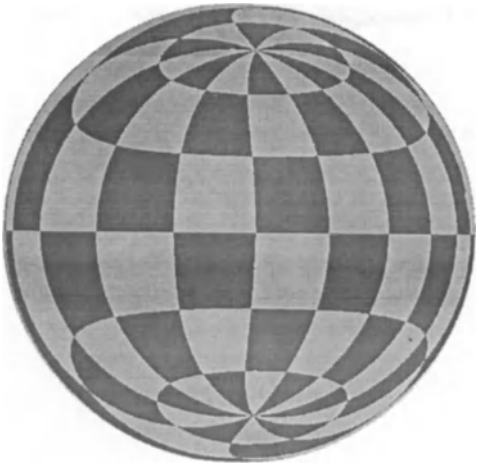

Fig. 14. 4D ray-tracing picture of a slowly rotating neutron star

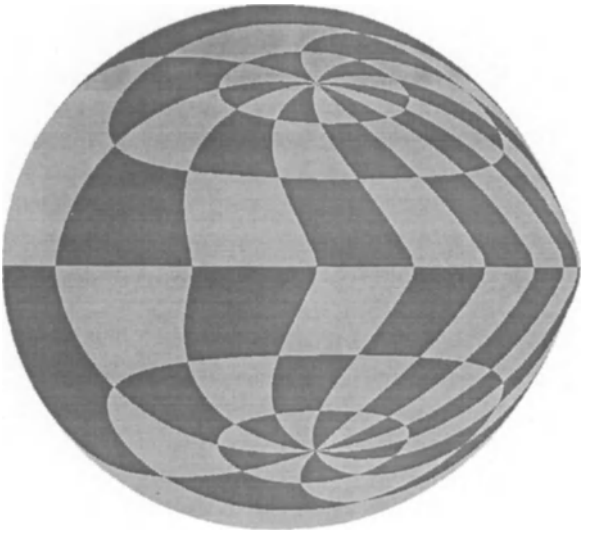

Fig. 15. 4D ray-tracing picture of a fast rotating neutron star near the mass shedding limit

shows that in this picture time-of-flight effects and Lense-Thirring frame dragging of photon paths appear in combination.

References

Arnett, W.D., Bowers, R.L. (1977): Astrophys. J. Suppl. **33** 415

Backer, D.C., Kulkarni, S.R., Heiles, C., Davis, M.M., Goss, W.M. (1982): Nature **300** 615

Baym, G., Pethick, C., Sutherland, P. (1971): Astrophys. J. **170** 299

Butterworth, E.M., Ipser, J.R. (1976): Astrophys. J. **204** 200

Canuto, V., Chitre, S.M. (1974): Phys. Rev. D**9** 1587

Datta, B. (1988): Fund. Cosmic Phys. **12** 151

Friedman, J.L., Ipser, J.R., Parker, L. (1986): Astrophys. J. **304** 115

Friedman, J.L., Ipser, J.R., Parker, L. (1989): Phys. Rev. Lett. **62** 3015

Glendenning, N.K. (1988): Phys. Rev. C **37** 2733

Hartle, J.B. (1967): Astrophys. J. **150** 1005

Hartle, J.B., Thorne, K.S. (1968): Astrophys. J. **153** 807

Neugebauer, G. (1985): in *Proceedings of the Balatonszéplak Relativity Workshop*, ed. by Z. Perjes (KFKI Publ., Central Research Institute for Physics, Budapest) p. 103

Neugebauer, G. (1988): in *Relativity Today, Proceedings of the 2nd Hungarian Relativity Workshop*, ed. by Z. Perjes (World Scientific, Singapore) p. 134

Neugebauer, G., Herlt, E. (1984): Class. Quant. Grav. **1** 695

Neugebauer, G., Herold, H. (1992): contribution in these proceedings (paper I)

Nollert, H.-P., Ruder, H., Herold, H., Kraus, U. (1989): Astron. Astrophys. **208** 153

Shapiro, S.L., Teukolsky, S.A. (1983): *Black Holes, White Dwarfs, and Neutron Stars: The Physics of Compact Objects* (Wiley, New York)

Weber, F., Glendenning, N.K., Weigel, M.K. (1991): Astrophys. J. **373** 579

Weinberg, S. (1972): *Gravitation and Cosmology* (Wiley, New York)

Weise, W., Brown, G.E. (1975): Phys. Lett. B **58** 300

Wu, X., Müther, H., Soffel, M., Herold, H., Ruder, H. (1991): Astron. Astrophys. **246** 411

Zienkiewicz, O.C. (1977): *The Finite Element Method* (McGraw-Hill, London)

A NEW LABORATORY EXPERIMENT FOR TESTING NEWTON'S GRAVITATIONAL LAW

J. Schurr, H. Meyer, H. Piel and H. Walesch

Fachbereich Physik, Bergische Universität Wuppertal, 5600 Wuppertal, Germany

ABSTRACT

We have performed an experiment to test the $1/r^2$ dependence of Newton's law of gravitation and to determine the gravitational constant G, using a novel experimental method. The two mirrors of a Fabry-Perot microwave resonator are suspended as a pair of pendula. The gravitational force of a laboratory test mass alternating between two positions acts on this resonator and changes the distance between the two mirrors. The resulting frequency change is used to determine the gravitational force. The test mass is moved periodically from a reference position to a distance r of the resonator. Analysis of the resulting periodicity in the frequency change allows a strong suppression of the random noise and thermal drift phenomena. This article describes the experimental method and discusses the results of first measurements in the range of distances of about 0.6 m to 3.6 m between the Fabry-Perot resonator and the test mass. In a series of experiments we have investigated the inverse square law and determined the gravitational constant with a relative error of $1.5 \cdot 10^{-3}$. No deviations from Newton's $1/r^2$ law or the CODATA value of G were found.

1. INTRODUCTION

Many aspects of Newton's law of gravitation have been investigated experimentally during the three centuries which passed after its formulation and the precision of these experiments has increased continuously [1-4]. Until today the most precise determination of the gravitational constant G is possible by means of the famous Cavendish torsion balance. It has been highly developed in the last century and G can now be determined with a relative precision of $1 \cdot 10^{-4}$ [5]. It is however remarkable, that none of the experiments which have been performed to obtain a precision value for G have at the same time tested the inverse square law.

Looking at the existing experiments performed to determine G and/or to investigate the inverse square law one observes a few classes of experimental methods characterized by the distance between the interacting masses. This distance is called in the following the "range" of the experiment.

Only a few experiments in the range of a few millimeter up to 10 m have been performed to test the inverse square law [1,6-10]. These experiments which used torsion balances or other types of detectors achieved a relative accuracy in the determination of the gravitational forces of typically a few percent but did not result in values for the gravitational constant.

Many astronomical experiments for distances of the interacting masses of 10^7 to 10^{11} m have been performed [1-3,12]. The motion of natural satellites like planets, asteroids and binary pulsars as well as artificial satellites like LAGEOS have been studied. There trajectories were investigated carefully and these

studies resulted in an excellent confirmation of Newton´s inverse square law. However, no value for the gravitational constant can be deduced from the motions of celestial bodies.

In the intermediate geophysical range of distances between 10^2 m and 10^4 m other types of experiments have been performed. In some experiments the local gravity gradient of the earth was measured using a gravimeter whose distance from the center of gravity of the earth was changed. These experiments were performed by moving appropriate gravimeters inside of mines, in the Greenland ice, in the earths atmosphere and in the ocean. The latter experiments were performed from a tower or a submarine as a work station [1-4,13-16]. In other experiments the gravitational force produced by a mountain or the changing content of a storage lake was studied. In many of these geophysical experiments deviations have been observed which were later identified as systematic errors [13-16]. In general geophysical experiments have achieved a high level of precision in the determination of the gravitational constant (up to 0.2 %) and have also tested the inverse square law.

During the last two decades, theories to unify the four fundamental forces in nature have been developed. Some of these models propose short-range forces that depend on the composition of the interacting matter [2,4,17-21] in contradiction to Einstein's principle of the equivalence of the gravitational and the inertial mass, which can be tested in experiments using torsion balances. The most famous of these experiments was performed by R. v. Eötvös in 1907. Until today no deviation from Einsteins principle has been found on a relative difference level of 10^{-12} [1-3,11].

Since 1986 Newton's gravitational law has been tested in a series of new experiments with different levels of precision [4]. These experiments were stimulated by reports of the observation of possible short-range deviations from Newton's gravitational law. This force (known as "fifth force") is assumed to depend on the Baryon- or Lepton number of the interacting matter and contradicts Einstein's weak equivalence principle. Many of the new experiments have a high sensitivity to detect such a material dependent force but have revealed no evidence for such an effect [1,4].

The fifth-force discussion has nevertheless motivated us to develop a pendulum gravimeter using a Fabry-Perot resonator [22,23]. The gravimeter was designed to measure the gravitational acceleration of a test mass as a function of distance and from this to determine the gravitational constant G. A possible short-range force can be investigated without assuming an explicit dependence on the material in use. Systematic influences will in general result in deviations from the inverse square law. Their dependence on the distance of the interacting masses is a useful information to identify and to eliminate the source of these effects. The very sensitive check of systematic influences allows to enhance the precision of the measurement of the gravitational force and of the determination of the gravitational constant.

The main objective of this article is to introduce our experimental method and to present the first results. In section 2 the basic principle of the Fabry-Perot gravimeter is explained. In section 3 the gravimeter and its main part, the microwave Fabry-Perot resonator, is described in detail. The procedure used to obtain a high precision measurement of the resonators resonant frequency is explained in section 4 and the experimental results concerning the stability of the resonant frequency of the Fabry-Perot resonator are presented. The presently dominating systematic effects due to external forces are outlined in section 4. In section 5 we describe the modulation technique applied to measure the gravitational force. Section 6 gives the results for the gravitational force as a function of distance in the range of 0.6 m to 3.6 m and our present result for the gravitational constant.

2. THE BASIC PRINCIPLE OF THE EXPERIMENTAL METHOD

The central part of the gravimeter consists of two Fabry-Perot mirrors suspended as a pair of pendula from a suspension platform (Fig. 1). The pendula have a length ℓ of approximately 3 m and the distance b between the two mirrors is about 20 cm. The gravitational acceleration of a test mass M displaces both pendula of approximately equal masses m_1 and m_2.The pendulum next to the test mass is displaced by a larger amount than the other pendulum. The test mass is brought to a distance r to the Fabry-Perot resonator and removed from it to a reference position in a periodic motion (r is the distance between the centers of gravity of the test mass and the closest mirror). It rests in each position for 15 min, a time which is much longer than the time constant of the pendula. The motion of the pendula is damped by eddy current brakes and they reach their new equilibrium position in a few seconds. The two Fabry-Perot mirrors form a microwave resonator with a resonant frequency f of approximately 20 GHz. The gravitational force from the moving test mass results in a change Δb of the distance between the two mirrors of the resonator and therefore changes its resonant frequency by Δf. A typical value for the change of the mirror separation due to the gravitational force is 20 nm.

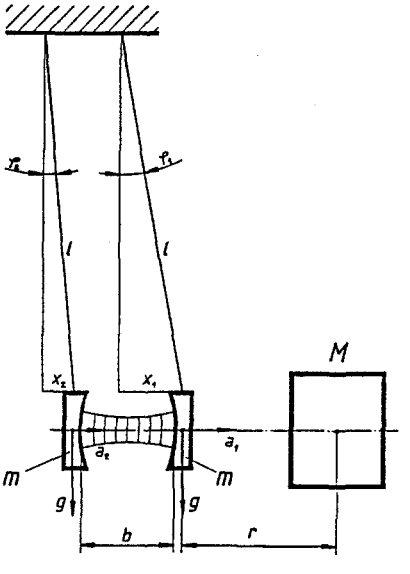

Fig.1 The principle of the Fabry-Perot gravimeter.

The deflection angles of the pendula (typically 10^{-8} rad) and the relative change of the separation of the mirrors ($\Delta b/b$ of typically 10^{-7}) are both very small and therefore Δf and Δb are proportional as described in more detail in section 3.1. Furthermore, the displacement of each mirror can be approximated as a horizontal translation proportional to the gravitational acceleration of the test mass. The quasi static change in the mirror separation Δb is therefore directly proportional to the difference Δa of the accelerations of the two pendula (a_1 and a_2)

$$\Delta f = \frac{df}{db}\Delta b = \frac{df}{db}\omega_o^{-2}(a_1 - a_2) = \frac{df}{db}\omega_o^{-2}\Delta a \qquad (1)$$

with ω_0 being the eigenfrequency of the pendula. The distance to frequency conversion factor of the Fabry-Perot resonator df/db is given in section 3.1. Equation (1) shows the relation between the quantities of interest, the gravitational acceleration Δa, and the measured frequency shift Δf.

The horizontal gravitational acceleration of each pendulum is calculated by a numerical integration of Newton's inverse square law over the mass distributions of the test mass (M) and the resonator (m_1, m_2):

$$m_i \cdot \vec{a}_i = \iint_{m_i M} G \cdot dm_i \cdot dM \cdot \frac{\vec{\rho}}{\rho^3} \qquad \text{for } i=1,2 \qquad (2)$$

ρ is the distance between the mass elements. Assuming a symmetric mass distribution and a homogeneous density, the 6-fold integral (2) can be transformed to a 3-fold integral, which can be evaluated easily using a Gaussian integration formula.

The combination of equation (1) and (2) results in our basic relation:

$$\Delta f(r) = \frac{df}{db} \cdot \omega_0^{-2} \cdot G \cdot M \cdot \left(\left[\frac{1}{r^2} - \frac{1}{(r+b)^2} \right] \cdot K(r) - \left[\frac{1}{r_{Ref}^2} - \frac{1}{(r_{Ref}+b)^2} \right] \cdot K(r_{Ref}) \right) \qquad (3)$$

Δf is the measured frequency shift of the Fabry-Perot resonator obtained by moving the test mass from a position r to a reference position r_{Ref}, and vice versa. The terms in the angular brackets correspond to the difference of the gravitational accelerations of each pendulum. The function K(r) takes the finite dimensions of the masses into account (K = 1 corresponds to point masses or to masses with infinite distance). The determination of the gravitational constant G from equation (3) requires the knowledge of the sensitivity of the Fabry-Perot resonator df/db, the eigenfrequency of the pendula ω_0, the mass of the test mass M, the distances r, r_{Ref} and b, and the mass distribution function K(r).

At first sight one might be tempted to suspend only one mirror as a pendulum and to fix the other one. A larger shift of the mirror distance created by the gravitational force of the test mass would be obtained. In this case the frequency shift Δf due to the gravitational force varies like $1/r^2$. However, the distance between the mirrors would be affected by tidal forces and by disturbing effects like microseismic vibrations and other movements and tilts of the ground. Furthermore, a change of the mass distribution in the far surroundings of the gravimeter would have a significant influence.

In order to cancel out these disturbing effects, both mirrors are suspended as pendula of equal length. The strong suppression of disturbances is much more important than the decrease of the gravitational effect. In this configuration the change in the mirror separation from gravitational forces varies approximately like $1/r^3$ and only the immediate surrounding of the gravimeter has to be controlled.

3. THE GRAVIMETER

We have started our experiments to test Newton's law with measurements of the gravitational force of a test mass having a distance of about 10 cm between its center of mass and the closest mirror of the Fabry-Perot resonator [22,23]. This prototype set-up was used to get an idea of the achievable resolution and accuracy and to develop solutions for various foreseen or unforeseen problems. Based on the experience

with this pilot experiment an improved experimental set-up was designed. It is used for measuring the gravitational force exerted by a test mass in a distance of 0.6 to 3.6 m.

Fig.2 shows the schematic arrangement of the experimental set-up. Its main part, the two Fabry-Perot mirrors suspended as pendula, is placed inside of a vacuum tank. This tank is mounted into a supporting steel construction called "the tower". The test mass is positioned on a guide rail outside of the tower and is aligned to the same height as the resonator. In the following paragraphs the individual components of this set-up are discussed.

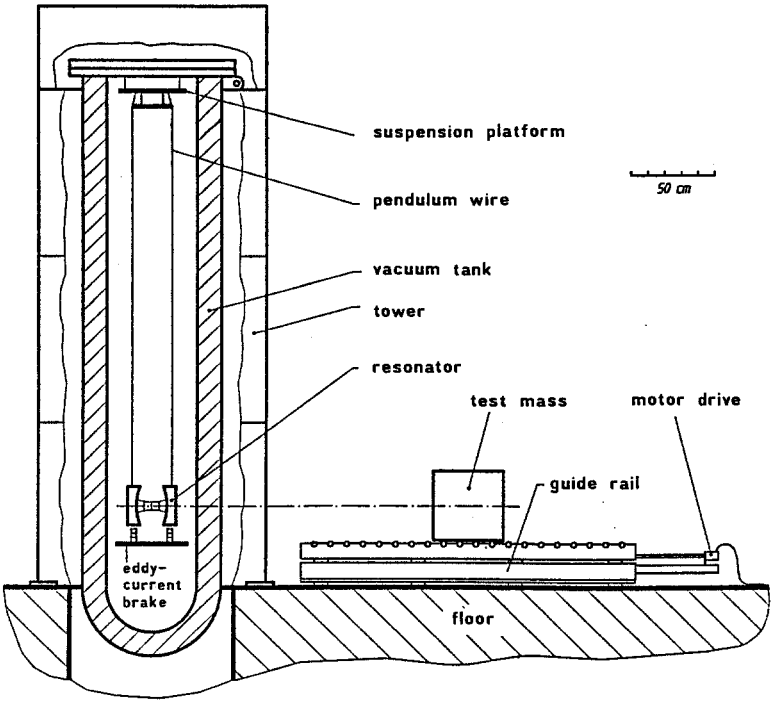

Fig.2 Schematic arrangement of the experimental set-up.

3.1 THE FABRY-PEROT MICROWAVE RESONATOR

The Fabry-Perot resonator (Fig.3) consists of two spherical mirrors separated by a distance b of 241 mm. The radius of curvature R of the circular mirrors is 580 mm, their diameter is 192 mm and their thickness at the center is 9 mm. They are fabricated from OFHC copper and the roughness of their diamond machined surfaces is 50 nm. For this combination of mirror separation and radius of curvature stable electro-magnetic modes exist in the Fabry-Perot resonator [24, 25].

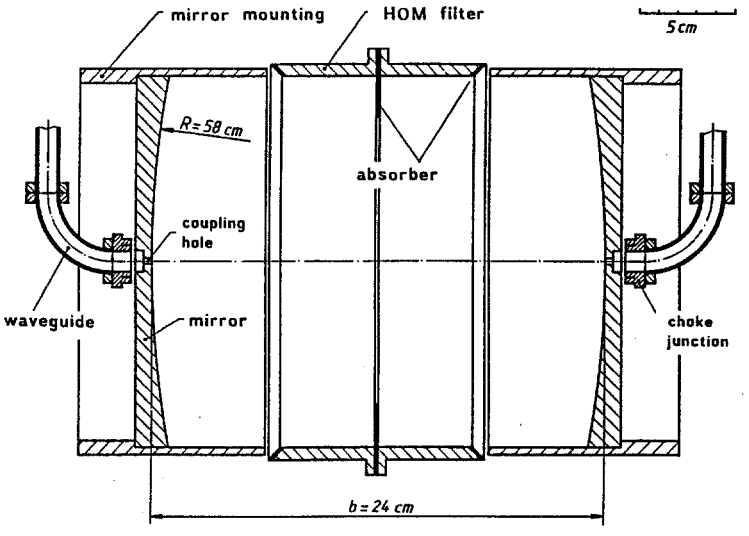

Fig.3 The Fabry-Perot microwave resonator.

The Fabry-Perot resonator is operated at microwave frequencies in the range of 20 GHz to 26 GHz. The field distribution of the resonant modes can be analytically computed with sufficient accuracy using the so called "complex-source-point theory" [26,27]. In first approximation the modes are transverse electromagnetic (TEM_{pmq}). The field distribution is a standing wave pattern with q being the number of half-wavelengths in axial direction. The number of half-wavelengths in radial and azimuthal direction is determined by p and m. In the case of the "fundamental modes" TEM_{00q}, which are preferred in most applications, the field distribution is a Gaussian beam with waist radius $w_0 \approx 3$ cm and length b. The beam axis aligns itself along a line connecting the centers of curvature of both mirrors. Due to the focussing effect of concave mirrors the field distribution of a Fabry-Perot resonator is stable, even if one mirror is tilted.

The resonant frequencies f_{pmq} are determined by the boundary conditions of the electromagnetic fields at the mirror surfaces [24-27] and are given by

$$f_{pmq} = \frac{c}{2b} \cdot \left[q + \frac{1}{\pi} \cdot n \cdot a\cos(1 - b/R) + \text{higher order terms} \right] \qquad (4)$$

The transverse order n is given by: $n = (2\,p + m + 1)$. This formula is accurate except for terms of the order $(2\pi \cdot f \cdot w_0 / c)^{-6}$ which amount to only 1 part in 10^7 of the frequency.

The first term in equation (4) gives the resonance frequency of an optical resonator with plane-parallel mirrors. In this approximation the distance between the mirrors is a multiple of half a wavelength - usually called axial order q. The second term takes into account the mirror curvature and depends on the transverse order n of the mode. This term is much smaller than the first one. The terms labelled 'higher order' include a correction due to the finite wavelength (calculated by means of the complex-source-point theory) and they also include the frequency shift caused by the small holes in the mirrors through which the microwaves are coupled into the resonator.

We have obtained very good agreement between the computed and the measured frequencies of the Fabry-Perot modes with a relative accuracy of $5 \cdot 10^{-6}$. This justifies the application of equation (4) to convert a measured frequency shift df into a change in distance db:

$$\frac{df}{db} = -\frac{f}{b} \cdot \left[1 - \frac{nc}{2\pi f} \cdot \left(2Rb - b^2 \right)^{-1/2} + \text{higher order terms} \right] \qquad (5)$$

This equation can be applied if the Fabry-Perot mirrors move in the direction of the Gaussian beam inside the resonator, for example due to the gravitational force of the test mass. Equation (5) now gives the distance to frequency conversion factor for the gravimeter which was introduced in equation (1) in section 2. A typical value for df/db is about 100 Hz/nm.

At the beginning of every gravitation experiment b and df/db are determined using the equations (4) and (5). We estimate the relative uncertainty in the determination of the two quantities to be $1.4 \cdot 10^{-5}$ and therefore no special calibration of frequency versus distance is necessary.

3.2 THE QUALITY FACTOR AND THE FREQUENCY RESOLUTION

The quality factor Q of a resonator is defined as 2π times the stored energy divided by the energy loss per cycle. For small energy losses Q is equal to the ratio of the resonant frequency to the full-width at half maximum (FWHM) of the resonance curve. If coupling, diffraction and scattering losses are sufficiently small, the Q of an open resonator is limited by reflection losses of the mirrors. In this case Q is independent of the transverse order of a mode excited in the resonator and the sensitivity limit δb in terms of the smallest detectable frequency shift δf is at its maximum:

$$\delta b = \left(\frac{df}{db} \right)^{-1} \delta f \equiv -\frac{\delta f}{FWHM} \cdot \frac{b}{Q} \qquad (6)$$

The sensitivity δb is to a good approximation not explicitly dependent on the resonant frequency and proportional to the ratio b/Q [22,23]. A shift of one mirror with respect to the other one by a distance of b/Q results in a frequency shift which is equal to the FWHM of the resonance curve. The experimentally obtained Q-value of 210 000 is dominated by reflection losses and the resulting value for b/Q is approximately 1 μm.

The experimental determination of the resonant frequency and the Q of the Fabry-Perot resonator is discussed in detail in section 4. In order to outline the principle a short description is given in the following. Waveguides are used to feed the microwave power from a synthesized sweep generator with a stability of 4 Hz to the resonator. A certain fraction of this power is coupled into the resonator and excites the desired mode. The power coupled out of the resonator is transmitted by a waveguide to a microwave detector diode, i.e. the Fabry-Perot resonator is operated in transmission. The coupling of the waveguide system to the resonator is achieved by means of small coupling holes to avoid a reduction of the Q-value due to scattering losses at these disturbances of the mirror surface. These small coupling holes lead to a very weak coupling with a typical transmission coefficient of about 10^{-4}. The sweeping of the generators output

frequency and the registration of the signal of the microwave detection diode are performed by computer control and the resonance curve of the Fabry-Perot resonator is obtained. From this resonance curve the resonance frequency and the FWHM is deduced. Only a single mode is excited in the resonator and all other modes are clearly separated in frequency by about thousand times the halfwidth of a resonance.

The relative frequency resolution $\delta f/FWHM$ in equation (6) is a quantity which is specific to the data analysis and is determined by the precision with which the shape of the resonance curve is measured. As will be shown later a relative resolution of $4 \cdot 10^{-6}$ is achieved experimentally which then corresponds to a precision in the determination of the change of the mirrors distance of $\delta b = 4 \cdot 10^{-12}$ m. A Fabry-Perot resonator is therefore well suited for measuring small displacements. The resolution is high enough to measure the mirror displacement of 12 nm due to the gravitational force of our test mass with a relative accuracy of about $4 \cdot 10^{-4}$ as a function of distance and to determine the gravitational constant with an accuracy of about $1 \cdot 10^{-4}$. At this moment it is worthwhile to mention that the absolute Q value of a Fabry-Perot resonator can be improved possibly by the use of superconducting mirrors. This would then in turn lead to an even higher accuracy in the determination of G.

3.3 SOME MICROWAVE ASPECTS OF THE FABRY-PEROT RESONATOR

In order to secure the proper functioning of the Fabry-Perot resonator as a gravimeter it has to be decoupled electrically and mechanically from its surrounding as perfectly as possible. The mechanical arrangement of the waveguide coupling to the resonator and the electrical coupling of the resonator to the surrounding metallic enclosure of the vacuum tank therefore needs special attention.

The two mirrors are mounted into a cylindrical copper shield in order to minimize the amount of microwave radiation leaving or entering the Fabry-Perot resonator (Fig. 3). These cylinders act at the same time as a mechanical mount of the mirrors, as a microwave shield and as an eddy-current brake (section 3.5). The mirrors and their mountings are designed to result in a simple and rotationally symmetric mass distribution which can be integrated with high precision (equation (2)).

An additional cylindrical shield is arranged between both mirrors. It reflects or absorbs microwave radiation which would otherwise be radiated from the surrounding into the resonator or vice versa. For the same reason the edge of the cylinder is coated with microwave absorber. An absorbing ring is mounted into the cylinder (Fig. 3) in order to damp selectively spurious higher order modes in the Fabry-Perot resonator. (The inner diameter of this ring is large enough to avoid a reduction of the quality factor of the main mode). Thus, the "higher order mode" filter (in the following abbreviated as HOM filter) results in a strong electromagnetic decoupling between the desired field distribution inside the resonator and undesired fields in its surrounding. For the same reason the resonator is surrounded by absorber plates. As will be described below the waveguides coupling the Fabry-Perot to the microwave drive and detection system couple also to the space surrounding the resonator. Any field exited in this space will therefore form a background signal underneath the Lorentzian shaped resonance curve. This background is a priori influenced in an unpredictable way by the movements of the two pendula (section 4.3).

The mechanical behaviour of the pendula would be strongly affected by a physical contact between the resonator and the waveguides. Therefore the waveguides have to be mechanically decoupled from the resonator. The gap between the waveguide and the back sides of the Fabry-Perot mirrors should be narrow

to avoid a strong coupling of the waveguides to the space outside of the Fabry-Perot. To achieve this the waveguides are connected to the Fabry-Perot by a choke-junction [28]. Such junctions are generally used to couple rotating radar antennas to their drive systems. The waveguides are carefully aligned by an externally adjustable positioning system. A continuing waveguide is machined into the back of each mirror and has a length of a quarter-wavelength (Fig.3). Due to this arrangement, the electromagnetic field in the gap between the waveguide and the back of the mirror is small and so are the radiation and reflection losses. Furthermore, the power transmitted through the resonator is nearly independent of the relative position of the mirrors and the waveguides and therefore insensitive to vibrations and to tilt. Typical dynamical displacements during the experiment are of the order of 1 μm and result in a relative change of transmitted power of less than 10^{-4}. In addition the determination of the resonant frequency is essentially independent of the transmitted power.

3.4 THE SUSPENSION OF THE FABRY-PEROT MIRRORS

Both mirrors and the HOM filter are suspended each in two loops of tungsten wire (Fig.4). The diameter of the wire is 0.2 mm. Tungsten is chosen because of its beneficial magnetic, elastic and thermal properties. The bodies of the pendula (Fig.3) are dimensioned such that the center of mass is located at the mirror surface. The center of mass is positioned halfway between the supporting wire loops. Therefore, in first approximation, the thermal expansion of the copper due to a change in temperature does not result in an change of the effective distance between the two Fabry-Perot mirrors and the resonance frequency is insensitive to a drift of the resonators temperature.

The tungsten wires are mounted to a special suspension platform (Fig. 4) and can be adjusted in order to properly align the resonator and the HOM filter. The horizontal separation of the wires (and thus the mirror separation) is held constant by means of a quarz plate. The length ℓ measured from this quarz spacer to the center of mass of the mirrors is about 2.62 m. Quarz is used because of its small thermal expansion coefficient in order to reduce the thermal drift of the resonance frequency of the Fabry-Perot resonator. The change in mirror separation due to thermal expansion of the quarz plate is about 60 nm/K and requires a good thermal stability of the suspension platform.

3.5 THE EDDY-CURRENT BRAKE

The system of the two pendula acts as a mechanical band-pass filter. Ground oscillations with a frequency much lower than the eigenfrequency of the pendula (0.308 Hz) move both pendula in the same way and the mirror separation b is affected negligible. The transmission of high frequency vibrations of the ground (for example vibrations caused by industrial activity, by traffic or by machines within our institute) is strongly suppressed due to the inertial mass of the pendula. Only ground vibrations in a frequency band centered at the resonant frequency of the pendula are transmitted and increased to a level which is determined by the damping of the pendula.

Ground oscillation in this frequency band (around 0.3 Hz) are dominated by microseismic noise which originates from waves caused by the surges of the Mediterranean Sea and the Atlantic Ocean [29]. These

waves propagate through the whole continent without strong damping. A typical time variation of seismic activity during the course of the year is obtained which is correlated with the intensity of cyclones at the ocean. The intensity of microseismic vibrations varies only weakly between different locations.

To damp the pendulum oscillations excited by microseismic noise we use eddy-current brakes (Fig.4).

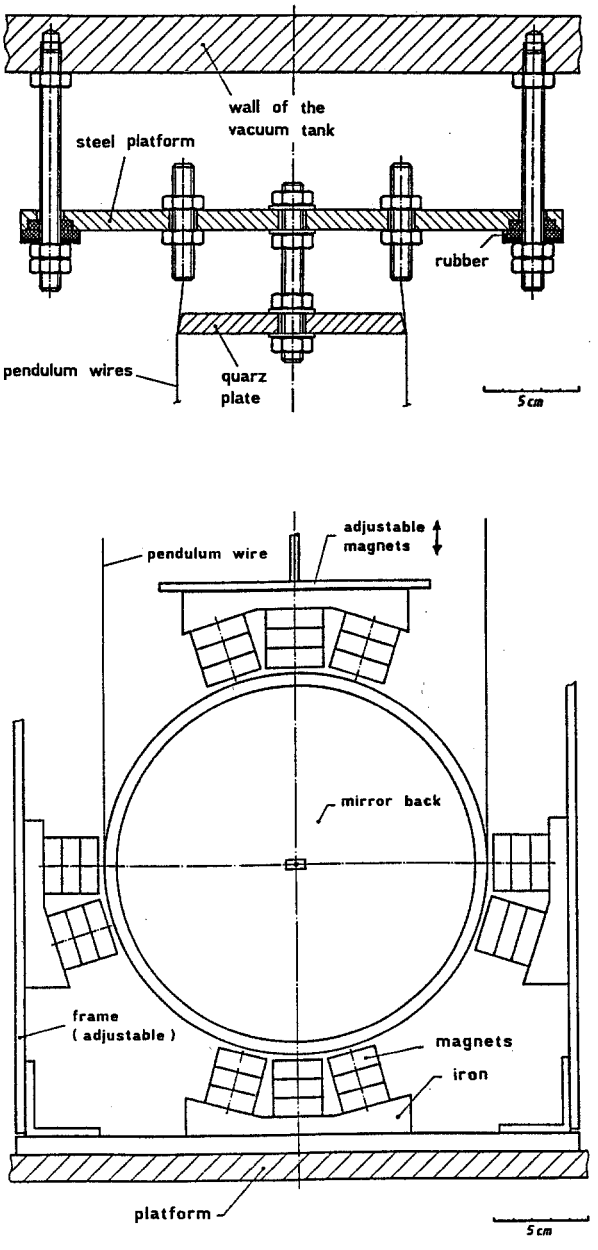

Fig.4 The suspension of the pendula and the eddy-current brakes.

The brakes ensure a damping of the oscillations of the pendula without mechanical contact and the damping force is strictly proportional to the velocity of the pendula. The eddy-current brakes consist of iron plates supporting an array of permanent magnets with alternating polarization (chess-board pattern). They are arranged around the cylindrical enclosures of the two mirrors and ensure a strong damping. The time constant of the damping is about 2 s. The pendula move with nearly the same phase and an amplitude that is inversely proportional to the eddy-current damping. One of the arrays with permanent magnets can be adjusted in situ in order to synchronize the oscillations of the two pendula and thereby to minimize the oscillations of their separation. The differential amplitude is found to be about 1/2000 times the amplitude of an individual pendulum.

3.6 THE VACUUM TANK AND THE TOWER

The gravimeter is placed inside a vacuum tank (Fig.2). Dielectric effects and convection in the residual gas which disturb the resonant frequency of the Fabry-Perot resonator can be avoided if the vacuum pressure variations can be kept below 0.01 Pa. To avoid gas pressure forces [23] without going to ultra high vacuum the experiment is operated above the threshold of this effect, at a pressure of 2 Pa. Good thermal insulation is required to keep thermal expansion effects small. To achieve this we use a vacuum tank with an additional vacuum insulation and a superinsulation shield. This vacuum tank which acts as a thermostat was in fact used in an earlier experiment as a large cryostat for liquid helium.

As the test mass is moved horizontally over the laboratory floor the Fabry-Perot resonator has to be placed at the same height as the center of the test mass. To achieve this the vacuum tank which encloses the gravimeter and which has a weight of 2 t has to be supported in a sturdy steel construction called "the tower". The tower is built from strong steel girders to obtain such a rigidity that the lowest eigenfrequencies of the mechanical vibrations of the tower are higher than the eigenfrequencies of all gravimeter modes which contribute significantly to the vibrations of the mirror separation. This is in fact achieved and no coupled or excessively increased oscillations of the pendula are observed. The transmission of high frequency oscillations of the tower to the pendula is well below the detection limit.

The outer surface of the tower is shielded with aluminum plates against thermal radiation. All components in the close surroundings of the experiment which produce heat, that is the vacuum pumps, the microwave source, the microwave amplifier and the motor drive of the test mass, are installed outside of the tower and are additionally cooled or shielded. The only heat generated inside the thermostat originates from the microwave losses in the waveguides and in the resonator itself. These losses are low and constant in time.

3.7 THE TEST MASS AND ITS POSITIONING SYSTEM

The test mass is a cylinder with a diameter of 440 mm and a length of 430 mm (Fig. 5). The mass distribution of this body can be easily integrated (equation (2)). Its dimensions were chosen in a way that the gravitational force between test mass and resonator is nearly the same as the gravitational force of point masses positioned at the centers of gravity (the correction function K(r) which was introduced in equation (3) assumes values between 0.98 and 1.0001). The mass is 575.8 kg (Table 1b).

Because of its very small magnetic susceptibility and its good mechanical properties, we have chosen a special brass for the material of the test mass (an alloy of 90% copper and 10% zinc with a magnetic susceptibility of $4 \cdot 10^{-5}$). Magnetic forces between the test mass and the resonator could result in a systematic error and have to be avoided. The magnetic field at the position of the test mass is only about 0.05 mT. It is given by the magnetic field of the earth distorted by the steel tower and the guide rails. In contrast the magnetic field of the eddy-current brake at the location of the resonator is strong (about 0.3 T). This causes a magnetic force due to the interaction of the induced magnetic dipole moments of test mass and resonator. Thus, only materials are allowed to be used with a low diamagnetic or paramagnetic susceptibility and a small percentage of ferromagnetic impurities. We have chosen copper for the resonator, tungsten for the pendulum wires and brass for the test mass. Because of this choice the magnetic forces between the masses were found to be below the detection level.

The mass of the test mass was determined according to the following procedure: During the fabrication of the testmass a part of the material was used to precisely machine small samples to determine the density ρ of the test mass and its uniformity. With the obtained density ($\rho = 8.8066 \pm 0.0004 \pm 0.0017$) g/cm^3 and with the dimensions of the test mass its mass M was found to be ($575.80 \pm 0.03 \pm 0.11$) kg.

The test mass rests on a special guide rail and glides on rollers which are rotating on ball bearings around axels fixed to this rail (Fig.5). The test mass can be positioned precisely by means of a spindle. A motor drive rotates the spindle in steps of 50 μm. The test mass and the spindle are connected by means of a fitting counterpart. Only the small mass of this counterpart has to be added to the mass distribution of the test mass (equation (2)). All other movable parts perform a strictly rotational motion.

The error in the determination of the distance between test mass and resonator becomes non critical if it is about 100 μm at $r = 0.6$ m and 2 cm at $r = 3.6$ m. The actual systematic positioning error is less than 80 μm per meter of covered distance relative to the closest position for ($r = r_1$). This distance of nearest approach between the Fabry-Perot resonator and the test mass is determined with a statistical accuracy of 20 μm and with systematical positioning errors of about 100 μm which are due to the inaccuracy of the instruments used for measuring the respective distances. These errors can be reduced by using calibrated standards. Each position can be reproduced within a statistical error of less than 20 μm (a summary of the positioning errors is compiled in Table 1).

4. THE MEASUREMENT OF THE RESONANT FREQUENCY

The measurement of the resonant frequency of the Fabry-Perot resonator as a function of the position of the test mass is the basic experimental task. In this context it is very important to ensure that the change in frequency which is registered is caused only by the gravitational force of the test mass in its different positions relative to the gravimeter. Even if the measurement of the resonance frequency is affected by various systematic errors, the measurement of the frequency shift due to the gravitational force of the test mass is not influenced as long as these effects are independent of the position of the test mass.

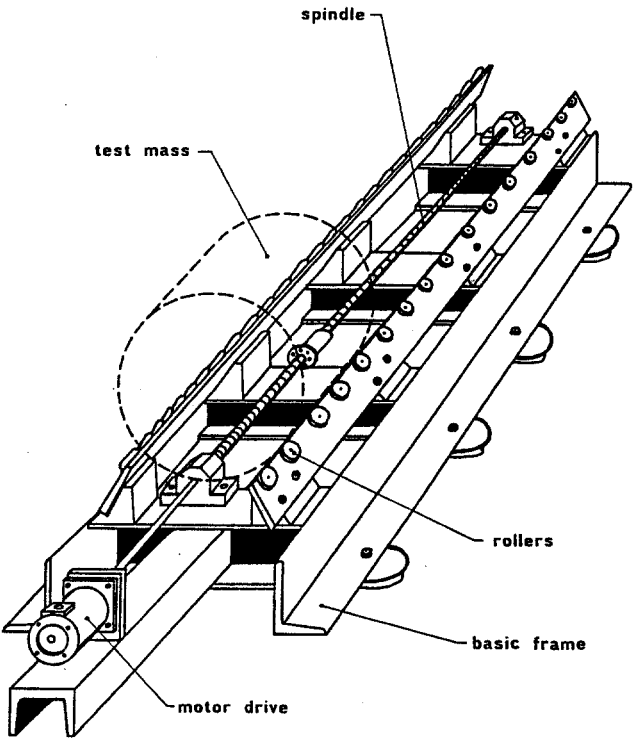

Fig.5 Test mass with positioning system.

4.1 EXPERIMENTAL PROCEDURE

Experimentally the resonance frequency of the Fabry-Perot resonator is determined as follows [23]. The microwave source used in this experiment is a synthesized sweep generator (Hewlett-Packard Model 8340 B) with a frequency stability of 4 Hz. The microwave power of this oscillator is amplified by means of an amplifier with a low noise figure. The level-controlled microwave power of this amplifier (Avantek, AMT 26136) is then transmitted through the Fabry-Perot resonator. A crystal detector transforms the transmitted power into a voltage signal which is proportional to the microwave power. To measure the detector signal (typical value: 5 mV) with a relative accuracy of about 10^{-5} the high frequency noise of the oscillator and of the detector are both reduced by a low pass filter. The filtered output signal of the diode is measured by means of a digital voltmeter (Hewlett-Packard Model 3457A) which integrates the voltage during one power line cycle (20 ms) to further suppress noise. A computer scans in a digital sweep the generators frequency across the resonant frequency of the chosen Fabry-Perot mode (for most measurements TEM_{0035}). The filtered detector signal is read from the voltmeter and the resonant frequency is calculated by a least-squares fit to the Lorentzian-shaped resonance curve (Fig.6). Due to this procedure the measured

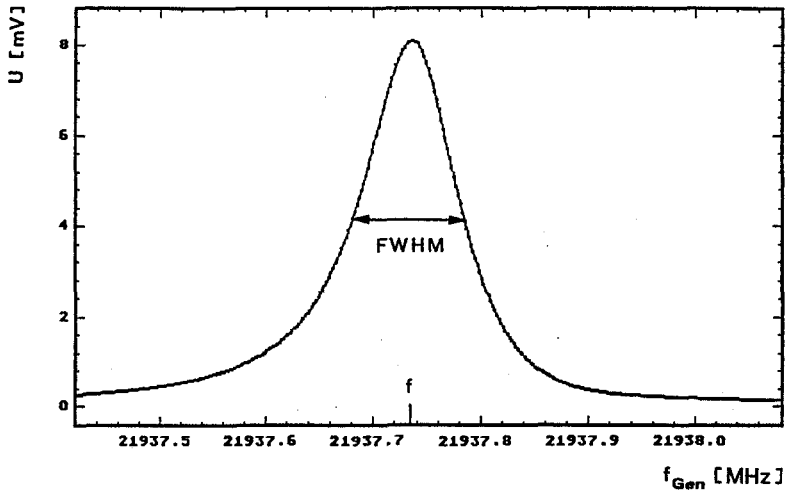

Fig.6 The detector voltage U which is proportional to the microwave power transmitted through the Fabry-Perot resonator as a function of the drive frequency f_{Gen}.

resonant frequency is independent of the microwave power as long as it does not change with a frequency which is close to the scanning frequency (or its harmonics). The digital sweep is repeated in equidistant time steps of 450 ms (with a time jitter of 0.15 ms) and results in a time series of resonant frequencies which can be further analysed and processed. Fig.7 shows the resonant frequency and its fluctuation as a function of time.

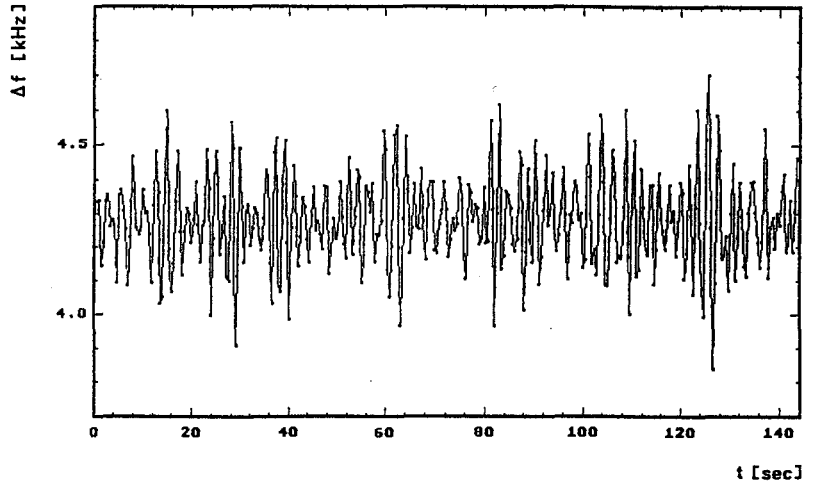

Fig.7 Fluctuation of the measured resonant frequency f of the Fabry-Perot resonator as a function of time.

4.2 FLUCTUATIONS OF THE RESONANT FREQUENCY

To a first approximation, changes of the resonator frequency can be described as a superposition of harmonic oscillations. Fig. 8 gives a typical example of a power density spectrum of the resonant frequency of the Fabry-Perot resonator. Oscillations of the distance between the pendula are driven by microseismic vibrations of the ground and are most pronounced around the resonant frequency of the pendula (in a frequency band from 0.15 Hz to 0.4 Hz). The structure of this resonance can be explained well by a convolution of the frequency response function of the pendula and the power spectrum of ground oscillations [30]. The torsion and the seesaw mode of the two suspended mirrors are not excited with any detectable amplitude. During periods of low microseismic activity, mainly from March to September, typical amplitudes of the oscillations of b are 0.5 nm, while during microseismic storms in the other half of the year, the amplitude is larger by a factor of 4 to 10 (see also [29]).

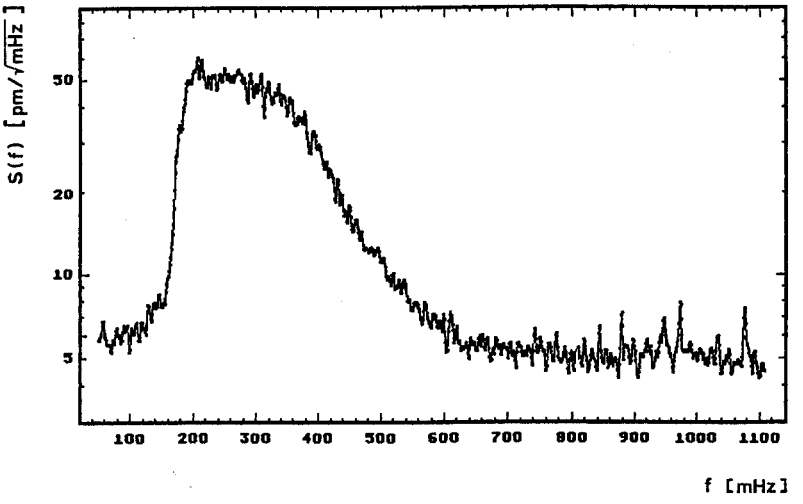

Fig.8 The power spectrum of the fluctuations of the distance b between the Fabry-Perot mirrors at low frequencies (with full eddy-current damping).

The average frequency ω_0 of the two pendula is determined after reducing the eddy current damping coefficient by about a factor of 100. Then a pronounced and sharp resonance is observed in the spectrum of the pendulum fluctuations (Fig.9). ω_0 is obtained from this measurement as 2π (0.3078006 ± 0.0000044) Hz. The fact that only one sharp resonance peak is observed indicates that the frequencies of the pendula are equal within a relative error of $2 \cdot 10^{-4}$.

The noise in the frequency range from 5 mHz up to 0.1 Hz is white and has an amplitude of presently about 5 pm/mHz$^{1/2}$ (Fig.8). This white noise is caused by electronic noise, detector noise and by fluctuations of the microwave power and of the background signal. To determine the frequency shift produced by the test mass in its different positions, the numerical values of the resonant frequency (Fig.7) are averaged by a digital low-pass filter with a time constant of 60 s which reduces white noise and microseismic oscillations with a frequency above 16 mHz. The filtered data are stored for further analysis.

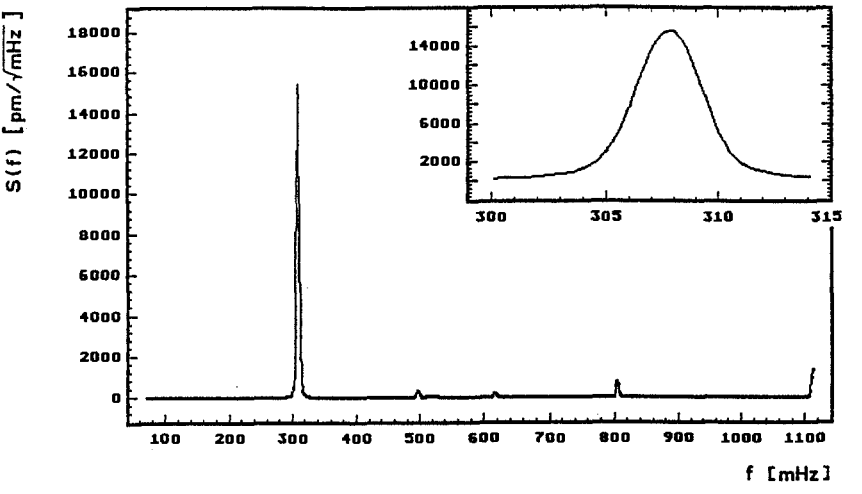

Fig.9 The power spectrum of the fluctuations of the distance between the Fabry-Perot mirrors at low frequencies (with reduced damping).

The remaining microseismic variations of the distance between the mirrors are less than 1 pm and the remaining white noise amplitude is 0.1 nm.

The noise with frequencies below 5 mHz is flicker noise, i.e. the noise amplitude is inverse proportional to its frequency (Fig. 10). The lowest flicker noise frequency which appears in a power spectrum is the inverse of the duration of the measurement. Some unimportant sources of flicker noise have been identified, but until today the dominating source remains unknown. This flicker noise is neither rejected by the digital filter nor by the least squares fit. It is the dominating source for the statistical error in the measurement of the gravitational force. This noise and the low frequency white noise is shown in Fig.11 as a function of time.

The long-term stability of the resonator is given by the drift rate of the Fabry-Perot frequency. Carrying out all provisions of thermal isolation and power control, we achieved a drift rate in the mirror separation of 0.2 nm/h to 0.5 nm/h. A period of 1 day and a long-term effect are observed. The drift is constant in time within a few hours. The thermal stability is better than 10 mK/h.

4.3 SYSTEMATIC PHENOMENA INFLUENCING THE RESONANT FREQUENCY OF THE FABRY-PEROT RESONATOR

As described in section 3.3 there exists a weak coupling between the fields in the Fabry-Perot resonator as well as the traveling waves in the two waveguides and the very small field exited in the metallic surrounding (vacuum tank) of the gravimeter. This results in a background signal at the detector diode and a slightly deformed shape of the resonance curve. The influence of this background signal on the resonance

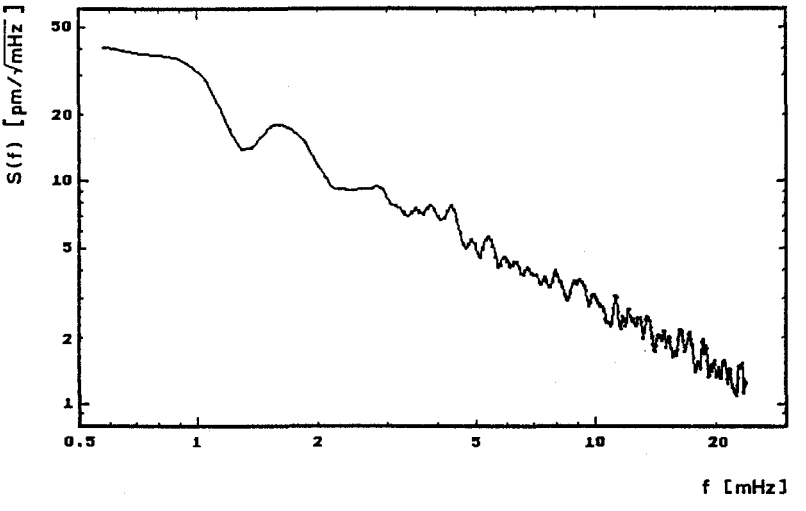

Fig.10 The power spectrum of the flicker noise at very low frequencies (after subtracting the white noise).

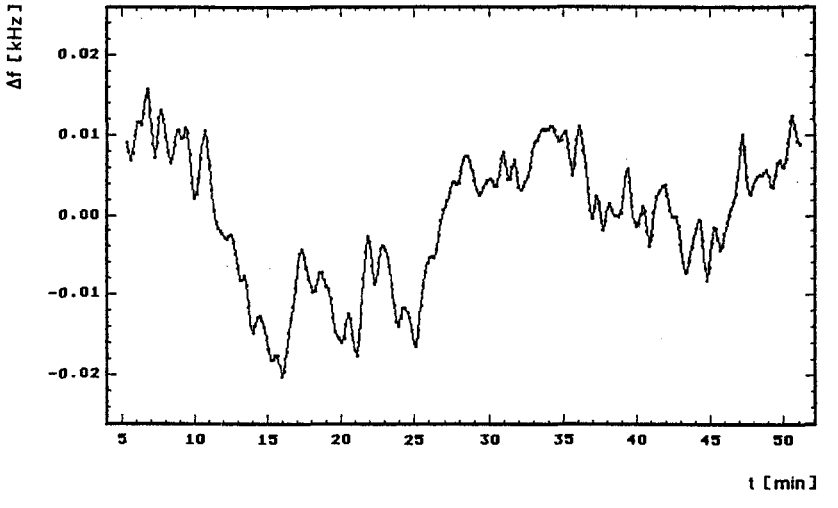

Fig.11 Low frequency fluctuations of the filtered time series of the resonant frequency f versus time.

frequency would be negligible if the ratio of this disturbance to the maximum of the resonance curve would be less than $4 \cdot 10^{-6}$. By means of the choke-junction, the copper shield around the mirrors and the HOM filter a ratio of background to transmitted signal of about 10^{-3} is achieved. This is indeed much higher than the required value of $4 \cdot 10^{-6}$. The background signal is however nearly constant in time and independent of the position of the test mass. The measurement of the shift of the resonance frequency in a gravitational experiment is influenced only on a level of $2 \cdot 10^{-7}$ times the FWHM. This corresponds to a relative systematic error in the measurement of the gravitational force of about $1 \cdot 10^{-5}$.

Another type of systematic error is due to the aging rate and the noise of the quarz oscillator used to lock the microwave source. Again, the measurement of a frequency shift is not affected which has been shown experimentally by using a rubidium frequency standard.

The microwave power necessary to measure the frequency shift of the Fabry-Perot can by itself influence the resonant frequency. The microwave power dissipated in the waveguides and in the resonator is the dominant internal heat source and leads to a thermal drift of the separation b of the two mirrors. Another problem arises from the radiation pressure exerted by the field in the resonator onto the two mirrors. To avoid both effects the resonator is operated at a power level as low and as constant in time as possible. A tolerable stored energy in the resonator is $1 \cdot 10^{-11}$ J. It results in a shift of the distance b between the pendula of 0.7 nm. This stored energy is related to a transmitted power of 2 μW which is well above the detection limit of 0.01 nW. Dissipation and microwave pressure are nearly constant in time due to a control system, which keeps the microwave power constant to a relative accuracy of about 10^{-3}. The fluctuations of the microwave power are therefore estimated to result in a jitter of 1 pm in b.

The dominant systematic error in the measurement of the gravitational effect due to an external force is a tilt of the tower. The load of the test mass results in an elastic deformation of the ground and thus in a tilt of the tower. The top of the tower is shifted maximally by an amount of 0.22 μm, depending on the position of the test mass. Even so the tilt sensitivity of the gravimeter is small (about 0.2 nm mirror shift per μm tilt), a considerable part of 0.04 nm is transferred to the mirror distance. The detailed mechanism of this transfer as well as the tilt sensitivity is not yet understood. The results presented in the next chapter have been corrected by means of independent measurements of this effect. The inaccuracy of this correction corresponds to about 0.008 nm and is added to the error of the modulation amplitude which results in a significant contribution to the total error budget (see table 2). Not only for that reason we will try to eliminate this effect as soon as possible.

5. EXPERIMENTAL PROCEDURE FOR MEASURING THE GRAVITATIONAL EFFECT

The gravitational acceleration between the test mass and the Fabry-Perot resonator is measured by moving the test mass from the position r for which the gravitational acceleration is to be determined to a reference position r_{Ref} and vice versa. This procedure is repeated periodically and results in a square wave modulation of the resonators frequency [23] (Fig.12a and 13a).

A period of 30 minutes is chosen which is much longer than the time needed to move the test mass (≤ 50 s) as well as the time constant of the digital low-pass filter (60 s), and again much longer than the relaxation time of the pendula (approx. 2 s). The modulation amplitude is proportional to the difference of the differential gravitational acceleration of the test mass (equation (3)). The number of periods per position and the number of test mass positions is usually chosen to be about 10. In the case of the measurement of the small gravitational force in distances larger than 2 m, a higher resolution is desired and 4 test mass positions with each 40 periods are more suitable.

The two extrem positioning examples of the test mass shown in Fig. 12 and 13 are superimposed by a slow drift of the Fabry-Perot frequency. This predominantly thermal drift is subtracted from the data prior to further analysis, described below. To visualize the result of this analysis the obtained data (Fig.12 and 13a) are superimposed period by period and averaged. The combined time averaged signals obtained this way are shown in Fig.12b and 13b for the two extreme positioning cases.

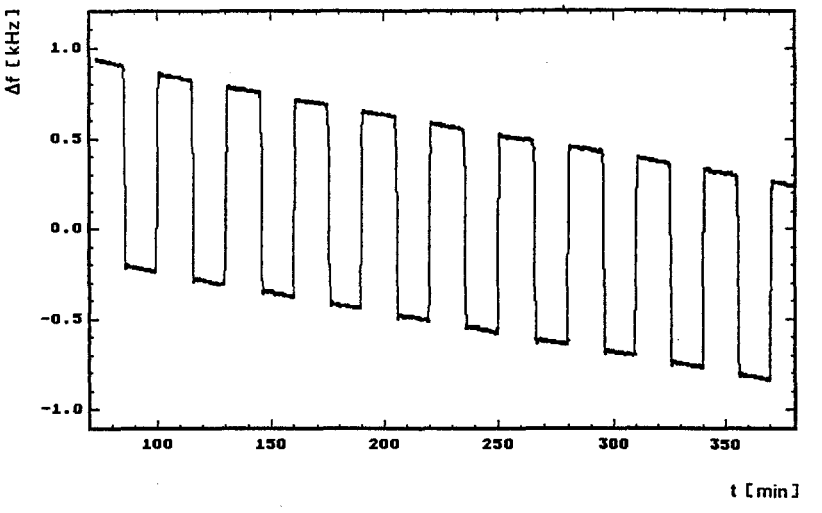

Fig.12a Modulation signal Δf of the Fabry-Perot frequency f produced by the test mass alternating between the position of closest approach $r = r_1$ and $r = r_{Ref.}$

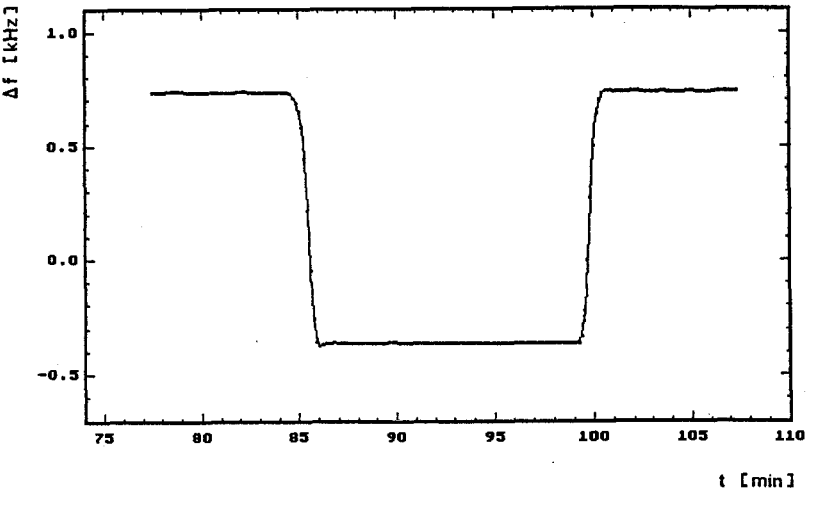

Fig.12b Combined time averaged modulation signal after subtracting the thermal drift of the mirror separation.

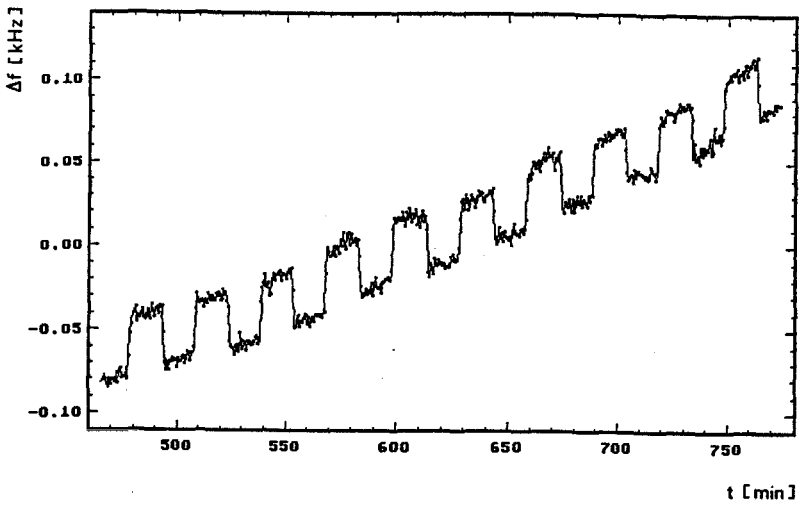

Fig.13a Modulation signal Δf of the Fabry-Perot frequency f produced by the test mass alternating between the position of largest distance $r = r_2$ and $r = r_{Ref}$.

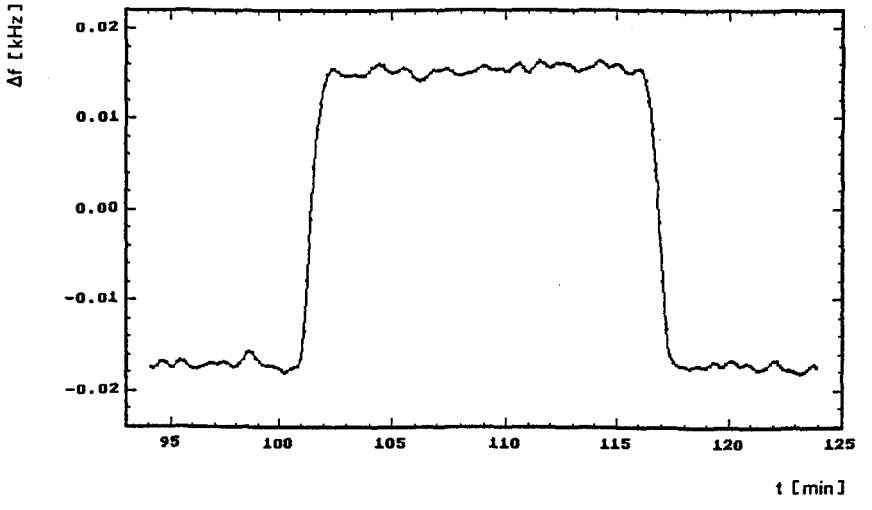

Fig.13b *Combined* time averaged modulation signal after subtracting the thermal drift of the mirror separation.

In a more precise analysis the modulation amplitude is determined by a demodulation technique. This method is based on the computation of a correlation of the measured time series of Fabry-Perot frequencies f(t) with a periodic function y(t) which is a square wave or a sine wave function:

$$C_{fy}(\omega) = \frac{1}{T} \cdot \int_0^T f(t) \cdot y(t) \cdot dt$$

$$\text{with} \quad y(t) = \sqrt{2} \cdot \sin(\omega t) \quad \text{or} \quad y(t) = \begin{cases} +1, & \text{if } \sin(\omega t) > 0 \\ -1, & \text{if } \sin(\omega t) < 0 \end{cases} \tag{7a}$$

T is the duration of the measurement and ω is the frequency of the modulation. The measured time series f(t) is approximated as a superposition of an ideal signal with an amplitude A and uncorrelated noise and drift effects N(t). The modulation amplitude is determined by evaluating the correlation function:

$$f(t) = A \cdot y(t) + N(t) \quad \text{with} \quad \lim_{T \to \infty} C_{Ny} = 0 \quad \rightarrow \quad \lim_{T \to \infty} C_{fy} = A \tag{7b}$$

The integration in equation (7a) is carried out using an integration routine or a digital low-pass filter. In the case of a square wave demodulation function, the integral is reduced to a mean value which can be determined numerically or graphically.

Various other methods are based on the computation of correlation functions. A prominent method is the Fourier series which can be carried out by computing Fourier integrals or by means of a narrow digital band-pass filter which lets only one Fourier component pass (usually the first component). All these numerical methods are mathematically equivalent. They differ only in the achievable resolution, the inherent errors and the level where they are affected by systematical errors and nonideal effects.

The common property of these methods is that statistical errors of the data, low- and high-frequency noise and the thermal drift of the mirror distance can be separated, and the change in the pendulum separation due to the gravitational acceleration can in principle be determined with high accuracy. The achieved accuracy in Δb depends on the duration of the measurement. A typical accuracy is 4 pm, corresponding to an integration time of 4 h, and results in a statistical error of the gravitational constant of $1.7 \cdot 10^{-4}$. The accuracy can be further improved to 1.3 pm, if the integration time is increased to 15 h. This is done for the range of distances above 2 m. An even longer integration time seems not to be practical. 1.3 pm has been the highest accuracy to measure a change in mirror separation achieved so far. It corresponds to a statistical error of the gravitational constant of $5 \cdot 10^{-5}$. This seems to be the fundamental limit of a normal conducting Fabry-Perot resonator and determined by its Q value.

6. RESULTS

The modulation procedure (Fig.12a and 13a) described above has been repeated with the test mass in different positions (but with the same reference position), and the shift of the resonant frequency has been measured as a function of distance between resonator and test mass. Then the modulation amplitude Δf is determined and converted into a shift of the separation of the pendula by means of the conversion factor df/db of the chosen TEM_{0035} mode which is calculated using equation (5) and given in Table 1a. Other important parameters which are necessary to determine the gravitational force are listed in table 1b.

Table 1. List of parameters necessary to determine the gravitational constant together with their systematical and statistical errors.

a)	b	= (0.241 195 ± 0.000 003) m
	df/db	= (-90.53297 ± 0.0012) Hz/nm
b)	ω_0	= 2π (0.307 800 6 ± 0.000 004 4) Hz
	M	= (575.80 ± 0.03 ± 0.11) kg
	r_1	= (0.365 46 ± 0.000 02 ± 0:000 11) m
	r_{Ref}	= (1.850 46 ± 0.000 02 ± 0.000 23) m
	r_2	= (3.333 76 ± 0.000 02 ± 0.001) m

Table 2. Listing of the statistical (a) and systematical (b) errors in the determination of the gravitational constant.

SOURCE OF UNCERTAINTY	UNCERTAINTY in $\Delta G/G / 10^{-3}$
a)	
noise and drift of mirror distance	0.17
microseismic noise	<0.02
relative test mass position	<0.02
electronic noise	<0.01
b)	
correction of the tilt effect	0.40
absolute test mass position	0.18
misalignment of the test mass	0.14
mass of the test mass	0.30
mass distribution of the resonator	0.22
mass distribution of the test mass	0.10
numerical data reduction	<0.1
microwave background	<0.02
misalignment of the resonator	0.02
average frequency of the pendulum	0.014
relative deviation of the pendula frequencies	<0.2
conversion factor	0.014
tilt of the Gaussian beam	<0.01
magnetic forces	<0.001

Table 2 gives a summary of the important statistical and systematical errors which finally determine the accuracy by which G or the $1/r^2$ dependence can be determined. Most of these errors have already been discussed in the proceeding sections and a few are given as estimates without further discussion.

The theoretical shift of the mirror distance due to the gravitational force is computed from the integration of the inverse square law over the mass distribution of test mass and resonator. The value of the gravitational constant used for this computation is the CODATA-value $G_C = 6.6726 \cdot 10^{-11}$ N m^2 kg^{-2} [5].

The distances r between test mass and resonator are measured for convenience from the front side of the test mass to the front side of the resonator, as well as the reference position r_{Ref}. (The distance of the centers of gravity of the test mass and the pendulum next to it is obtained by adding 0.2541 m). r_1 is the distance from the test mass in the front position on the guide rail to the resonator and thus the smallest distance for which the gravitational force could be measured. In a first experiment the gravitational force has been measured in the range from r_1 to r_{Ref}. In a second experiment the distances have been increased to the range from r_{Ref} to r_2. This experiment was carried out with the guide rail shifted by $r_{Ref} - r_1$. The chosen set of distances r in both ranges is not given here and can be read from the corresponding figures.The values of r_1, r_2 and r_{Ref} are listed in Table 1b).

6.1 THE GRAVITATIONAL FORCE IN THE RANGE OF 0.6 M TO 3.6 M

In Fig.14 the values of the measured shift Δb of the mirror distance b for the test mass positioned in the range from 0.6 m to 2.1 m are plotted versus the computed values. The data points are normally distributed and well on a straight line as expected in the case of an inverse square law. No significant deviations are observed. The full line is a least-squares fit to the data using relation (3). The gravitational constant G is determined from the slope of this line to be

$$G = (6.6613 \pm 0.0011 \pm 0.0093) \cdot 10^{-11} \text{ N m}^2 \text{ kg}^{-2}$$

and

$$\frac{G - G_c}{G_c} = (-1.69 \pm 0.17 \pm 1.39) \cdot 10^{-3}.$$

The first error is the statistical error as determined from the least-squares fit (Table 2 a) and amounts to only $1.7 \cdot 10^{-4}$. In addition systematical errors of about $1.4 \cdot 10^{-3}$ have to be taken into account (Table1 and Table 2 b). The systematical error is the second error quoted.

The value of the gravitational constant G as determined in this experiment and the CODATA-value G_C [5] differ by $1.7 \cdot 10^{-3}$. If the total relative error of our experiment ($1.6 \cdot 10^{-3}$) is considered, both values agree very good.

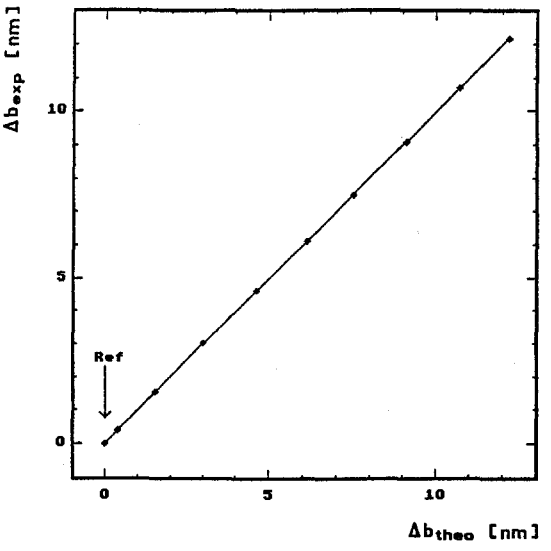

Fig.14 The measured modulation amplitude Δb_{exp} of the distance b between the two Fabry-Perot mirrors due to the gravitational force of the test mass in different distances to the resonator versus the theoretical values Δb_{theo} which were computed from equations (1) and (2) using the CODATA value G_C. In this plot a gravitational force following Newtons law is a straight line. The slope of this line is our experimental value of G in units of G_C.

In Fig.15 the measured shift of the mirror separation is plotted versus the distance r. Fig.16 gives the deviation from the theoretical values which are based on the measured value of the gravitational constant G. The results measured in the range of 2.1 m to 3.6 m are included in Fig.15 and 16. A more detailed figure is inserted to Fig.15. As can be seen from these figures, the measured results agree very well with the inverse square law and no significant deviations are observed.

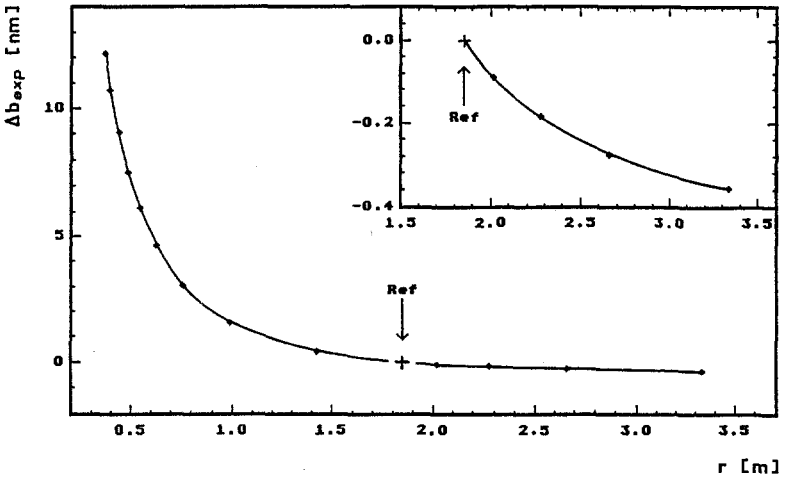

Fig.15 The measured modulation amplitude Δb_{exp} versus the distance r of the test mass from the Fabry-Perot resonator. The full line corresponds to Newton's $1/r^2$ law.

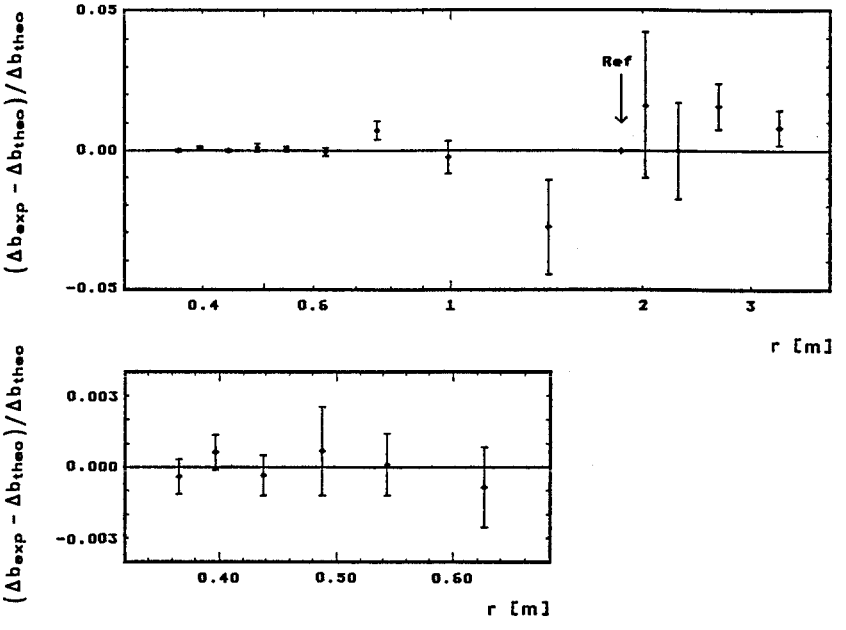

Fig.16 The relative difference between the experimental results and the computed values using the gravitational constant G as determined in this experiment.

Possible deviations from the inverse square law can be described using a modified law of the following form:

$$F(r) = \frac{G\,M\,m\,r_0^{\delta}}{r^{2+\delta}} \sim \frac{1}{r^2}\cdot\left(1-\delta\cdot\ln(r/r_0)+\ldots\right) \quad \text{for} \quad \delta << 1 \tag{8}$$

The gravitational force is assumed to decrease with a power (2+δ) of the distance of the interacting masses. In Newton's gravitational law the parameter δ equals zero. The parameter r_0 has an absulute value of 1 and is introduced only to keep the conventional units of the gravitational constant.

This modified gravitational force was integrated over the mass distribution of test mass and resonator, and the shift of the mirror distance Δb was calculated as a function of distance. The parameter δ was then determined from a least squares fit to be

$$\delta = (2.1 \pm 1.8)\cdot 10^{-3}$$

The quoted error includes both statistical and systematical errors. The measured value of δ is consistent with zero, that is the gravitational force measured in this experiment agrees well with the inverse square law.

7. SUMMARY

We have tested a new method to experimentally determine the gravitational force produced by a test mass in a range of distances between 0.6 m and 3.6 m from its center of mass. The two mirrors of a Fabry-Perot microwave resonator are suspended as a pair of pendula and form the gravimeter. The gravitational force of a laboratory test mass exerted on this resonator changes the distance between the two mirrors. The resulting frequency change is used to test the gravitational law. The test mass is brought to a distance r of the gravimeter and removed from it into a reference position by a periodic motion. The analysis of the resulting periodicity in the change of the resonant frequency allows a strong suppression of the influence of the random noise and and thermal drift phenomena. The change in separation of the two mirrors by the action of the gravitational force was measured as a function of distance with an absolute accuracy of $5 \cdot 10^{-12}$ m. The relative deviations of the experimental data from the predictions obtained using Newton´s inverse square law were between about $1 \cdot 10^{-3}$ and $3 \cdot 10^{-2}$ and well within the estimated errors of our experiment. The gravitational constant was determined with a relative accuracy of $1.5 \cdot 10^{-3}$ and its value is consistent with the CODATA result.

The presently achieved precision is not inherent to the chosen experimental method. The ultimate sensitivity of the gravimeter, which is determined by the finite Q of the Fabry-Perot resonator, limits the relative precision by which G can be determined to $5 \cdot 10^{-5}$. Our present uncertainty is limited by systematic errors which we intend to reduce in our future work. The obtained sensitivity however is large enough to increase the distances between test mass and gravimeter into the 5 m range. The inherent limitation of the method can be most likely overcome by replacing the copper mirrors of the Fabry-Perot resonator by superconducting ones. We have demonstrated in an earlier experiment, that at a temperature of 4.2 K a Q value of $1.8 \cdot 10^7$ can be obtained with niobium mirrors [22]. Recent results of the microwave surface resistance of the high T_c superconductor $YBa_2Cu_3O_7$ at about 20 GHz [31] seem to open the possibility to use high T_c mirrors in the gravimeter for an improved sensitivity already at 77 K.

ACKNOWLEDGEMENTS

We thank B. Langensiepen for his valuable contributions to this experiment. We are very obliged to SIEMENS, Bensberg for its contribution of the cryostat and the microwave sweep generator. This work was funded in part by the Deutsche Forschungsgemeinschaft under grant number Pi 118/4-1.

REFERENCES

[1] S.W. Hawking, W. Israel (Ed.), "Three Hundred Years of Gravitation" Cambridge University Press, Cambridge 1987.

[2] Proceedings of the Moriond Workshop 1988 to 1991, Les Arcs, France, Ed.: O. Fackler and J. Tran Thanh Van, Edition Frontieres, Gif-sur-Yvette.

[3] G.T. Gillies, "The Newtonian Gravitational Constant. An Index of Measurements", Metrologia 24, 1-56 (1987).

[4] A. De Rujula, Phys. Lett. 180 B, 213 (1986).

[5] E.R. Cohen, B.N. Taylor, Physics Today BG 9 (1991).

[6] V.P. Mitrofanov, O.I. Ponomareva, Sov. Phys. JETP 67, 10 (1988).

[7] J.K. Hoskins, R.D. Newman, R. Spero, J. Schultz, Phys. Rev. D 32, 3084 (1985).

[8] Y. Ogawa, H. Hirakawa, K. Kuroda, N. Mio, Phys. Rev. D 26, 729 (1982), Phys. Rev. D 32, 342 (1985), Phys. Rev. D 36, 2321 (1987) and KEK Preprint 88-99 H, (1988).

[9] V.I. Panov, Sov. Phys. JETP 50, 5 (1979).

[10] H.T. Yu, W.T. Ni, C.C. Hu, F.H. Liu, C.H. Yang, W.N. Liu, Phys. Rev. D 20, 1813 (1979).

[11] V.B. Braginskii, V.I. Panov, Soviet Physics JETP 34, 463 (1972).

[12] D.R. Mikkelsen, Phys. Rev. D 16, 919 (1977).

[13] C. Jekeli, D.H. Eckhardt, A.J. Romaides, Phys. Rev. Lett. 64, 1204 (1990).

[14] J. Thomas, P. Vogel, Phys. Rev. Lett. 65, 1173 (1990) and references therein.

[15] M.E. Ander, M.A. Zumberge, T. Lautzenhiser, R.L. Parker, C.L.V.Aiken M.R. Gorman, M.M. Nieto, A.P.R. Cooper, J.F. Ferguson, E. Fisher, G.A. McMechan, G. Sadagawa, J.M. Stevenson, G. Backus, Aland D. Chave, J. Greer, P. Hammer, B.Lyle Hansen, J.A. Hildebrand, J.R. Kelty, C. Sidles and J. Wirtz, Phys. Rev.Lett. 62, 985 (1989).

[16] M.A. Zumberge, J.A. Hildebrand, J.M. Stevenson, R.L. Parker, A.D. Chave, M.E. Ander, F.N. Spiess, Phys. Rev. Lett. 67, 3051 (1991).

[17] Y. Fujii, Ann. Phys. 69, 494 (1972).

[18] John O'Hanlon, Phys. Rev. Lett. 29, 137 (1972).

[19] J. Scherk, Phys. Lett. 88 B, 265 (1979).

[20] T. Goldman, R.J. Hughes, M.M. Nieto, Phys. Lett. 171 B, 217 (1986).

[21] R.D. Peccei, J. Sola, C. Wetterich, Phys. Lett. 195 B, 183 (1987).

[22] N. Klein, thesis, Universität Wuppertal, Germany, 1989, WUB-DIS 89-3; J. Schurr, diplom thesis, University of Wuppertal, Germany, 1988, WUD-88-11.

[23] J. Schurr, N. Klein, H. Meyer, H. Piel, H. Walesch, Metrologia 28, 5 (1991).

[24] H. Kogelik, T.Li, Appl. Opt. 5, 1550 (1966).

[25] G.D. Boyd, J.P. Gordon, Bell Syst. Tech. J. 40, 489 (1961).

[26] A.L. Cullen, F.R.S. and P.K. Yu, Proc. R. Soc. Lond. 366 A, 155 (1979).

[27] K.-M. Luk, P.-K. Yu, IEE Proceedings 132, 105 (1985).

[28] G.L. Ragan, "Microwave Transmission Circuits", McGraw-Hill Company, Inc., New York 1948, p. 100 ff and 291 ff.

[29] E. Hardtwig, "Geophysikalische Monographien, Band 3", Ed.: W. Buchheim, Akademische Verlagsgesellschaft Geest und Portig KG, Leipzig 1962.

[30] K. Akl, P.G. Richards, "Quantitative seismology", W.H. Freeman and Company, New York 1980

[31] H. Piel, H. Chaloupka and G. Müller, to be published in "Proceedings of the 4th Int. Symposium on Superconductivity", Tokyo, (1991).

Matter Wave Interferometry and why Quantum Objects are Fundamental for Establishing a Gravitational Theory

by

Jürgen Audretsch*, Friedrich W. Hehl◇, and Claus Lämmerzahl*

*Department of Physics, University of Constance, D(W)-7750 Konstanz
Germany (e-mail: claus@spock.physik.uni-konstanz.de)

and

◇Institute for Theoretical Physics, University of Cologne, D(W)–5000 Köln 41
Germany (e-mail: hehl@thp.uni-koeln.de)

Abstract

More recently, a number of interferometric experiments with electrons, neutrons, and atoms have been performed in the gravitational field of the earth and in non-inertial frames of reference. In atomic interferometry, additional high-precision experiments are expected to be done in the near future. The results obtained with electrons, neutrons, and atoms, respectively, can be understood by means of the Schrödinger or, in the polarized case, by means of the Pauli equation, both of which are coupled to the external gravito-inertial field. Based on the characteristic features read off from these experiments, one can set up a constructive axiomatic approach for establishing an appropriate spacetime geometry and can, independently, develop a gauge theoretical formalism for gravity. Both constructions make the Riemann-Cartan geometry of spacetime manifest. This geometry carries *torsion* as well as *curvature*. The Riemannian geometry of Einstein's gravitational theory can be recovered as a limiting case for the motion of classical point particles and light rays. We put the Dirac equation, formulated in a non-inertial frame of reference, into an arbitrary gravitational field represented by the spacetime geometry obtained. We compute the consequences for interferometric experiments and provide thereby a theoretical basis for future experiments.

*) Supported by the Deutsche Forschungsgemeinschaft and the Commission of the European Community, DG XII.

◇) Supported by the German–Israeli Foundation for Scientific Research and Development (GIF), Jerusalem and Munich.

1. Introduction

In these lectures we want to illustrate and to support the following thesis: Quantum objects are fundamental for the establishment of the structure of spacetime and thereby, also, for the theory of gravity.

To do so, we start our reasoning by drawing attention to the experimental foundations of the interaction between quantum objects and gravity. Today, matter wave interferometry with electrons, neutrons, and atoms provides an ever increasing number of experiments in which the influence of gravity and inertia on quantum objects can be studied in a very direct and precise way. A simple description of these fundamental experiments makes use of the Pauli equation in a homogeneous *gravitational* field and in a *non-inertial* frame of reference. There is presently still a big gap between this experimental level on the one hand side and the theoretical level at which gravity is quantized on the other hand side. But if one restricts oneself to classical gravity, it is possible to read off from the experimental results some characteristic traits on which a theory of the structure of classical spacetime can be founded.

Matter wave interferometry convincingly demonstrates the importance of gravity and inertia in the quantum domain. Accordingly, we need a corresponding theory which, however, presupposes a framework for the spacetime structure in the quantum domain. How can it be established? We know from the corresponding situation in classical physics that it would be unsatisfactory to put the postulate 'Spacetime obeys a Riemannian geometry' at the outset of a theory of spacetime. What we rather need is a physically conclusive reasoning which will eventually lead to a statement of this type. For classical physics, there have been many efforts to establish such a statement as a result of basic postulates related to the equivalence principle for classical matter fields or, in the constructive axiomatic approach, to the behavior of point particles and light rays. In experiments point particles and light rays are typically realized by satellites and radar signals and, conversely, they characterize the domain of application of this approach. It is evident that, for instance, the interior of the hydrogen atom including its nucleus cannot be explored in this way. On the other hand, to refer for the quantum domain once more to the postulate cited above, would be as unsatisfactory as it has already been in the classical domain. Instead, we have to take into account quantum mechanical experience right from the beginning. This is what we are going to do.

A further reason for this procedure is the following: There is a hierarchy within the theories of matter. Quantum physics is more fundamental than classical physics. The latter is contained in the former as a limiting case. Matter, classically described, such as satellites, stones, and other candidates for point particles, is composed of quarks, leptons, and their gauge bosons. The gravitational and inertial behavior of the complex objects should be a consequence of the behavior of the more elementary objects. It is therefore reasonable, if not compelling, to relate a theory of the structure of spacetime to the more fundamental theoretical framework of quantum mechanics, which means, by the same token, to base it on the more primitive objects, namely on the elementary particles. And this even more so, because the influence of gravity

on classically described matter can be derived, as a limiting case, from the more fundamental quantum approach.

This is not to be confused with the fact that quantum mechanical experiments are performed using classical measuring devices. For a quantum based constructive axiomatics, for example, only empirical facts will be used which can be read off in a geometry-free way.

Finally we mention still another reason for relying on quantum objects as primitive objects when exploring the structure of space-time: quantum objects, as compared to classical point particles and light rays, are the deeper searching probes. The interference experiments demonstrate that massive fields with spin couple to gravito-inertial fields in accordance with the strong equivalence principle. The experimental results depend on the parameters mass and spin. Based on this richer structure of the primitive objects, additional physical structures can be geometrized yielding more specific statements on spacetime geometry. For example, the torsion of spacetime can be "sensed" if spacetime is explored by particles carrying spin. This makes it implausible to restrict gravitational theory to the torsion-free Einstein theory from the outset without giving a physical justification for this restriction of vanishing torsion. Only after having established a theory of spacetime with torsion, one can try to look for experiments which may show that torsion is negligibly small in certain domains.

Guided by these heuristic considerations and founded on an empirical basis, the following two different procedures for establishing the geometrical structure of spacetime seem to be natural: (i) The gauge approach to gravity which represents the generic gravitational theory for quantum mechanical matter fields and which incorporates the equivalence principle in an essential way. Following Einstein, the structure and form of the gravitational potentials are read off in flat spacetime from the inertial forces arising in non-inertial frames. Following Cartan (1986), in a second step, arbitrary non-inertial reference frames are identified with a field of orthonormal (anholonomic) tetrads. (ii) The constructive axiomatics, as an alternative approach, does not refer to special relativity. Instead, nearly all elements of spacetime geometry are built up by reformulating as postulates experience largely gained from matter wave interferometry. Both approaches independently result in a *Riemann-Cartan* spacetime, carrying *torsion* as well as *curvature*, thus validating the thesis stated above.

The article is organized as follows: As experimental background, in Sect.2, matter wave interferometry is described with reference to the Pauli equation which is coupled to an external Newtonian gravitational field. In Sect.3, the fundamental physical consequences are pointed out. Based on this, in Sect.4, the constructive axiomatic approach and, in Sect.5, the gauge approach are concisely presented. Finally, in Sect.6, starting from the spacetime structure established, an approximation scheme is given for the description of interference experiments in gravitational and inertial fields. It can be used when searching for new measurable effects.

Acknowledgments: We are grateful to the W.&E.Heraeus-Stiftung and to Dr. Gerhard Schäfer for the invitation to present lectures at the Bad Honnef School on Gravitation.

For his conscientious reading of some parts of the manuscript, we would like to thank Prof. J. Dermott McCrea (Dublin); we also thank him for his permission to use some of his unpublished work [McCrea (1989)]. We also thank Dr. Olivier Carnal and Prof. Jürgen Mlynek (Konstanz) for helpful discussions and Prof. U. Bonse (Dortmund) and Dr. F. Hasselbach (Tübingen) for providing pictures of their apparatuses.

2. Experimental background

2.1. Interferometers

Interferometry belongs to the fundamental experiments in physics. By means of interferometry one can study the structure of the interferring light and matter waves as well as the type of interactions of these waves with external fields.

Interferometry of light waves has been known for quite some time. It can be described by means of the eikonal approximation of the Maxwell equations. Interferometry of matter waves can be understood only if one takes into account the quantum theory of matter (at least within a certain approximation). Therefore matter wave interferometry provides a tool for testing some principles of quantum theory as well as the influence of external fields on quantum matter.

Up to now there are three types of matter waves at our disposal[1] for which interference had been observed and which can be used to study the interaction with external fields. These are electrons, neutrons, and atoms. The corresponding interferometers are mostly of the Mach-Zehnder type, that is, there are spatially separated matter beams as, for instance, in the Bonse and Hart (1965) perfect crystal interferometer. Other types of interferometric setups are also possible, such as the atomic fountain setup of Kasevich and Chu (1991). Furthermore, by means of exciting trapped atoms, one can do interferometry of atoms which remain at the same place.

Electrons and neutrons are most conveniently described by means of the Pauli or the Dirac equation. Atoms are, of course, more complex objects and should be described in an n-particle approach. In some approximation, this yields a Pauli-type equation with magnetic and electric dipole moments or its respective relativistic version. This represents a center-of-mass motion with additional degrees of freedom. In the following we will restrict the external forces to gravitational and inertial forces, that is, we study the influence of the spacetime geometry on quantum matter.

Electrons are the first matter waves proper interferometry has been done with. Today charged particle interferometry is still based on electrons.

The advantages of *neutron interferometry* consist in the simplicity and the macroscopic dimensions of the interferometer. The comparatively large separation of the neutron beams provide a device to study quantum mechanics in macroscopic dimensions.

---

[1] As the earliest matter wave 'interferometer', sensitive to an external gravitational field, one may consider the $K^0\overline{K^0}$-meson system as described by Good (1961).

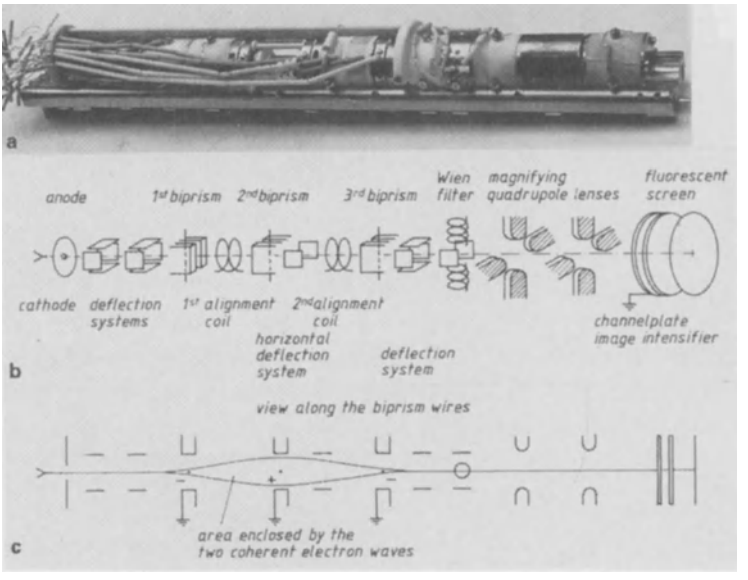

Figure 1. a) The triple biprism interferometer, b) the electron optical set up, c) the path of the electron beams [taken from Hasselbach and Nicklaus (1988)].

In comparison to neutron interferometry, *atomic beam interferometry* provides several advantages: (i) By means of laser cooling and trapping, atomic beams can be prepared with very low velocites, (ii) atoms have a larger mass and hence smaller deBroglie wavelengths [this together with (i) yields for a large class of interference experiments an increased accuracy], (iii) there are much more possibilitites to manipulate atomic beams because of their internal degrees of freedom, (iv) sources of atomic beams are much easier to handle, and (v) because of the internal degrees of freedom there are additional effects which may possibly be tested with atom beam interferometry based on new types of interferometer experiments. However, because of the low velocities of the atoms, the experiments are not well suited for testing relativistic effects.

2.1.1. Electron

Interference of electrons has first been observed by Marton *et al.* (1953) by using crystal plates as beam splitters where the electrons undergo Bragg scattering. An effective type of electron interferometer was built by Möllenstedt and co-workers (1954, 1961) using a triple biprism (see Fig.1). They achieved a beam splitting of about $100\,\mu m$ and a path length of about $10\,cm$. The electrons had an energy of 1 keV and hence a velocity of about $v \approx 0.06\,c$, where c is the velocity of light.

Also a double-slit interferometer for electrons has been built by Möllenstedt and Jönsson (1959).

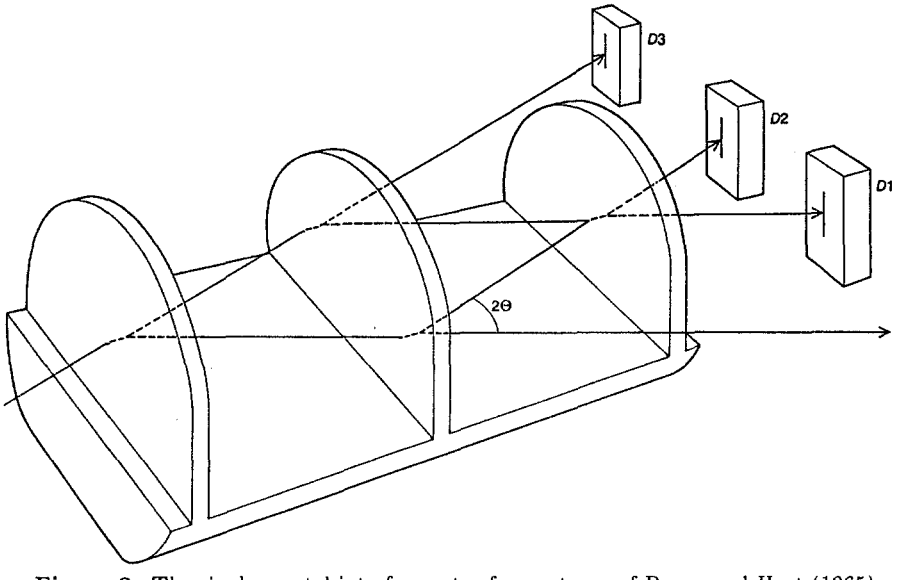

Figure 2. The single crystal interferometer for neutrons of Bonse and Hart (1965). The incoming neutron beam is spit at the first slab. The second slab serves as mirror and at the third slab the beam is recombined. The intensity of the interfering beams can be read off from the counters D1 and D2. Θ is the Bragg angle. D3 is a reference counter.

2.1.2. Neutron

The neutron interferometer designed by Rauch, Treimer, and Bonse (1974) is, because of its conceptual simplicity, a very successful interferometer for quantum matter waves. It demonstrates the wave aspect of matter on macroscopic scales.

The interferometer consists of a silicon single crystal (see Fig.2). Three slabs are cut from the crystal. The first two slabs serve as beam splitter and mirror, respectively, whereas the last one recombines the two beams such that the information related to the interference is coded onto the beams leaving the third slab of the interferometer. By means of this set-up, one does not observe any interference pattern directly as, for instance, in the case of a double slit experiment with light where the interference fringes are displayed on some screen. Instead, this neutron interferometer set-up is designed for observing phase shifts induced by varying external parameters, like the orientation in the gravitational field or the strength of some magnetic field influencing one of the neutron beams.

Neutron waves entering the crystal undergo Bragg scattering at the atomic planes. Within the crystal the neutron beams propagate perpendicular to the crystal face. When leaving the crystal they split into a forward and a backward beam.[2] The height and the length of the interferometer are of the order of 10 cm. This means that

[2] Actually, the propagation within the crystal is more complicated because there

one can do quantum mechanics on a macroscopic scale. The matter waves used are thermal neutrons with a wavelength of about 0.1 μm, which is equivalent to a velocity of about 2000 m/s. The coherence length of such a neutron is about 30 μm, whereas the extension of the wave packet is of the order of 1 cm.

A double slit set-up for neutron interferometry has been built by Klein *et al.* (1981) to confirm the Fizeau effect for neutrons.

2.1.3. Atom

Today there are five types of atomic beam interferometers working around the world.

(i) The first one, based on a double slit as mechanical beam splitter, was build by Carnal and Mlynek (1991) [see also Carnal (1992)]. The slit width is 1 μm and the two slits are 8 μm apart. The length of the path is about 1 m (see Figs.3 and 4). Helium atoms are used with a velocity of 500 m/s.

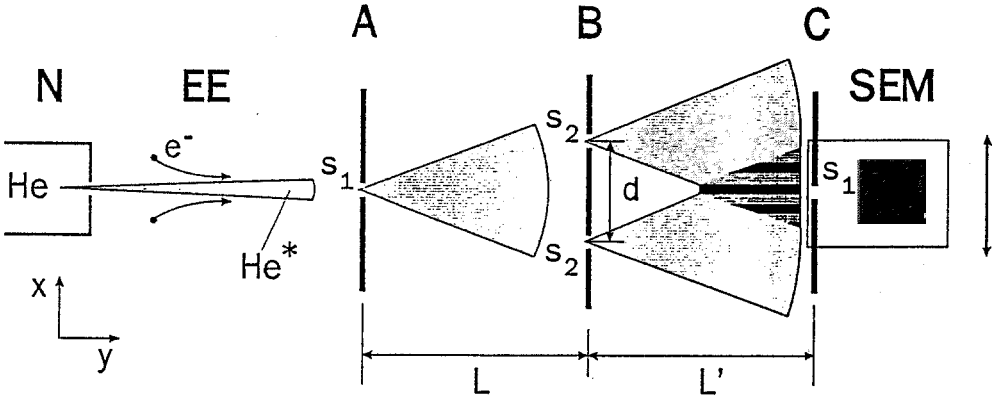

Figure 3. Schematic representation of the atomic beam interferometer of Carnal and Mlynek (1991). N: nozzle system and gas reservoir; EE electron impact excitation; A: entrance slit; B: double slit; C: detector screen; SEM: secondary electron multiplier. d=8 μm, L=L'=64cm, s_1=2 μm, s_2=1 μm.

(ii) Keith *et al.* (1991) took a grating as mechanical beam splitter with about the same geometric dimensions as the double slit.

(iii) Riehle *et al.* (1991) used a totally different device as beam splitter: As pointed out by Bordé (1989), one can use four travelling laser waves for transmitting to

---

are two slightly different propagation directions which, by interference, yield the so-called 'pendellösung'. Hence four beams leave one slab leading to eight possible paths within the interferometer, which all interfere at the last slab. Since the phase shifts as, for example, those induced by gravito-inertial effects, depend on the geometry of the paths, it influences the interpretation of the measuring data. This problem is still under investigation, see Horne (1986) and Werner and Kaiser (1990).

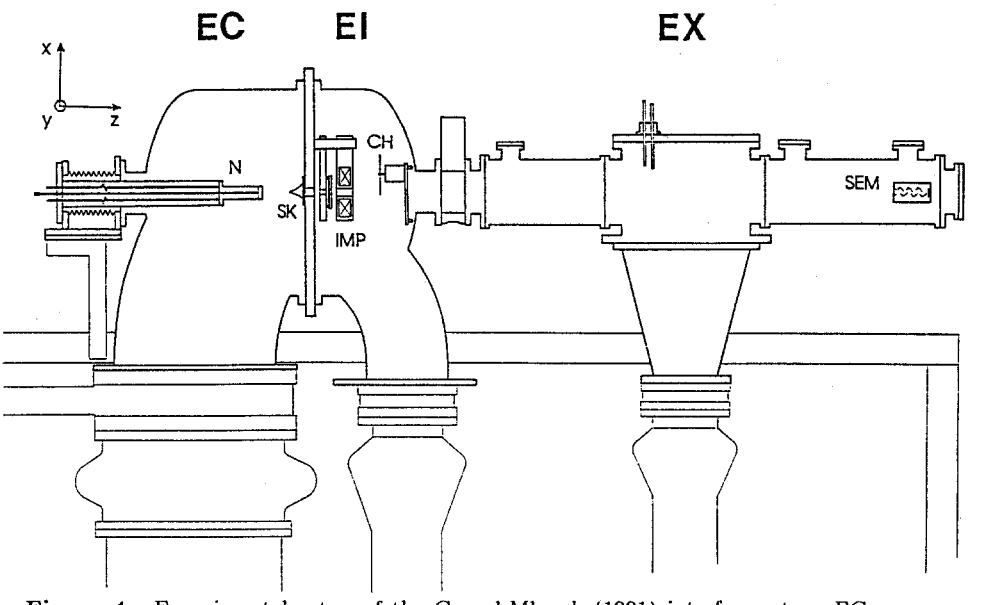

Figure 4. Experimental setup of the Carnal-Mlynek (1991) interferometer. EC: expansion chamber (N nozzle system, SK skimmer orifice); EI: electron impact excitation area; CH chopper (velocity selector); EX: experimental chamber (insertion of double slit, etc.); SEM: secondary electron multiplier.

the atoms well defined momenta. By resonant absorption and emission processes, energy and momentum is exchanged between atoms and photons (see Fig.5). A wave packet with momentum p and internal state $|a>$ is put at the splitter 1 into a superposition of states $|a, p_0>$ and $|b, p+\Delta p>$ by the momentum transfer

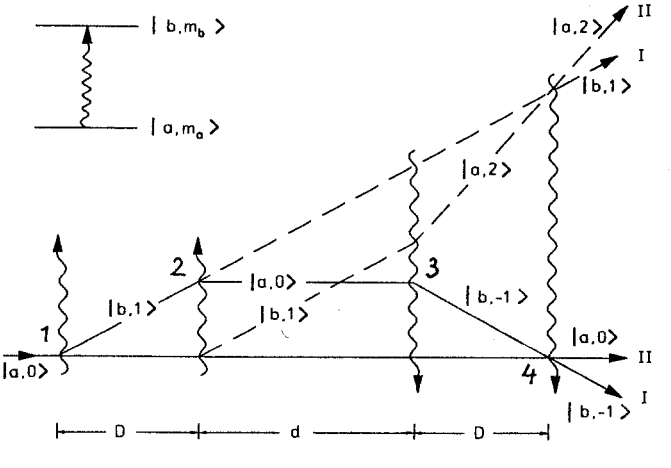

Figure 5. Beam splitting of an atomic beam by means of optical Ramsey excitation using four travelling laser fields. In the first (left) interaction zone the atomic matter wave is coherently split. The second and third interaction zones act as mirrors. In the last zone the beams recombine and interfere with one another [Bordé (1989)].

$\Delta p = \hbar k$ from the laser wave. After similar laser induced transitions $|b, p + \Delta p > \longrightarrow |a, p >$ and $|a, p > \longrightarrow |b, p - \Delta p >$ in points 2 and 3, the two atomic waves interfere at point 4 after interaction with a fourth laser beam. (A second possiblity is represented by the dashed lines in Fig.5. This optical beam splitter allows coherent separation and recombination of atomic beams. In the realisation of Riehle *et al.*, Calcium atoms are used with a velocity of about $700\, m/s$ and a momentum transfer of about $2 \times 10^{-28}\, kg\, m/s$ corresponding to a deflection angle of $22\, \mu$rad.

Note that because of the small beam separation it is not possible to put anything between the paths. However, because the position of the atom for the transitions between the internal states is not relevant, this configuration provides a high atomic beam flux.

The same type of interferometer, but using Magnesium atoms, has recently been built by Ertmer (1991).

(iv) Kasevich and Chu (1991) used Raman transitions of Sodium atoms to transmit a well defined momentum from the laser light to the atoms. They used laser cooled atoms with a temperature of 30 μK, which is equivalent to a velocity of $18\, cm/s$. They have done this for two configurations: the Mach-Zehnder and the atomic fountain configuration. Whereas in the Mach-Zehnder configuration the atom beams are spatially separated, the atoms move in one direction only in the fountain configuration (see Fig.6). They absorb and emit momenta from the laser light in the direction of motion in such a way that, before recombination, half of the atoms in the beam is travelling faster than the other half.

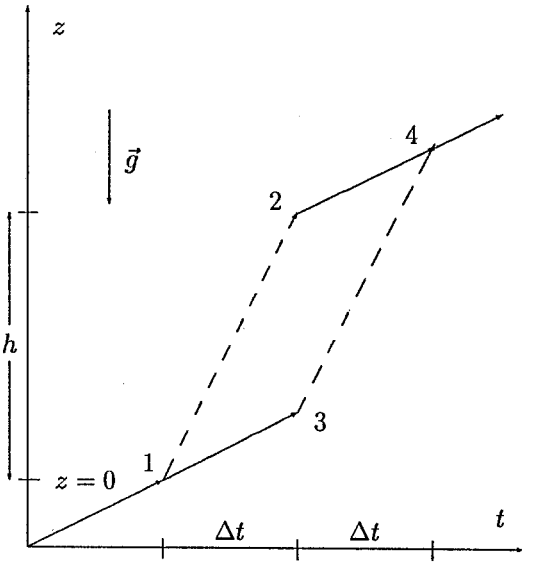

Figure 6. Space-time diagram of an atomic beam interferometer using an atomic fountain. Solid lines: state $|1,p>$. Dashed lines: state $|2,p+\Delta p>$.

(v) The most recent atomic beam interferometer was built by Shimizu, Shimizu, and Takuma (1992). They use ultra-cold $1s_3$ metastable Neon atoms the trajectories of which are determined only by the inital velocity of the atoms and by the gravitational acceleration (see Fig.7).

Other beam splitters and mirrors for atomic beams are under construction: Examples are beam splitters based on the Kapitza-Dirac effect [Kapitza and Dirac (1933)], on Bragg scattering of atoms from standing light waves [Martin *et al.* (1988)],

on the Stern-Gerlach effect [Miniatura *at al.* (1991)], or using the concept of velocity-tuned resonances [Glasgow *et al.* (1991)].

2.2. Gravito-inertial experiments and their simple theoretical description

In this section we describe the outcome of the interference experiments with the help of the simplest theory, that is, by means of the Pauli equation in a rotating and linearly accelerating frame under the influence of Newtonian gravitation. The measurability of the corresponding effects has been discussed for atomic beam interferometers by Clauser (1988) and Audretsch and Lämmerzahl (1992).

2.2.1. General formula for the phase shift

We use the Pauli equation as field equation for describing the propagation of matter waves with spin 1/2 in an external constant gravitational field $\vec{g}$:

$$i\hbar\frac{\partial}{\partial t}\psi = \left[-\frac{\hbar^2}{2m_i}\Delta + \varphi_{grav}\right]\psi = \left[-\frac{\hbar^2}{2m_i}\Delta - m_g\vec{g}\cdot\vec{r}\right]\psi\,. \tag{2.2.1}$$

Here we distinguish between the inertial mass m_i and the gravitational mass m_g. We transform to a frame with constant rotation $\vec{\omega}$ and linear accceleration $\vec{a}$. Then, to first order in $\vec{\omega}$ and $\vec{a}$, we find

$$i\hbar\frac{\partial}{\partial t}\psi = \left[-\frac{\hbar^2}{2m_i}\Delta + \vec{\omega}\cdot(\vec{r}\times i\hbar\nabla + \frac{\hbar}{2}\vec{\sigma}) - (m_g\vec{g} - m_i\vec{a})\cdot\vec{r}\right]\psi\,, \tag{2.2.2}$$

where $\vec{\sigma}$ represents the Pauli matrices.

The phase shift for a matter wave interference experiment can be calculated in the the semi-classical WKB approximation: $\psi = \varphi\exp\left(\frac{i}{\hbar}\phi\right)$, with $\nabla\varphi \sim 0$ and $\nabla\nabla\phi \sim 0$. Substitution of this ansatz into (2.2.2) yields

$$E\varphi = \left[\frac{p^2}{2m_i} + \vec{\omega}\cdot(\vec{L} + \vec{S}) - (m_g\vec{g} - m_i\vec{a})\cdot\vec{r}\right]\varphi\,, \tag{2.2.3}$$

with $\vec{L} := \vec{r}\times\vec{p}$ denoting angular momentum, $\vec{S}$ spin angular momentum, and $E := -\partial_t\phi$ and $\vec{p} := -\nabla\phi$ energy and momentum, respectively. We choose $\vec{r} = 0$ at the beam splitter.

An interference experiment must be done under quasi-stationary conditions, otherwise the interference fringes may wash out. For the theoretical description, however, we assume strict stationarity. Nevertheless, the results obtained may be used in order to describe adiabatic changes of parameters. In the calculation we take E to be constant. Now we solve (2.2.3) with respect to p and have, to first order in the interactions,

$$p = \sqrt{2m_i\left[E_{\rm kin} - \vec{\omega}\cdot\vec{L} + (m_g\vec{g} - m_i\vec{a})\cdot\vec{r}\right]} \approx p_0\left(1 - \frac{E_{\rm int}}{2E_{\rm kin}}\right)\,, \tag{2.2.4}$$

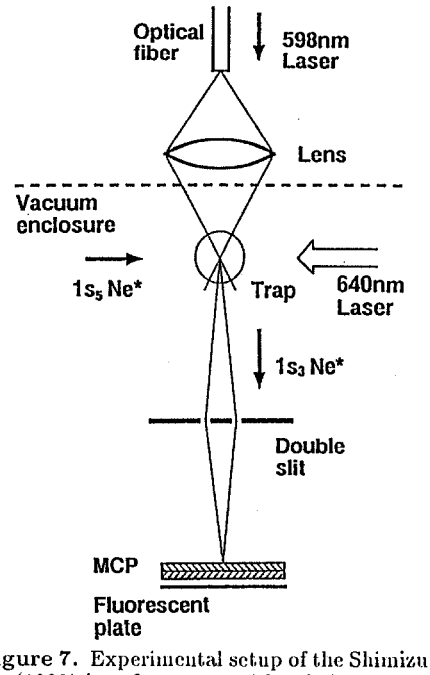

Figure 7. Experimental setup of the Shimizu *et al.* (1992) interferometer. After being trapped, the atoms fall freely through the double slit. The experiment is done for different positions (heights) of the screen (fluorescent plate).

with $E_{\rm kin} := \dfrac{p_0^2}{2m_i}$, $E_{\rm int} := \vec{\omega} \cdot (\vec{r} \times \vec{p}_0) + (m_g \vec{g} - m_i \vec{a}) \cdot \vec{r}$, and p_0 being the modulus of the momentum of the matter wave at the position of the beam splitter.

Then the resulting phase shift in an interference experiment with symmetric paths is given by

$$\delta\phi = \frac{1}{\hbar} \oint \vec{p} \cdot d\vec{r} = -\frac{1}{\hbar} \oint E_{\rm int} dt \,, \qquad (2.2.5)$$

where dt has to be calculated from the classical path and the group velocity of the wave packet as determined by the plane wave solutions. Herewith we obtain the well-known formula [see Heer (1961), Overhauser and Colella (1974), Page (1975), Anandan (1977)]

$$\delta\phi = \frac{1}{\hbar} \oint \left[\vec{\omega} \cdot (\vec{r} \times \vec{p}_0) + (m_g \vec{g} - m_i \vec{a}) \cdot \vec{r} \right] dt = \underbrace{\frac{1}{\hbar} \frac{(m_g g - m_i a) A}{v_0}}_{\text{grav.\& accel.effect}} + \underbrace{2 \frac{m_i}{\hbar} \vec{\omega} \cdot \vec{A}}_{\text{Sagnac type effect}} \,, \qquad (2.2.6)$$

where A is the interferometer area. Note as characteristic results that even for $m_i = m_g$ the mass parameters do not drop out and that $\vec{g}$ and $\vec{a}$ have an equivalent influence.

The interaction energy $\vec{\omega} \cdot \vec{S}$ of (2.2.3) can only be observed if for one path a spin flip is imposed after splitting and before recombination of the two beams, see Mashhoon (1988).

Effects of tidal forces, that is space variations of the eath's gravitational field $\vec{g}$, seem to be too small to be detectable in the laboratory by present-day interferometers. Some detailed derivations are given in Sect.6.

2.2.2. Effect of acceleration & gravity

Using the neutron interferometer, the effect of the earth's gravitational field [eq. (2.2.6) for $\vec{a} = \vec{\omega} = 0$] has been measured by Colella, Overhauser, and Werner (1975). This is usually called the COW-experiment. Bonse and Wroblewski (1983) have obained exactly the same interference pattern in a reference frame accelerated with

$\vec{a} = \vec{g}$ [for $\vec{\omega} = 0$ and no influence of gravity] (see Fig.8). Assuming the equality of inertial and gravitational mass, $m_i = m_g$, this experiment proves the complete equivalence of acceleration and gravity as reflected in (2.2.4). Nevertheless, the effect remains mass dependent. This will be important for the subsequent discussion of the strong equivalence principle for matter fields.

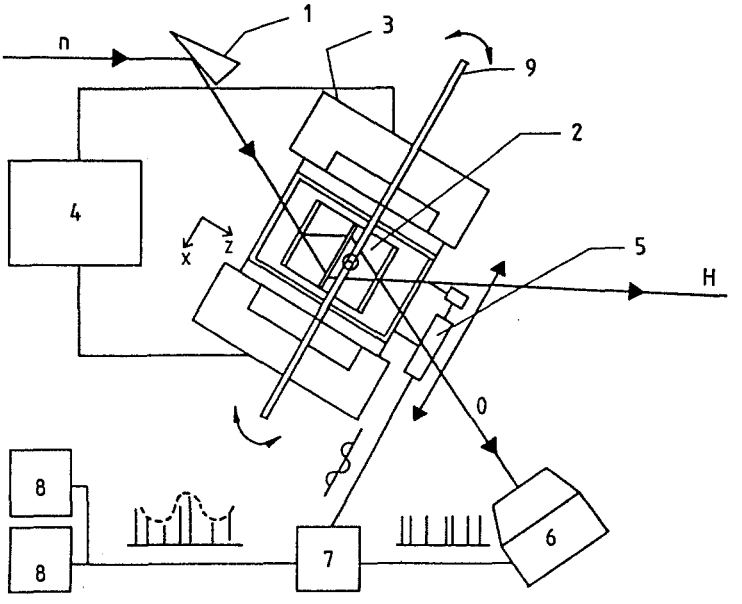

Figure 8. Experimental setup of Bonse and Wroblewski (1983). *n*: incoming neutron beam; 1: fore crystal; 2: interferometer on traverse; 3: loudspeaker magnets; 4: function generator; 5: position transducer; 6: neutron detector measuring the intensity; 7: position-to-pulse-height converter; 8: pair of single channel analyzer; 9: *Al* phase shifter.

Gravitational acceleration has also been observed for atomic beam interferometers by Kasevich and Chu (1991) and Shimizu, Shimizu, and Takuma (1992). Shimizu et al. measured by interference at a double slit the change of the deBroglie wavelength arising from the gain in energy during the free fall of the atoms in the earth's gravitational field (see Fig.9). In this way they tested not only the linear approximation in g but the complete expression,

$$\delta\phi = \oint E \sqrt{1 - \frac{2m_g\, g\, r}{m_i\, v_0^2}}\, dt\,, \tag{2.2.7}$$

which is a direct consequence of (2.2.3).

An important feature of this experiment is the fact that the center of mass of the wave packet moves in downward direction mass independently like a classical point particle ($\ddot{r} = -g$), i.e., in accordance with the weak equivalence principle ($m_i = m_g$). On the other hand, the double slit at the same time causes a quantum uncertainty to the vertical momentum component leading to a mass dependent interference pattern.

2.2.3. Sagnac type effect

The Sagnac effect for light has been verified by Michelson and Gale (1925). For matter wave interferometry, the Sagnac effect is a consequence of the coupling of the rotation of the reference frame to the angular momentum of the matter wave. For neutrons it was measured for the rotating earth by Werner *et al.* (1979) and for a rotating turntable by Atwood *et al.* (1984). Riehle *et al.* (1991) measured the influence of the rotation of a turntable on atomic beams. For electrons the effect was measured by Hasselbach and Nicklaus (1988, 1989, 1991).

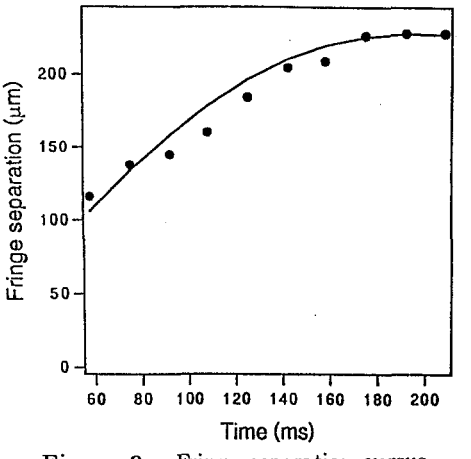

Figure 9. Fringe separation versus transit time of the atoms crossing the interferometer of Shimizu *et al.* (1992).

2.2.4. Spin-rotation coupling

On the non-relativistic level, by using matter wave interferometry, this coupling can only be measured by flipping the spin along one of the two paths. Let the length of the path, where the spin is in its flipped state, be Δl_{flip}. Then the phase shift will be

$$\delta\phi = \vec{\omega} \cdot \vec{S} \frac{\Delta l_{\text{flip}}}{v_0} . \tag{2.2.8}$$

This effect has not yet been detected. However, for neutron interferometry a phase shift of $10^{-2}\pi$ [Mashhoon 1988)] is expected and for the atomic fountain configuration of Kasevich and Chu (1991) a phase shift of π may be possible.

2.2.5. Linearity of matter field equations

In addition to the effects described by (2.2.6), we mention here a matter wave interference experiment which is important for the general structure of quantum mechanical field equations.

Into the Schrödinger equation non-linearities of the type $b \ln(|\psi|^2)\psi$, $b = const.$, have been introduced, which still allow to construct a conserved current. By bringing in attenuators at different positions into the neutron beams, Shull *et al.* (1980) found out that the parameter b, characterizing the strength of the non-linearity, has to be smaller than 4×10^{-13} eV.

3. Consequences of matter wave interferometry

Having the results of the interferometric experiments at hand, together with a simple theoretical description, we are now in a position to address the question of how to read off from these findings the appropriate description of the spacetime in which the matter fields propagate.

We will first turn our attention to the superposition principle and to the mass dependence of the phase shift of a matter field. These principles represent fundamental knowledge which is instrumental in a constructive approach to spacetime axiomatics. Then, more specifically, we will exploit the experimentally verified $\varphi_{grav}\psi$-coupling and compare it to the way a point particle couples to the gravito-inertial field. The relationship of the $\varphi_{grav}\psi$-coupling to the equivalence principle (EP) will be described in some detail.

3.1. Superposition principle

The linearity of quantum mechanics represents a very fundamental principle. In order to learn more about quantum mechanics, this principle has been questioned by Shimony (1979) and Weinberg (1989), amongst others. The experimental results, described in Sect.2.2.5, imply that matter waves are governed by linear field equations. Other types of experiments give even stronger estimates, see Physics Today The linearity of the field equations will be fundamental for establishing a spacetime structure by starting from quantum principles.

3.2. Mass dependence of phase shift

The phase shift in (2.2.6) is 'mass' dependent. Here 'mass' denotes a parameter m which is assigned to different types of matter fields, such as electrons or neutrons, for example.

If one used the same interferometer for different matter waves, the resulting phase shifts would differ in accordance to the 'mass' of the quantum objects. In particular, the positions of the interference fringes for different particle types can be compared without any reference to geometry. Results of this type we shall use in Sect.4.3.

What in point mechanics is called 'mass', attributes to a quantum object a certain property related to its wave character. The parameter m is essentially a proportionality factor between the phase shift $\delta\phi$ and the gravito-inertial field (multiplied by the interferometer area). Or, turning the argument around, one may *define* the mass of a matter wave by this kind of experiment.

3.3. Structure of the gravito-inertial coupling to the matter field

We now turn to the more specific form of the coupling to the gravito-inertial field. As shown above, it has been experimentally verified that the $\varphi_{grav}\psi$-coupling appropriately describes the behavior of a matter wave in an external gravitational field.

The Pauli equation (2.2.1), however, which we used in Sect.2.1.1 for the derivation of the general formula (2.2.6) for the phase shift, represents only the non-relativistic approximation of the gravitationally coupled Dirac equation. Consequently, we have to understand the coupling of gravity to the Dirac equation and how this is related to the way gravity influences the motion of a point particle according to Einstein's heuristic derivation of general relativity theory (GR).

3.3.1. Point particles and the $m\vec{\nabla}\varphi_{grav}$-coupling

Let us have a look at Table 1 (see next page). Consider a mass point with mass m and velocity $u^i := dx^i/ds = \gamma(1, \vec{v})$, where $\gamma := 1/\sqrt{1-v^2}$. In the flat Minkowski spacetime M_4, the point particle's force-free motion in an inertial frame of reference (i.e. in Cartesian coordinates) is governed by the equation

$$m\frac{du^i}{ds} \stackrel{*}{=} 0\,. \tag{3.3.1}$$

Observe that for the description of a world line $x^i = x^i(s)$ of a point particle (holonomic) coordinates x^i with $i, j, k \ldots = 0, 1, 2, 3$ is all we need from the geometrical backgroud. The natural (or coordinate) basis ∂_i, which is linked to the coordinates, will, however, not be sufficient for describing, say, a spinor field ψ in a non-inertial frame. Then we must turn to an orthonormal frame $\mathbf{e}_\alpha$ with (anholonomic) indices $\alpha, \beta, \gamma \ldots = 0, 1, 2, 3$. In general, $\mathbf{e}_\alpha$ will not be a natural frame, that is, the tetrad coefficients $e^i{}_\alpha$ in the decomposition $\mathbf{e}_\alpha = e^i{}_\alpha\, \partial_i$ will no longer be integrable. We will come back to this question in Sect.3.3.3.

The star on top of the equality sign in (3.3.1) means that the relation is valid only with respect to the specific basis under consideration, here an inertial reference frame represented by a (natural) Cartesian coordinate frame. In a non-inertial frame, eq.(3.3.1) becomes:

$$m\frac{du^i}{ds} + m\left\{ {i \atop jk} \right\} u^j u^k = 0\,. \tag{3.3.2}$$

The inertial forces emerge as additional terms $m\left\{ {i \atop jk} \right\} u^j u^k$ bilinear in the velocity u^i of the particle. For small velocities $v \ll 1$ and a static metric g_{ij}, which deviates from its inertial values $\eta_{ij} \stackrel{*}{=} \mathrm{diag}(-1, 1, 1, 1)$ only weakly, (3.3.2) yields

$$m\frac{d\vec{v}}{dt} + m\vec{\nabla}\left(\frac{g_{00}}{2}\right) \approx 0\,. \tag{3.3.3}$$

Table 1.

	Einstein's approach	**gauge approach**
elementary object in SR	mass point m	Dirac spinor $\psi(x)$
inertial frame	Cart.coord. x^i $ds^2 \stackrel{*}{=} \eta_{ij}\, dx^i dx^j$	orthon. hol. tetrads $e_\alpha = \delta^i_\alpha \partial_i, \quad e_\alpha \cdot e_\beta = \eta_{\alpha\beta}$
force-free motion in IF	$\frac{du^i}{ds} \stackrel{*}{=} 0$	$(i\gamma^i \partial_i - m)\psi \stackrel{*}{=} 0$
non-inertial frame	arb. curvilinear coord. $x^{i'}$	orthon. anhol. tetrads $e_\alpha = e^i{}_\alpha \partial_i$ coframe $\vartheta^\alpha = e_i{}^\alpha dx^i$
force-free motion in NIF	$\frac{du^i}{ds} + \left\{ {i \atop jk} \right\} u^j u^k = 0$	$[i\gamma^\alpha e^i{}_\alpha (\partial_i + \Gamma_i) - m]\psi = 0$ $\Gamma_i := \frac{1}{4}\Gamma_i^{\beta\gamma} \gamma_\beta \gamma_\gamma$
non-inertial objects	$\left\{ {i \atop jk} \right\}$ 40	$\vartheta^\alpha, \quad \Gamma^{\alpha\beta} = -\Gamma^{\beta\alpha}$ 16 + 24
constraints in SR	$R(\partial\{\}, \{\}) = 0$ 20	$T(\partial e, e, \Gamma) = 0, \quad R(\partial\Gamma, \Gamma) = 0$ 24 + 36
global IF	$g_{ij} \stackrel{*}{=} \eta_{ij}\,, \quad \left\{ {i \atop jk} \right\} \stackrel{*}{=} 0$	$(e_i{}^\alpha,\ \Gamma_i^{\alpha\beta}) \stackrel{*}{=} (\delta_i^\alpha, 0)$
switch on gravity	$R \neq 0$ *Riemann*	$T \neq 0, \quad R \neq 0$ *Riemann − Cartan*
local IF	$g_{ij}\vert_P = \eta_{ij}\,, \quad \left\{ {i \atop jk} \right\}\vert_P = 0$	$(e_i{}^\alpha,\ \Gamma_i^{\alpha\beta})\vert_P = (\delta_i^\alpha, 0)$
field equations	$Ric - \frac{1}{2} tr(Ric) \sim mass$	$Ric - \frac{1}{2} tr(Ric) \sim mass$ $Tor + 2\ tr(Tor) \sim spin$

Accordingly, in Newtonian parlance, the Christoffel symbols $\left\{ {i \atop jk} \right\}$ subsume the gravitational field strength whereas the metric $g_{00}/2$, with a suitable additive constant, represents the gravitational potential.

In Newtonian mechanics, the equation of motion for a point particle with inertial mass m_i and gravitational mass m_g reads

$$m_i \frac{d\vec{v}}{dt} = -m_g \vec{\nabla} \varphi_{grav} \,. \tag{3.3.4}$$

If and only if the proportionality (in suitable units: equality) of inertial and gravitational mass is valid (weak equivalence principle, or weak EP), then (3.3.3) and (3.3.4) coincide, provided we have approximately

$$g_{00} \approx -1 + 2\,\varphi_{grav} \,. \tag{3.3.5}$$

This line of reasoning is represented in the 2nd column of Table 1 and, for historical reasons, named as "Einstein's approach". As is evident from (3.3.3) and (3.3.4), the $m\vec{\nabla}\varphi_{grav}$-coupling or, synonymously, the $m\{\,\}$-coupling, can be considered as its characteristic feature. And then the weak EP directly yields the universality of the free-fall, that is, the mass independence of the equation of motion (3.3.4) or (3.3.2), respectively; compare the discussion in Einstein (1955).

It should be clear, however, that even if the equation of motion is mass independent, the (time-independent) Hamilton-Jacobi equation for the point particle in the gravitational field

$$-\frac{\left(\vec{\nabla} S\right)^2}{2m_i} - m_g \vec{g} \cdot \vec{r} = E \tag{3.3.6}$$

does depend on the mass also after the application of the weak EP (Dehnen, private communication). The analogous effect is then expected to occur in Schrödinger type equations for matter fields.

3.3.2. Different equivalence principles

We applied in (3.3.4) the weak EP to a point particle. However, it should also be possible to formulate it for matter fields. Since a point particle is localized whereas a field is spread over spacetime, the EP has to be discussed separately for these two cases. Conventionally, an EP is called *weak*, if, within some theoretical framework, it leads to the universality of free fall (see Sect.3.3.1). It is called *strong*, if it implies a special form for the equation of motion of a point particle and for dynamical equations in general. We will use this terminology.

If we abstract from the Newton-Einstein type of equation of motion of Sect.3.3.1, then for a classical point particle the weak and the strong EPs read, respectively:

(i) In the absence of any interaction other than gravitation, point particles, with the same prescribed velocity in some point of spacetime, move along the same

path irrespective of their mass. This gives rise to a *path structure* of spacetime according to [cf. Ehlers and Köhler (1977) and Coleman and Korte (1980)]

$$\dot{u}^i + H(x,u) = \alpha\, u^i\,, \qquad u^i := \frac{dx^i}{d\lambda}\,, \tag{3.3.7}$$

for some parameter λ and some function α. In the context of Newtonian physics, this means the equality of inertial and gravitational mass.

(ii) Strong EP or the equivalence of gravity and acceleration: Locally, the acceleration caused by gravity can be transformed to zero for a point particle provided there are no fields present other than gravity. In other words, locally the point particle is not accelerated in some specific coordinate system and for some parametrization of the path: $\dot{u}^\mu \stackrel{*}{=} 0$. In the general-relativistic case, this leads to a *projective structure*,

$$\dot{u}^i + \Gamma_{jk}{}^i u^j u^k = \alpha\, u^i \tag{3.3.8}$$

for some connection $\Gamma_{jk}{}^i$.

3.3.3. Matter waves and minimal coupling

Turning now to matter waves, we will follow up the discussion of the 'gauge approach' column in Table 1. Taking a Dirac field in the Minkowski spacetime M_4 as the generic case, we have to study the Dirac equation in inertial and in non-inertial frames according to the standard formalism [cf.McCrea (1987, 1989)]:

A *spinor* field is linked essentially with the notion of orthonormality, since it derives ultimately from the representations of the Lorentz group. By contrast, a tensor can easily be generalized to linear transformations. To define a spinor field in an M_4, we need an *orthonormal reference frame* at each event, i.e. a basis of four vectors $\{\mathbf{e}_\alpha\}$ $(\alpha = 0,1,2,3)$ such that

$$g_{\alpha\beta} := g(\mathbf{e}_\alpha, \mathbf{e}_\beta) = \mathrm{diag}(-1,1,1,1). \tag{3.3.9}$$

The vector basis $\mathbf{e}_\alpha$ can be decomposed with respect to the tangent vectors ∂_i of the coordinate lines according to $\mathbf{e}_\alpha = e^i{}_\alpha\, \partial_i$. The 1-form basis ϑ^α will be defined, in the usual way, by

$$\vartheta^\beta(\mathbf{e}_\alpha) = \delta^\beta_\alpha. \tag{3.3.10}$$

Its decomposition reads $\vartheta^\beta = e_j{}^\beta\, dx^j$. In an M_4, both torsion and curvature vanish,

$$T^\alpha := d\vartheta^\alpha + \Gamma_\beta{}^\alpha \wedge \vartheta^\beta = 0, \qquad R_\alpha{}^\beta := d\Gamma_\alpha{}^\beta - \Gamma_\alpha{}^\gamma \wedge \Gamma_\gamma{}^\beta = 0, \tag{3.3.11}$$

with $\Gamma_\alpha{}^\beta = \Gamma_{i\alpha}{}^\beta dx^i$ as the connection 1-form and $\Gamma^{(\alpha\beta)} = 0$ (metricity). Therefore there exist global frames for which

$$\Gamma^{\alpha\beta} = 0 \tag{3.3.12}$$

and

$$d\vartheta^\alpha = 0. \tag{3.3.13}$$

These are the so-called inertial frames. By (3.3.13), such frames are holonomic in an M_4, i.e. there exist coordinates systems $\{x^i\}$ (spatial Cartesian coordinates + time) such that

$$\vartheta^\alpha = \delta^\alpha_i dx^i. \tag{3.3.14}$$

Thus, in an M_4, the Cartesian coordinate bases already provide us with the global orthonormal frames necessary for the description of spinor fields. However, for the transition to gravitational theory via the strong EP, we have to use non-inertial frames and these must be anholonomic ($d\vartheta^\alpha \neq 0$ and $\Gamma^{\alpha\beta} \neq 0$) if they are to remain orthonormal.

Let be given the Dirac equation in an inertial frame of reference $(i\gamma^i \partial_i - m)\psi \stackrel{*}{=} 0$. The γ^i denote the Dirac matrices fulfilling $\gamma^{(i}\gamma^{j)} = \eta^{ij}$. The important step consists in the transition to a non-inertial frame. And here we take recourse to the COW-experiment [Colella et al.(1975)] and to the BW-experiment [Bonse & Wroblewski (1983)], see the discussion in Sect.2.2.2. The COW-experiment verifies the $\varphi_{grav}\,\psi$-coupling of the Newtonian potential to the matter field, as given in (2.2.2). The BW-experiment, on the other hand, shows that the gravitational field in (2.2.2) can be simulated by means of a linear acceleration. The corresponding $(m_i \vec{a} \cdot \vec{r})\psi$-term, however, had been derived by formulating the Pauli equation with respect to an accelerated frame. Consequently, this procedure of transforming a matter field equation from an inertial into a non-inertial frame of reference has been verfied by COW in conjunction with BW. Needless to say that also the Sagnac term in (2.2.2), which we won by evaluating the Pauli equation on a turntable, belongs to the established results of matter wave interferometry, see Sect.2.2.3. In other words, the 'Mashhoon'-term is the only hypothetical one in (2.2.2).

Returning to the Dirac equation, we transform it into a non-inertial frame in an analogous way as we did it for the Pauli equation, namely by rotating the local frames $\mathbf{e}_\alpha$ into a non-inertial position. Thus we relax the conditions (3.3.12) and (3.3.13). This results in (see Table 1):

$$\left[i\gamma^\alpha e^i{}_\alpha \left(\partial_i + \frac{1}{4}\Gamma_i{}^{\beta\gamma}\gamma_\beta\gamma_\gamma\right) - m\right]\psi = 0\,. \tag{3.3.15}$$

We recognize that the gravitational potentials $e^i{}_\alpha$ (or $e_j{}^\beta$) and $\Gamma_i{}^{\beta\gamma}$, which become manifest in non-inertial frames, deviate from their inertial values

$$\left(e_i{}^\beta,\, \Gamma_i{}^{\beta\gamma}\right) \stackrel{*}{=} \left(\delta^\beta_i,\, 0\right). \tag{3.3.16}$$

In other words, the $\left(e_i{}^\beta,\, \Gamma_i{}^{\beta\gamma}\right)$ describe the gravito-inertial field, or rather its potentials. In an M_4 the potentials can be 'trivialized' globally, since both torsion T^β and curvature $R^{\beta\gamma}$ vanish.

The leading additional terms, picked up by the Dirac equation in a non-inertial frame with $e^i{}_\alpha = \delta^i_\alpha + h^i{}_\alpha$ and $h^i{}_\alpha$, $\Gamma_i{}^{\beta\gamma} \ll 1$, read,

$$i\gamma^\alpha\Big[h^i{}_\alpha\,\partial_i + \delta^i_\alpha\frac{1}{4}\Gamma_i{}^{\beta\gamma}\gamma_\beta\gamma_\gamma\Big]\psi \approx \Big(ih^i{}_\alpha\,\gamma^\alpha\Big)\partial_i\psi + \Big(\frac{i}{4}\Gamma^{\alpha\beta\gamma}\gamma_\alpha\gamma_{[\beta}\gamma_{\gamma]}\Big)\psi\,. \qquad (3.3.17)$$

It is of the general type of coupling the matter wave function ψ to the gravitational potentials. In non-relativistic approximation, compare, for instance, Hehl *et al.* (1990, 1991) or Lämmerzahl (1991), it degenerates to the $\varphi_{grav}\psi$-coupling of (2.2.2). And this coupling has been experimentally verified, as we saw in Sect.2.2.2.

Instead of putting the Dirac equation into a non-inertial frame, as in Table 1 or in (3.3.15), we start in Sect.5 directly with the ordinary special-relativistic Dirac *Lagrangian.* In non-inertial frames, the Lagrangian picks up terms of the type given in (3.3.17). Thus for fields, the strong EP can be formulated as the *principle of minimal gravitational coupling* to the matter Lagrangian. It is no longer necessary to speak about objects which one observes directly, like the point particles in the Einstein approach, but rather about the corresponding material Lagrangian. Apparently, this version of the strong EP is general enough for accommodating matter fields carrying spin and the corresponding equations of motion for spinning particles [see Sexl & Urbantke (1983)]. A detailed presentation of this principle of minimal coupling has first been given by Sciama (1962).

To sum up: We only need to know the behavior of a (first order) Lagrangian in a *non-inertial reference frame,* then the coupling to gravity is determined. Violations of the strong EP would require the existence of non-minimal (Pauli type) terms in the Lagrangian containing the gravitational field strengths torsion and/or curvature explicitly. This ends our heuristic considerations.

4. Constructive axiomatic approach to spacetime geometry

We now turn to the first of the two independent procedures for establishing the geometrical stucture of spacetime: the constructive axiomatic approach to spacetime geometry using elements of quantum mechanics.

The aim of a constructive axiomatics is to discover and to describe the geometrical structure of spacetime by means of the behavior of appropriately selected physical systems, called *primitive objects,* and of particular physical effects, called *basic experience.* The intended final theorem is of the type: "If spacetime is the entity which dictates the particular primitive objects their typical behavior, then spacetime mathematically is ..." The method thereby is to enrich the manifold step by step with mathematical structures read off from experience. The postulates used must be formulated in a geometry-free manner. Our procedure will be analogous to the one followed by Ehlers, Pirani, and Schild (1972), who have used free point particles and light rays as primitive objects. Instead, we will use matter fields $\psi : \mathcal{M} \to \mathbf{C}^n$ as primitive objects. Mass and spin, as their degrees of freedom, will be essential in our scheme in order to establish the full richness of spacetime geometry. In the

following we will briefly summarize a simplified version of the scheme as developed by Audretsch and Lämmerzahl (1991a, b) and Lämmerzahl (1990):

4.1. Establishing the matter field equation

Because it is difficult to operationally justify from basic observations particular field equations for quantum objects (like the Dirac equation), we will start from the fundamental experience related to the dynamics of quantum matter fields. We will derive a general partial differential equation governing the dynamics of the matter field considered. It will turn out that the structure of this field equation is essentially determined by demanding a deterministic evolution with finite propagation speed as well as a superposition principle.

(i) We postulate a deterministic evolution of the field as an ordered behavior 'in time': There is a (1+3)-slicing of the 4-dimensional manifold $\mathcal{M}$ with monotonically increasing parameter t such that, given a field on some hypersurface, the field will be determined uniquely on a subsequent or 'later' hypersurface. The hypersurfaces for which this statement holds true are called spacelike.

(ii) For introducing the superposition principle, we require the evolution of an arbitrary sum of initial data to result in the sum of the separately propagated fields. Hence the evolution must be linear. One finds an abstract Cauchy problem $\frac{d}{dt}\psi_t = G_t\psi_t$, where ψ_t is the field ψ for fixed t and G_t the generator of the dynamical evolution. If as initial data derivatives of the field are needed in order to uniquely determine the field on a 'future' hypersurface, then we arrive at a higher order Cauchy problem, see Audretsch and Lämmerzahl (1991a).

(iii) According to experience, signals cannot propagate with infinite velocity. Therefore for all initial data with compact support we demand that, after some time, the propagated field still has compact support. This requirement implies for the generator G_t to be local. This has the important mathematical consequence that the evolutionary system reduces to a partial differential equation of first order

$$i\gamma^i(x)\partial_i\psi(x) - M(x)\psi(x) = 0\,, \tag{4.1.1}$$

where γ^i and M are some complex $n \times n$ matrices (not necessarily Dirac matrices). In addition, this first order system can be shown to be weakly hyperbolic, that is, the spacelike hypersurfaces are non-characteristic and all zeros of the characteristic equation $H(x,k) := \det(\gamma^i k_i) = 0$ are real; for a weaker version of postulate (iii), see Audretsch and Lämmerzahl (1991a).

(iv) The probability interpretation of quantum mechanics is based on a real current j^i which is bilinear in the fields. Its zeroth component j^0 is interpreted as probability density for finding a particle at a certain location. The only object in our theory which carries a contravariant vector index is γ^i. Therefore we require $j^i(x) = \psi^+ A\gamma^i\psi$ to be real for some matrix A and for all ψ's. This implies that $A\gamma^i$ is hermitian: $(A\gamma^i)^+ = A\gamma^i$.

4.2. Establishing the conformal structure

(v) The shock waves (that is, the singularities or characteristics) of the field equation (4.1.1) represent our first class of primitive elements. Jumps of lowest order along a hypersurface $\phi = 0$ obey

$$0 = A\gamma^i k_i a\,, \tag{4.2.1}$$

with some $a \in \mathbf{C}^n$ describing the helicity states on the hypersurface and $k_i = \partial_i \phi$. The solvability condition of (4.2.1) is the characterstic polynomial

$$H_c(x,k) = \det(A\gamma^i k_i) = g^{i_1 \cdots i_n} k_{i_1} \cdots k_{i_n} = 0\,, \tag{4.2.2}$$

with some real tensor $g^{i_1 \cdots i_n}$. Postulate (v) formalizes the experience that there is only one light cone at any point x (that is one future and one past light cone) and there are only two helicity states. The latter imply that the multiplicity of the zeros of the characteristic polynomial H_c must be two:

$$H_c(x,k) = \big(H_0(x,k)\big)^2\,. \tag{4.2.3}$$

The uniqueness of the light cone leads to

$$H_0(x,k) = g^{ij}(x) k_i k_j = 0\,. \tag{4.2.4}$$

Accordingly, $n = 4$ (that is, ψ has 4 complex components) and γ^i and M are 4×4-matrices.

The important consequence is that there must exist a class of real second rank tensors $g^{ij}(x)$. Because elements of this class of g^{ij} are fixed by the procedure given above only up to a positive scalar function, we are led to the notion of a conformal structure.These g^{ij} can be proven to be non-singular and to have signature ± 2. By means of the conformal structure we can construct orthotetrads e^i_a fulfilling $g_{ij} e^i_\alpha e^j_\beta = \eta_{\alpha\beta}$, where $\eta_{\alpha\beta}$ is the Minkowski metric. Therefore we are able to represent Lorentz-transformations.

4.3. Establishing the Riemannian structure

(vi) In the next step we return to (4.1.1). We select a special class of matter wave solutions, so-called approximate plane wave solutions, by making, according to the WKB-procedure, the ansatz $\psi = a \exp(iS)$. We demand that derivatives of a, that is, variations in the amplitude, are negligible. We then arrive at

$$0 = (\gamma^i p_i - M^{(0)})a\,, \tag{4.3.1}$$

$$i\gamma^i \partial_i a = M^{(1)} a\,, \tag{4.3.2}$$

for some 4×4-matrices $M^{(0)}$ and $M^{(1)}$. The solvability condition of the first equation gives a polynomial of fourth order in p, the Hamilton-Jacobi equation: $H(x,p) = (g^{ij} p_i p_j)^2 + \mathcal{O}(p^3) = 0$.

The subclass of *free* matter waves obeying (4.3.1) and (4.3.2) will be our second type of primitive elements. The property 'free' is represented by the requirement that $H(x,p) = 0$ exhibits the symmetry given by the conformal structure. Accordingly, if a momentum p is solution of the given Hamilton-Jacobi equation, then another momentum p', which results from the first one by an active Lorentz transformation $p' = Lp$, should also solve the Hamilton-Jacobi equation. By means of the fundamental theorem for vector invariants it follows that $H(x,p)$ must be of the form

$$H(x,p) = \left(g^{ij}p_ip_j - V_1(x)\right)\left(g^{ij}p_ip_j - V_2(x)\right), \tag{4.3.3}$$

with two *scalar mass functions* V_1 and V_2 and the metric g^{ij} as introduced above. Then $H_\kappa(x,p) = g^{ij}p_ip_j - V_\kappa(x) = 0$, $\kappa = 1,2$ gives the equation of motion for the group velocity $v^i = g^{ij}p_j$ of a wave packet. The mass functions $V_\kappa(x)$ turn out to be real.

(vii) Up to now it is not excluded that different types of quantum objects (denoted by the index λ), which all obey the aforementioned requirements, may lead to different scalar mass functions $V_{\lambda_\kappa}(x)$. Based on experience with matter wave interferometry, we require for free matter waves the following: For the same physical set-up (the same interferometer apparatus under identical conditions), we perform interference experiments at all points of spacetime with different quantum objects, such as electrons, neutrons, etc.. Then the pattern of the interference fringes, up to a constant factor, must be identical. This means that $V_{\lambda_\kappa} = m^2_{\lambda_\kappa}V_{1_1}$, with $m_{\lambda_\kappa} = const.$. Therefore dividing H_{λ_κ} by V_{1_1} and introducing $\bar{g}^{ij} := \frac{1}{|V_{1_1}|}g^{ij}$, we find $H_{\lambda_\kappa}(x,p) = \bar{g}^{ij}p_ip_j - m^2_{\lambda_\kappa}$. Causality requires m^2_κ to be positive. With $\bar{g}^{ij}$ we arrived at a *Riemannian metric*. Note that this does not mean that torsion is vanishing, it has simply not yet been established.

4.4. Establishing axial torsion

In the aforementioned reasoning we have used so far the properties of the Hamilton-Jacobi equation only. Using further properties of the matter fields, as displayed in equation (4.3.2) governing the differential behavior of the amplitude a, it is possible to introduce torsion [see e.g. Lämmerzahl (1990), compare Audretsch and Lämmerzahl (1987)]. For this purpose, from (4.3.2), we can derive an equation of motion for the amplitude a, which is of the form $v^i\partial_i a = v^i\Gamma_i(x)a$. In addition, we can show that the γ-matrices obey the Clifford algebra rule $\gamma^{(i}\gamma^{j)} = g^{ij}$ which the usual Dirac algebra can be derived from. Then one can prove that the only independent bilinear expressions are the probability current $\bar{\psi}\gamma^i\psi \sim v^i$ and the spin current $\bar{\psi}\gamma_5\gamma^i\psi$. By means of the propagation equation for the amplitude it follows that the propagation equation for the spin current reads

$$v^i(\overset{\{\}}{D}_iS^k + \epsilon_{ij}{}^{kl}K_lS^j) \sim S^k \tag{4.4.1}$$

for some axial vector K. Thus the propagation of the spin vector introduces an axial torsion.

To sum up, we have shown: *If spacetime is the entity which prescribes the behavior of the characteristics, of the free matter waves, and of the spin states in the way specified above, then spacetime is a Riemann-Cartan spacetime with axial torsion.*

By building wave packets out of free matter waves, it is possible to obtain the paths of the maxima of the wave packets thus introducing a path structure. These paths are the geodesics

$$v^j \partial_j v^i + \left\{ \begin{matrix} i \\ jk \end{matrix} \right\} v^j v^k = \alpha\, v^i \tag{4.4.2}$$

of a Riemannian spacetime with the metric $\bar{g}_{ij}$ defined above. Indices are moved with $\bar{g}^{ij}$ and its inverse and the Christoffels are also built from this metric. Hence the equation of motion (4.4.2) is, with respect to the Christoffel connection, the same for all types of quantum objects. Therefore it defines a Riemannian spacetime. It demonstrates that the results of the axiomatics of Ehlers, Pirani, and Schild (1972) and, in addition, the restriction from Weyl geometry to a Riemann geometry [Audretsch (1983), and Audretsch, Gähler, and Straumann (1984)] is obtained as limiting case of our axiomatics based on matter fields.

In retrospect we can specify the elements of quantum mechanics which the appraoch is based on: They essentially agree with the elements which are necessary to physically describe matter wave interference.

5. Gauge approach to spacetime geometry

Having in the last section been led to a specific spacetime geometry by the axiomatic approach, which is itself based on experience extracted from matter wave interferometry, we are now turning our attention to a gauge approach of gravity. These considerations will be independent from those of Sect.4. However, a gauge approach is fundamentally based on the notion of a matter field and its invariance properties. In other words, the notion and the existence of matter waves (or fields) is the connecting element of both approaches. In this sense, they are *not* independent but rather both based on the quantum mechanical ψ-field. Accordingly, it is not by chance that the letter ψ (with an analogous meaning) features in Sect.4 as well as in Sect.5. In the following we will resume the considerations of Sect.3.3.

Soon after Yang and Mills (1954) coupled the conserved isospin current and its Noether-related SU(2)-invariance to their newly introduced B-gauge-field, Utiyama (1956, 1980) extended the Yang-Mills idea to other non-Abelian groups [see O'Raifeartaigh (1979)] and applied it, in particular, the the Lorentz group SO(1,3). In the context of 'gauging' the Lorentz group, Utiyama was able, using some additional hypotheses, to recover general relativity (GR). Since in Newton-Einstein gravity the source of the gravitational field is the mass, i.e. the momentum current, and the corresponding symmetry the *translation* invariance, clearly a gauging of the full Poincaré

group as the semidirect product of the translation group T_4 and the Lorentz group SO(1,3) was desirable. This was carried out by Sciama (1962) and Kibble (1961). They found that the Riemannian spacetime of GR must be enriched by a *torsion field* [Cartan (1986)] such that the connection remains metric-compatible, in other words the length of a vector, as in general relativity, still stays constant under parallel transport.

In the Sciama-Kibble approach [see also Trautman (1973, 1980) and Hehl *et al.* (1976)], which used structures investigated earlier by Cartan (loc.cit.), the analog of the Hilbert Lagrangian of the underlying Riemann-Cartan spacetime was used in the gravitational action function. As a consequence, the torsion field is confined to matter, i.e. in vacuo the spacetime remains Riemannian, as in general relativity. Subsequently gravitational models with propagating torsion were proposed, leading finally to a general framework of a Poincaré gauge theory with a gravitational Lagrangian quadratic in torsion and curvature.[3]

Here we will merely sketch the appropriate gauge field-theoretical formalism for a matter field, represented by a spinor- or tensor-valued p-form, interacting with the gravitational potentials $(\vartheta^\alpha, \Gamma^{\alpha\beta})$ of (3.3.16) using the calculus of exterior differential forms [see Thirring (1986)].

5.1. Poincaré invariance in Minkowski spacetime

Let us firstly suppose that there is no gravitational field and that consequently the spacetime is Minkowskian, or M_4. In an M_4 the group of motions is the $(4+6)$ parameter Poincaré group, which is generated by translations and Lorentz rotations. The state of a particle is associated with an irreducible unitary representation of the Poincaré group and is characterized by its mass and spin, as well as by its momentum.[4] If one realizes a representation of the Poincaré group by means of a matter field over a Minkowskian spacetime, the matter field $\psi(x)$ transforms as a spinor or tensor under Lorentz transformations, depending on whether we are dealing with a fermion or a boson, respectively. The spinorial matter field is fundamental — leptons and quarks are described in this way — and its characteristic features are essential in our later considerations.

---

[3] One may consult in this context the articles of Goenner (1987), Ivanenko & Sardanashvily (1983), Kibble & Stelle (1986), Kopczyński (1990), Lord & Goswami (1986), Meyer (1982), McCrea (1987, 1989), Mielke (1987), Ne'eman (1980), Ne'eman & Regge (1978), Nester (1984), and Hehl (1980).

[4] Compare, for example, the very clear and intuitive description of Sexl and Urbantke (1982).

5.2. First order Lagrangian of a matter field

The dynamical behaviour of fermionic and bosonic matter in an M_4 is determined by their Lagrangian 4-form L which we here refer to Cartesian coordinates and hence to the inertial frames of (3.3.14),

$$L \stackrel{*}{=} L(\psi, d\psi)\,, \tag{5.2.1}$$

where the p-form ψ is the matter field. In the sense of conventional Lagrangian field theory, L may depend at most on first derivatives of the matter field ψ. Via the action principle one finds the matter field equation depending at most on second derivatives:

$$\frac{\delta L}{\delta \psi} \equiv \frac{\partial L}{\partial \psi} - (-1)^p d(\frac{\partial L}{\partial d\psi}) = 0\,. \tag{5.2.2}$$

For an isolated system, i.e. if external fields do not act, the action function associated with (5.2.1) is invariant under Poincaré transformations. This implies, via Noether's theorem, the conservation of momentum and angular momentum:

$$d\Sigma_\alpha = 0. \qquad d\tau_{\alpha\beta} - \vartheta_{[\beta} \wedge \Sigma_{\alpha]} = 0\,. \tag{5.2.3}$$

where

$$\Sigma_\alpha = \mathbf{e}_\alpha \rfloor L - (\mathbf{e}_\alpha \rfloor d\psi) \wedge (\frac{\partial L}{\partial d\psi}) - (\mathbf{e}_\alpha \rfloor \psi) \wedge (\frac{\partial L}{\partial \psi}) \tag{5.2.4}$$

is the canonical (and asymmetric) energy-momentum 3-form and

$$\tau_{\alpha\beta} = G_{[\alpha\beta]}\psi \wedge (\frac{\partial L}{\partial d\psi}) \tag{5.2.5}$$

is the canonical spin angular momentum 3-form. If ψ is a 0-form then $\mathbf{e}_\alpha \rfloor \psi = 0$ and the last term of (5.2.4) drops. In equation (5.2.5) $G_{[\alpha\beta]}$ are the spin generating operators. The "inertial currents" $(\Sigma_\alpha,\ \tau_{\alpha\beta})$ represent in field-theoretical language the particle's mass and spin — accordingly the appropriate names are *momentum* current and *spin* current, respectively. As is evident from (5.2.3) as well as from the labels of the irreducible unitary representations of the Poincaré group, the inertial behavior of, say, a fermion is not only characterized by its mass and its momentum current, but also a spin concept is necessary for a complete representation of the inertial properties of the fermion.

5.3. Minimal coupling to gravity

If we now introduce non-inertial reference frames in the M_4, (5.2.1) reads

$$L = L(\vartheta^\alpha,\ \psi,\ D\psi) = L(\vartheta^\alpha,\ \Gamma^{\alpha\beta},\ \psi,\ d\psi) \tag{5.3.1}$$

with the covariant exterior derivative $D\psi := (d + \Gamma^{\alpha\beta} G_{\alpha\beta})\psi$. Tetrad and connection, but not their derivatives, appear explicitly in the Lagrangian.

As we saw in our discussion in Sect.3.3.3, the strong EP amounts to the following: Viewed locally, special relativistic matter in a non-inertial frame behaves in the same way as in a corresponding gravitational field.

To which quantity characterizing a matter field do we apply the strong EP? Certainly to the Lagrangian (5.3.1). Since the strong EP is a *local* principle we ought to apply it to the differentiation level of the lowest possible order. In general, the matter field equation (5.2.2) is of second differentiation order so that in the presence of a gravitational field curvature terms could already emerge. Moreover, we know that the Lagrangian enters explicitly the Feynman-quantization of matter fields and determines the transition amplitudes. Consequently in the theory of matter *fields* the Lagrangian possesses nearly the quality of an observable.

We stress again that in the original Einsteinian argumentation the EP has been applied directly to the acceleration or the equation of motion of a point particle, i.e. on the level of (5.2.2). We dispensed with the point particle concept and found only the level (5.2.1) or (5.3.1) suitable for the application of the EP. Consequently we indeed recognize $(\vartheta^\alpha, \Gamma^{\alpha\beta})$ as the *gravitational potentials*. According to (3.3.12) and (3.3.14), they can be globally trivialized in an M_4 since they are only induced by the choice of non-inertial reference frames.

5.4. Riemann-Cartan geometry of spacetime

Following McCrea (1989), the transition to gravitational theory is executed by requiring only the local validity of special relativity and of conditions (3.3.12) and (3.3.14):

$$(\vartheta^\alpha, \Gamma^{\alpha\beta}) \stackrel{*}{=} (\delta^\alpha_i dx^i, 0), \quad \text{locally.} \tag{5.4.1}$$

Thus we arrive at a Riemann-Cartan spacetime U_4 which is characterized by a local Minkowski metric (3.3.9) and a connection $\Gamma^{\alpha\beta}$ which is metric (compatible): $\Gamma^{(\alpha\beta)} = 0$. Its deviation from a Minkowskian M_4 is measured by the torsion T^α and the curvature $R^{\alpha\beta}$, which are both defined in (3.3.11). In fact, it is possible to prove that in a U_4, in a suitable tetrad and in suitable coordinates, condition (5.4.1) can always be fulfilled. Consequently, in such a 'local inertial system' the gravitational potentials are trivialized and can no longer be perceived, it represents the 'Einsteinian elevator' of the U_4. Now the Lagrangian (5.3.1) has its special-relativistic form only locally in a suitable tetrad. The constraints (3.3.11) may be relaxed since torsion and curvature cannot appear in the Lagrangian(5.3.1):

$$T^\alpha \neq 0, \qquad R_\alpha{}^\beta \neq 0. \tag{5.4.2}$$

Thereby we recognize torsion and curvature as *gravitational field strengths*, and a *Riemann-Cartan spacetime* U_4 emerges as appropriate for the description of gravitational phenomena. Thus, the gauge approach and the constructive axiomatic approach independently lead to the same result.

To support our line of reasoning, we add the following remarks: The strong EP is a heuristic principle, since the notion 'local' used in its formulation is not exact. In

exploiting the strong EP, we required stronger locality as compared to the one used in 'deriving' GR. The following procedure is also conceivable. We rewrite the constraint of vanishing torsion $(3.3.11)_1$ in terms of the connection

$$T^\alpha = 0 \iff \Gamma_{\alpha\beta} = \frac{1}{2}\{\mathbf{e}_\alpha \rfloor d\vartheta_\beta - \mathbf{e}_\beta \rfloor d\vartheta_\alpha - (\mathbf{e}_\alpha \rfloor \mathbf{e}_\beta \rfloor d\vartheta_\mu)\vartheta^\mu\}. \tag{5.4.3}$$

When this connection $\Gamma^{\alpha\beta} = \Gamma^{\alpha\beta}(\vartheta, d\vartheta)$ is substituted into (5.3.1), then, in non-inertial systems, L depends on ϑ^α as gravitational potential and only the constraint of vanishing curvature can be relaxed, whereas torsion remains zero. Accordingly one arrives at the Riemannian spacetime V_4 of GR. However, since it is always possible to make a tetrad and coordinate transformation at a point so that (5.4.1) is fulfilled at that point, it follows from (3.3.13) that in a V_4 the resulting tetrad satisfies the constraint $d\vartheta^\alpha = 0$ at the point, a constraint that is non-local and hence contrary to the spirit of the EP. Provided we do not wish to stay in the context of point particles and their trajectories in an 'Einsteinian elevator', there seems to be no reason for requiring the constraint $d\vartheta^\alpha = 0$. Hence everything speaks in favor of leaving the adjustment of torsion to the dynamics of the gravitational field and not to rule it out in the context of kinematics. Alternatively, one could have rewritten the constraint $(3.3.11)_2$ of vanishing curvature in terms of the connection:

$$R_\alpha{}^\beta = 0 \iff \Gamma^{\alpha\beta} \overset{*}{=} 0 \qquad \text{(globally)} \tag{5.4.4}$$

i.e. in the case of vanishing curvature there exist global parallel frames. Then, in a way analogous to above, we would have been led to a teleparallel spacetime T_4, i.e. to a U_4 with vanishing curvature. But a T_4 for fermions is no more convincing than the V_4 of GR. For a gravitational theory of *fermionic matter*, the U_4 with its 'Einstein elevator' (5.4.1), which is indispensable for a correct application of the strong EP, offers the appropriate geometrical framework, but for scalar or macroscopic matter a V_4 or a T_4 is already sufficient.

We complete our presentation with a short discussion of the appropriate Lagrange-Noether formalism. As in any other gauge theory, now that the interaction has been switched on by means of (5.4.2), the Lagrange-Noether formalism, which originally, in the 'pure gauge' case, leads to (5.2.2–5), has to be redone. By standard methods, we find in the U_4 the matter field equation

$$\frac{\delta L}{\delta \psi} = \frac{\partial L}{\partial \psi} - (-1)^p D\left(\frac{\partial L}{\partial D\psi}\right). \tag{5.4.5}$$

and the Noether identities

$$D\Sigma_\alpha = (\mathbf{e}_\alpha \rfloor T^\beta) \wedge \Sigma_\beta + (\mathbf{e}_\alpha \rfloor R^{\beta\gamma}) \wedge \tau_{\beta\gamma}\,. \qquad D\tau_{\alpha\beta} - \vartheta_{[\beta} \wedge \Sigma_{\alpha]} = 0\,. \tag{5.4.6}$$

The new definitions of the momentum current

$$\Sigma_\alpha := \frac{\delta L}{\delta \vartheta^\alpha} = \mathbf{e}_\alpha \rfloor L - (\mathbf{e}_\alpha \rfloor D\psi) \wedge \frac{\partial L}{\partial D\psi} - (\mathbf{e}_\alpha \rfloor \psi) \wedge \frac{\partial L}{\partial \psi} \tag{5.4.7}$$

and of the spin current

$$\tau^{\alpha}{}_{\beta} := \frac{\delta L}{\delta \Gamma_{\alpha}{}^{\beta}} = G^{\alpha}{}_{\beta}\,\psi \wedge \frac{\partial L}{\partial D\psi} \tag{5.4.8}$$

show that these inertial currents are coupled to the corresponding gravitational potentials.

In an M_4, the Noether laws (5.4.6) reduce to (5.2.3). The volume force densities showing up on the right-hand side of the momentum identity in (5.4.6), namely (torsion $\times$ momentum) and (curvature $\times$ spin) i.e. (field strength $\times$ current) are remarkable. There is a close analogy here with the U(1)-gauge field theory (Maxwell's theory) where the Lorentz force is exactly of this type.

These Noether laws we will use in the following section for constructing a Conserved Energy-like quantity.

6. Testing spacetime geometry by interference experiments

In Sect.2.2.1, we had discussed the experimentally verified interference effects to lowest order. Based on results of Sect.5.4. we are now in a position to continue in a more rigorous fashion in order to include possible curvature and torsion effects. This can be done using a WKB-approximation of the Dirac equation [Audretsch and Lämmerzahl (1983)] or by deriving a Hamilton operator for the energy in a stationary spacetime using the energy-momentum current of the respective field. This last mentioned procedure mainly relies on Lämmerzahl (1992) using unpublished results of Hecht (1986) and of Hehl, McCrea, Mielke, and Ne'eman (1989). For other recent work one may consult Huang (1992).

6.1. Conserved energy

Let us take the Hamilton operator method which describes the energy of the considered quantum system in a curved spacetime for rotating and accelerating interferometers. We start from the general material Lagrangian (5.3.1), which is minimally coupled to the gravitational field. We suppose that the matter field equations (5.4.5) are fulfilled.

We assume the existence of a symmetry of spacetime, that is, of a Killing vector $\xi = \xi^{\alpha}\mathrm{e}_{\alpha}$ which fulfills the symmetry conditions

$$\mathcal{L}_{\xi} g = 0\,. \qquad \mathcal{L}_{\xi}\Gamma_{\alpha}{}^{\beta} = 0\,. \tag{6.1.1}$$

We introduce the abbrevations $D_{\alpha} := \mathrm{e}_{\alpha}\rfloor D$ and $D^{\beta} := g^{\beta\gamma}D_{\gamma}$ and the transposed connection $\widetilde{\Gamma}_{\alpha}{}^{\beta} := \Gamma_{\alpha}{}^{\beta} + \mathrm{e}_{\alpha}\rfloor T^{\beta}$ together with its covariant exterior derivative $\widetilde{D}$. Furthermore, we note for later use

$$\widetilde{D}_{\alpha}\xi^{\beta} = \overset{\{\}}{D}_{\alpha}\xi^{\beta} - \xi\rfloor K_{\alpha}{}^{\beta}\,. \qquad \text{with} \qquad K_{\alpha}{}^{\beta} := \overset{\{\}}{\Gamma}_{\alpha}{}^{\beta} - \Gamma_{\alpha}{}^{\beta}\,. \tag{6.1.2}$$

Here $K^{\alpha\beta} = -K^{\beta\alpha}$ denotes the contortion of spacetime. Then the symmetry conditions (6.1.1) can be rewritten as

$$\widetilde{D}^{(\alpha}\xi^{\beta)} = 0 \quad \text{or} \quad \overset{\{\}}{D}{}^{(\alpha}\xi^{\beta)} = 0\,. \qquad \text{and} \qquad D\widetilde{D}_\alpha\xi^\beta = -\xi\rfloor R_\alpha{}^\beta\,. \tag{6.1.3}$$

Note that the last relation is equivalent to $\mathcal{L}_\xi K_\alpha{}^\beta = 0$.

With these preparations we can find the energy expression $\mathcal{E}$ according to

$$d\mathcal{E} = 0\,, \qquad \mathcal{E} := \xi^\alpha\Sigma_\alpha + \left(\widetilde{D}_\alpha\xi^\beta\right)\tau^\alpha{}_\beta\,, \qquad E := \int_\Sigma \mathcal{E} = const.\,, \tag{6.1.4}$$

where Σ_α and $\tau^\alpha{}_\beta$ denote the material momentum and spin currents of (5.4.7) and (5.4.8), respectively. The conservation law $d\mathcal{E} = 0$ can be proved by applying the Noether identities (5.4.6) and the symmetry conditions (6.1.3). We always integrate over a space-like hypersurface Σ.

The transition to quantum mechanics is achieved by defining a scalar product by means of a conserved quantity. Accordingly, we additionally assume the Lagrangian to be invariant against phase transformations. This yields the conserved quantity

$$dj = 0 \qquad j(\bar\psi, \psi, x) := \psi \wedge \frac{\partial \mathcal{L}}{\partial D\psi} \qquad Q = \int_\Sigma j(\bar\psi, \psi, x)\,. \tag{6.1.5}$$

This construction will allow the definition of an energy operator in Sect.6.2.

Aa a simple example, we display the results for a Dirac field. Its Lagrangian depends only on the axial torsion. For the momentum, the spin, and the Dirac currents we find, respectively.

$$\Sigma_\alpha = \frac{i}{2}\left(\bar\psi\,{}^*\gamma D_\alpha\psi - D_\alpha\bar\psi\,{}^*\gamma\psi\right), \quad \tau^\alpha{}_\beta = -\frac{1}{8}\bar\psi\left\{{}^*\gamma\,, \sigma^\alpha{}_\beta\right\}\psi\,, \quad j = -\bar\psi\, i\,{}^*\gamma\psi\,, \tag{6.1.6}$$

where $*$ denotes the Hodge star and $\gamma := \gamma_\alpha\,\vartheta^\alpha$ the Dirac matrix 1-form; moreover, $\sigma_{\alpha\beta} = \frac{i}{2}[\gamma_\alpha, \gamma_\beta]$.

6.2. Hamilton operator

Explicitly, we introduce a scalar product by means of the current j of (6.1.5):

$$< \psi_1 \mid \psi_2 > := \int_\Sigma j(\bar\psi_1, \psi_2, x)\,. \tag{6.2.1}$$

A classical field observable $A_{\rm cl} := \int_\Sigma j(\bar\psi, \mathcal{A}_{\rm cl}\psi, x)$ is identified with the expectation value $A_{\rm qu}$ of a quantum measurement $A_{\rm qu} = < \psi \mid \widehat{\mathcal{A}}_{\rm qu} \mid \psi > = \int_\Sigma j(\bar\psi, \mathcal{A}_{\rm qu}\psi, x)$. Here $\mathcal{A}_{\rm qu}$ denotes the position representation of the quantum operator $\widehat{\mathcal{A}}_{\rm qu}$. This identification implies $\mathcal{A}_{\rm qu} = \mathcal{A}_{\rm cl}$.

Therefore the energy operator $\mathcal{H}$, using also (6.1.4), can be defined according to

$$\int_\Sigma j(\bar{\psi}, \mathcal{H}\psi, x) := E = \int_\Sigma \mathcal{E} = \int_\Sigma \left[\xi^\alpha \Sigma_\alpha + \left(\tilde{D}_\alpha \xi^\beta\right)\tau^\alpha{}_\beta\right]$$
$$= \int_\Sigma \left[\xi^\alpha n_\alpha (n^\beta \Sigma_\beta) + \xi^\alpha (k_\alpha{}^\beta \Sigma_\beta) + \left(\tilde{D}_\alpha \xi^\beta\right)\tau^\alpha{}_\beta\right], \qquad (6.2.2)$$

where we introduced the projection tensor $k_\alpha{}^\beta := \delta^\beta_\alpha - n_\alpha n^\beta$ of the hypersurface Σ with n as its normal. We can identify $n^\beta \Sigma_\beta$ as the energy flux density and $k_\alpha{}^\beta \Sigma_\beta$ as momentum flux density. Then the corresponding terms in (6.2.2) can be assigned to a suitably symmetrized operator expression $\frac{1}{2}\{\xi^\alpha, n_\alpha \mathcal{H}_0\}$ and $\frac{1}{2}\{\xi^\alpha, \mathcal{P}_\alpha\}$ in the conventional way. Moreover, we assign to the spin flux density $\tau^\alpha{}_\beta$ the operator $s^\alpha{}_\beta$. This yields the symmetrized Hamilton operator

$$\mathcal{H} = \frac{1}{2}\{\xi^\alpha, n_\alpha \mathcal{H}_0\} + \frac{1}{2}\{\xi^\alpha, \mathcal{P}_\alpha\} + (\tilde{D}_\alpha \xi^\beta) s^\alpha{}_\beta. \qquad (6.2.3)$$

Each term in this Hamiltonian is hermitian by construction. Note that the components of the canonical spin current describe the spin flux density, the spin density, the energy-dipole-moment flux density, and the energy-diplole-moment density [cf. Hehl (1976)].

According to (6.1.2), the last term in (6.2.3) splits into a spin-contortion term and a coupling of the spin to the kinematical properties of the Killing field ξ the trajectories of which later on will be identified with the trajectories of the different pieces of the interferometer.

The velocity u^i of these pieces is proportional to the Killing vector $\xi^i = e^\chi u^i$, with χ as the gravi-stationary potential. The projection tensor $h^j_i = \delta^j_i - u_i u^j$ transforms the vorticity $\partial_{[i} u_{j]}$ into the local rotation $\omega_{ij} := h^k_i h^l_j \partial_{[k} u_{l]}$. Eventually, the 'Christoffel' curl entering (6.2.3) can be decomposed according to $\overset{\{\}}{D}_{[i}\xi_{j]} = e^\chi(\omega_{ij} + 2a_{[i}u_{j]})$, where $a_i = u^j \overset{\{\}}{D}_j u_i = -\partial_i \chi$ denotes the acceleration of the Kiling field. Consequently, the spin of the matter wave couples to the acceleration and and to the rotation of the interferometer.

In particular, for the Dirac case the Hamiltonian (6.2.3) reads

$$\mathcal{H}_0 = mc\gamma^i n_i + \frac{i}{2}\gamma^{[i}\gamma^{j]}\mathcal{P}_i n_j + \frac{1}{2}\overset{\{\}}{D}_i n^i, \quad \mathcal{P}_i = k_i{}^j \overset{\{\}}{D}_j, \quad s^{ij} = \frac{i}{8}\gamma^{[i}\gamma^{j]}. \qquad (6.2.4)$$

As special case we recover the Hamilton operator in flat Minkowski space. There, the most general Killing vector field $\xi^i = b^i + \omega^i{}_j x^j$ with $\omega_{ij} = -\omega_{ji}$ leads to

$$\mathcal{H} = \frac{1}{2}\{b^0 + \vec{a}\cdot\vec{r}, \mathcal{H}_0\}\psi - (\vec{b} + \vec{a}t)\cdot\vec{p}\psi + \vec{\omega}\cdot(\vec{L} + \vec{S})\psi, \qquad (6.2.5)$$

with the 3-acceleration $\vec{a} \leftrightarrow \omega^i{}_j n^j$ and the 3-rotation $\vec{\omega} \leftrightarrow -\epsilon^{ijkl}\omega_{jk} n_l$ and $\vec{a} \leftrightarrow \omega^i{}_j n^j$ [see Lämmerzahl (1991) and Hehl & Ni (1990)].

6.3. Phase shift

In an interference experiment the matter field is localized, that is, it is different from zero only within a comparatively small region. Hence, in (6.2.2) and (6.2.3), we can approximate both, the Killing vector, which represents the external gravitational field, and its derivative $\widetilde{D}\xi^\beta$ by a Taylor expansion. At the beam splitter we may choose the origin O of our coordinate system. The center of mass of the matter wave packet, the motion of which we will follow up, will be denoted by C. The difference 'vectors' $\Delta x := \overset{c}{x} - \overset{o}{x}$, $\delta x := x - \overset{c}{x}$ are defined to be tangent to the hypersurface Σ: $\Delta x \rfloor n = \delta x \rfloor n = 0$. Then the Taylor expansion around O reads, if we use the relations (6.1.3):

$$\xi^\alpha(x) = \xi^\alpha(\overset{o}{x}) + (\Delta x + \delta x)\rfloor(\overset{\{\}}{D}\xi^\alpha)(\overset{o}{x}) + \Delta x\rfloor\xi\rfloor\overset{\{\}}{R}_\beta{}^\alpha(\overset{o}{x})\delta x^\beta + \dots , \tag{6.3.1}$$

$$(\mathbf{e}_\alpha\rfloor\widetilde{D}\xi^\beta)(x) = (\mathbf{e}_\alpha\rfloor\widetilde{D}\xi^\beta)(\overset{o}{x}) + (\Delta x + \delta x)\rfloor\xi(\overset{o}{x})\rfloor R_\alpha{}^\beta(\overset{c}{x}) + \dots . \tag{6.3.2}$$

In (6.3.1), we used the Christoffel derivative $\overset{\{\}}{D}$ since, in comparison with D, only higher order deviations are expected to arise in (6.2.2). Moreover, according to $(6.1.3)_2$, the Christoffel derivative is appropriate for the expansion of ξ. Consequently, $(6.1.3)_3$ motivates the use of D in expanding $\widetilde{D}\xi$. In (6.3.2), $R_\alpha{}^\beta(\overset{c}{x})$ will be replaced by $R_\alpha{}^\beta(\overset{o}{x})$, since $R_\alpha{}^\beta$, in the interferometer region, is assumed to be a slowly varying field. We also neglect products of the form $\delta x\, R\, \tau$.

We substitute (6.3.1) and (6.3.2) into (6.2.3). Then, to first order in the dimension of the extension δx of the field and in the distance Δx, the Hamiltonian reads:

$$\begin{aligned}
\mathcal{H} = {} & \xi(\overset{o}{x})\rfloor n\,\mathcal{H}_0 + (\overset{\{\}}{D}_\alpha\xi^\beta)(\overset{o}{x})\left(\Delta x^\alpha n_\beta\,\mathcal{H}_0 + \frac{1}{2}\{\delta x^\alpha, n_\beta\mathcal{H}_0\}\right) \\
& + \Delta x\rfloor(\xi\rfloor\overset{\{\}}{R}_\alpha{}^\beta)(\overset{o}{x})\,\frac{1}{2}\{\delta x^\alpha, n_\beta\mathcal{H}_0\} \\
& + \xi^\alpha(\overset{o}{x})\,\mathcal{P}_\alpha + (\overset{\{\}}{D}_\alpha\xi^\beta)(\overset{o}{x})\left(\Delta x^\alpha\,\mathcal{P}_\beta + \frac{1}{2}\{\delta x^\alpha, \mathcal{P}_\beta\}\right) \\
& + \Delta x\rfloor(\xi\rfloor\overset{\{\}}{R}_\alpha{}^\beta)(\overset{o}{x})\,\frac{1}{2}\{\delta x^\alpha, \mathcal{P}_\beta\} \\
& + s^\alpha{}_\beta\Big((\overset{\{\}}{D}_\alpha\xi^\beta)(\overset{o}{x}) - (\xi\rfloor K_\alpha{}^\beta)(\overset{o}{x}) + \Delta x\rfloor(\xi\rfloor R_\alpha{}^\beta)(\overset{o}{x})\Big) + \dots .
\end{aligned} \tag{6.3.3}$$

We neglected products of the contortion with the rotation or the acceleration, respectively. The results of this and the last section were achieved recently, a more detailed analysis will be presented in a forthcoming publication.

Our procedure is especially appropriate for evaluating interference experiments. If we are going to describe an atom by such a matter field and if the external field is not so strong as to be able to extract an electron from an atom, then the matter field is always localized. The operator $\mathcal{P}_\alpha$ will be interpreted as momentum, $l^\alpha{}_\beta := \frac{1}{2}\{\delta x^\alpha, \mathcal{P}_\beta\}$ as orbital angular momentum, $\varepsilon_\alpha := n_\alpha\mathcal{H}_0$ as energy flux and $\varepsilon^\alpha{}_\beta :=$

$\frac{1}{2}\{\delta x^\alpha, n_\beta \mathcal{H}_0\}$ as energy dipole-moment. Therefore $L^{\alpha\beta} := \Delta x^{[\alpha}\mathcal{P}^{\beta]}$ represents the angular momentum and $E^{\alpha\beta} := \Delta x^{[\alpha}\varepsilon^{\beta]}$ the energy dipole-moment of the center of mass with respect to O. Furthermore, $j^\alpha{}_\beta := l^\alpha{}_\beta + \varepsilon^\alpha{}_\beta + s^\alpha{}_\beta$ is the total angular momentum and energy dipole-moment with respect to the center of mass C. Thus we find for the Hamiltonian:

$$\mathcal{H} = \xi^\alpha(\overset{o}{x})\ (\varepsilon_\alpha + \mathcal{P}_\alpha) + \mathcal{H}_{\text{int}}\,, \tag{6.3.4}$$

$$\mathcal{H}_{\text{int}} := \overset{\{\}}{D}_\alpha \xi^\beta(\overset{o}{x})\ (\ \underbrace{L^\alpha{}_\beta}_{\text{Sagnac}} + \underbrace{E^\alpha{}_\beta}_{\text{accel.}} + \underbrace{j^\alpha{}_\beta}_{\substack{\text{inertial spin}\\ \text{coupling}}}\ ) \tag{6.3.5}$$

$$+\ \underbrace{\Delta x\rfloor(\xi\rfloor \overset{\{\}}{R}_\alpha{}^\beta)(\overset{o}{x})\,(l+\varepsilon)^\alpha{}_\beta}_{\substack{\text{orbital angular momentum-}\\ \text{Riemann curvature}}} + \underbrace{\Delta x\rfloor(\xi\rfloor R_\alpha{}^\beta)(\overset{o}{x})\, s^\alpha{}_\beta}_{\text{spin-curvature}} - \underbrace{(\xi\rfloor K_\alpha{}^\beta)(\overset{o}{x})\, s^\alpha{}_\beta}_{\text{spin-torsion}}\,.$$

The first term on the right hand side of (6.3.4) describes the energy and the translational momentum of the field with respect to the Killing field. In the second term $\mathcal{H}_{\text{int}}$, according to (6.3.5), the angular momentum and the energy dipole-moment of the total system are coupled to the rotation and acceleration of the Killing field. These terms result in the Sagnac type effect and its time-like analogue, the COW-effect. Additionally, spin plus orbital angular momentum, that is, the total angular momentum j, also couple to rotation and acceleration thereby generalizing the well known spin-rotation coupling [Schmutzer (1973), Collier (1977), Audretsch and Lämmerzahl (1983), Mashhoon (1988), Hehl & Ni (1990), Lämmerzahl (1991)]. The subsequent terms mediate the coupling of the orbital angular momentum to the Riemannian curvature as well as that of the spin of the matter field to the (total) Cartan curvature. The last term exhibits an explicit coupling between the material spin current to the contortion tensor. Note that only the spin current 'feels' torsion whereas all orbital terms react only to the Riemannian curvature, as worked out by Yasskin and Stoeger (1981).

For a pure *space-like* spin current (as in the case of Dirac matter) we introduce a more convenient notation by relating $s^\alpha{}_\beta \leftrightarrow \vec{S}$, $\omega^\alpha{}_\beta \leftrightarrow \vec{\omega}$, $a^\alpha \leftrightarrow \vec{a}$, $\xi\rfloor K_\alpha{}^\beta \leftrightarrow \vec{K}$, and $\Delta x \leftrightarrow \vec{x}$. Then we find, see Table 2, the lowest order inertial, curvature, and torsion effects of the matter field Hamiltonian (6.3.4).

The Hamilton operator can be used to calculate the outcome of an interference experiment with an interferometer of small extension. A particle beam is coherently split and brought to interference after having travelled along separate paths. The interferometer is assumed to travel along a Killing trajectory. For describing such experiments, the semi-classical approximation is appropriate. Accordingly, we assume that the field equation $\delta L/\delta\psi = 0$ possesses an approximate solution of the form $\psi \approx \varphi \exp(\frac{i}{\hbar}S)$, with S as the classical phase. Then, in the classical limit, the phase

shift for one round trip is given by [see Anandan (1977), Audretsch and Lämmerzahl (1983)]

$$\delta\Phi = \frac{1}{\hbar}\oint (e^{-\chi} E\, u_i\, dx^i + P_a\, dx^a) \approx \frac{1}{\hbar}\oint E_{\rm int} dt\,, \tag{6.3.6}$$

with $a = 1,2,3$. The energy E is the eigenvalue in $\mathcal{H}\psi = E\psi$. The expression $u_i dx^i$ is to be calculated according to the particle's group velocity, and $\vec{P}$ is the canonical momentum. Because E is constant by construction, we can simplify (6.3.6) somewhat:

$$\oint e^{-\chi} E\, u_i\, dx^i = E \int e^{-\chi}\, \omega_{ij}\, dA^{ij}\,. \tag{6.3.7}$$

However, it turns out to be very useful in some experiments (with spin-rotation coupling and spin-torsion coupling) to enforce a spin-flip after splitting and before recombining the matter wave [see Mashhoon (1988)]. In these cases E can be extracted from the integral only for certain parts of the path.

$m\,(\vec{a}\cdot\vec{x})$	Redshift (Bonse-Wroblewski → COW)
$-\vec{\omega}\cdot\vec{L}$	Sagnac type effect (Page-Werner et al.)
$-\vec{\omega}\cdot\vec{S}$	Spin-rotation effect (Mashhoon)
$\vec{p}\cdot(\vec{a}\cdot\vec{x})\,\vec{p}\,/\,(2m)$	Redshift effect of kin. energy
$\vec{S}\cdot(\vec{a}\times\vec{p})\,/\,(4m)$	Inertial spin-orbit coupling
$x\rfloor\xi\rfloor \overset{\{\}}{R}_\alpha{}^\beta\,(l+\varepsilon)^\alpha{}_\beta$	Orb.ang.mom.-{}-curvature coupling
$x\rfloor\xi\rfloor R_\alpha{}^\beta\, S^\alpha{}_\beta$	Spin-curvature coupling
$\vec{K}\cdot\vec{S} = \xi\rfloor K_\alpha{}^\beta\, S^\alpha{}_\beta$	Spin-torsion coupling

Table 2. Lowest order inertial, curvature, and torsion effects for a general matter field. The interior product is denoted by $\rfloor$.

After having performed one interference experiment, we cannot uniquely attribute parts of the total phase shift to the various terms in the Hamiltonian. Rather, we have to measure phase shifts under different physical situations. These can be realized, for example, by different boundary conditions (adiabatically changing the orientation of the interferometer) or by differently preparing the quantum system (selecting a system with a certain polarization or spin component).

If we insert the energy eigenvalues into the formulas of Table 2, we find a phase shift for each term. We denote the area of a COW-type interferometer by $\vec{A}$, its height by $\vec{h}$, and its length by l. Then we have explicitly:

$$\delta\Phi_{\rm acc} = \frac{mA}{\hbar v}\, a \qquad \text{acceleration effect}$$

$$\delta\Phi_{\rm Sagnac} = \frac{2m}{\hbar}\, \vec{\omega}\cdot\vec{A} \qquad \text{Sagnac (or rotation) coupling}$$

$$\delta\Phi_{\rm s-r} = \frac{2l_{\rm tot}}{v}\, \vec{\omega}\cdot\vec{J} \qquad \text{spin-rotation coupling}$$

$$\delta\Phi_{\rm s-o} = \frac{al}{c^2}\, J \qquad \text{inertial spin-orbit coupling}$$

$$\delta\Phi_{\rm o-c} = \frac{lh^a}{2\hbar v}\, \overset{\{\}}{R}_{0a\alpha}{}^{\beta}\, J^{\alpha}{}_{\beta} \qquad \text{orb.ang.mom.-}\{\}\text{-curvature-coupling}$$

$$\delta\Phi_{\rm s-c} = \frac{lh^a}{2\hbar v}\, R_{0a\alpha}{}^{\beta}\, S^{\alpha}{}_{\beta} \qquad \text{spin-curvature-coupling}$$

$$\delta\Phi_{\rm s-t} = \frac{l_{\rm tot}}{v}\, \vec{K}\cdot\vec{S} \qquad \text{spin-torsion coupling}$$

$$\delta\Phi_{\rm red} = \frac{v^2}{2c^2}\, \delta\Phi_{\rm acc} \qquad \text{redshift of kinetic energy}$$

The total phase shift is the sum of all of these contributions:

$$\delta\Phi = \delta\Phi_{\rm acc} + \delta\Phi_{\rm Sagnac} + \delta\Phi_{\rm s-r} + \delta\Phi_{\rm s-o} + \delta\Phi_{\rm o-c} + \delta\Phi_{\rm s-c} + \delta\Phi_{\rm s-t} + \delta\Phi_{\rm red} + \ldots . \quad (6.3.8)$$

Here $l_{\rm tot}$ is the total length of the particle's path in the interferometer and J and S are the eigenvalues of the total angular momentum (with respect to the center of mass) and the spin angular momentum, respectively. The other terms in the Hamiltonian do not contribute to the phase shift. The spin-rotation, the spin-orbit, and the spin-torsion phase shifts are only nontrivial, provided the particle's spin had been inverted shortly after splitting and, again, shortly before recombining the particle beam [Mashhoon (1988)]. The inertial effects have also been derived by Hehl & Ni (1990) by transforming the special-relativistic Dirac equation into an accelerated and rotating frame. For the spin-curvature phase shift we oriented the interferometer in such a way that the covector $\xi\rfloor R_{\alpha}{}^{\beta} S^{\alpha}{}_{\beta}$ has the same direction as the acceleration in the COW-case.

The phase shift $\delta\Phi_{\rm acc}$ is the leading contribution caused by the redshift of the energy. Note that the inertial spin-orbit coupling does not depend on the mass nor on the velocity of the particle, but only on its spin. The remaining terms, up to now,

have not yet been experimentally verified. If we employ in an atomic interferometer atoms of atomic weight 40, intrinsic spin 1, and velocity $0.1\,m/s$, then, for an effective interferometer area of about $10^{-4}\,m^2$, the following phase shifts are of the order $\delta\Phi_{s-o} \approx 10^{-20}$, $\delta\Phi_{s-c} \approx 10^{-25}$. Thus they are not measurable today.

7. Literature

Anandan, J. (1977): Gravitational and Rotational Effects in Quantum Interference, *Phys. Rev.* **D15**, 1448.

Atwood, D.K., Horne, M.A., Shull, C.G., Arthur, J. (1984): Neutron Phase Shift in a Rotating Two-Crystal Interferometer, *Phys. Rev. Lett.* **52**, 1673.

Audretsch, J. (1983): The Riemannian Structure of Space-Time as a Consequence of Quantum Mechanics, *Phys. Rev.* **D27**, 2872.

Audretsch, J., Gähler, F., Straumann, N. (1984): Wave Fields in Weyl-Space and Conditions for the Existence of a Preferred Pseudo-Riemannian Structure, *Comm. Math. Phys.* **95**, 41.

Audretsch, J., Lämmerzahl, C. (1983): Neutron interference: general theory of the influence of gravity, inertia and space-time torsion, *J. Phys. A: Math. Gen.* **16**, 2457.

Audretsch, J., Lämmerzahl, C. (1987): Neutron Interference: Influence of the Mirror on the Spin, *Ann. Phys.* (Leipzig) **44**, 145.

Audretsch, J., Lämmerzahl, C. (1988): Constructive Axiomatic Approach to Space-Time Torsion, *Class. Quantum Grav.* **5**, 1285.

Audretsch J., Lämmerzahl, C. (1991a): Reasons for a physical field to obey partial differential equations, *J. Math. Phys.* **32**, 1354.

Audretsch, J., Lämmerzahl, C. (1991b): Establishing the Riemannian structure of space-time by means of light rays and free matter waves, *J. Math. Phys.* **32**, 2099.

Audretsch, J., Lämmerzahl, C. (1992): New Inertial and Gravitational Effects made measurable by Atom Beam Interferometry, *Appl. Phys.* **B**, in press.

Bonse, U., Hart, M. (1965): An X-Ray Interferometer, *Appl. Phys. Lett.* **6**, 155.

Bonse, U., Wroblewski, T. (1983): Measurement of Neutron Quantum Interference in Noninertial Frames, *Phys. Rev. Lett.* **51**, 1401.

Bordé, Ch.J. (1989): Atomic interferometry with internal state labeling, *Phys. Lett.* **A140**, 10.

Carnal, O. (1992): *Beam Splitter, Lens and Interferometer for Metastable Helium Atoms by Diffraction from Microfabricated Transmission Structures*, Ph.D. Thesis, ETH Zürich.

Carnal, O., Mlynek, J. (1991): Young's Double Slit Experiment with Atoms: A Simple Atom Interferometer, *Phys. Rev. Lett.* **66**, 2689.

Cartan, É. (1986): *On Manifolds with an Affine Connection and the Theory of General Relativity*, English translation of the French original (Bibliopolis, Napoli).

Clauser, J.F. (1988): Ultra-heigh sensitivity accelerometers and gyroscopes using neutral atom matter-wave interferometers, *Physica* **B151**, 262.

Colella, R., Overhauser, A.W., Werner, S.A. (1975): Observation of Gravitationally Induced Quantum Interference, *Phys. Rev. Lett.* **34**, 1472.

Coleman, R.A., Korte, H. (1987): Jet bundles and path structures, *J. Math. Phys.* **21** 1340.

Collier, R. (1977): Pauli equation for a particle in metric and electromagnetic fields, *Czech. J. Phys.* **B27**, 991.

Ehlers, J., Köhler, E. (1977): Path Structure on Manifolds, *J. Math. Phys.* **18**, 2015.

Ehlers, J., Pirani, F.A.E., Schild, A. (1972): The Geometry of Free Fall and Light Propagation, **in**: L. O'Raifeartaigh (ed.): *General Relativity, Papers in Honour of J.L. Synge* (Clarendon Press, Oxford).

Einstein, A. (1955): *The Meaning of Relativity*, 5th ed. (Princeton University Press, Princeton).

Ertmer, W. (1991): Seminar given at the University of Konstanz.

Glasgow, S.; Meystre, P.; Wilkens, M.; Wright, E.M. (1991): Theory of an atomic beam splitter based on velocity-tuned resonances, *Phys. Rev.* **A43** 2455.

Goenner, H. (1987): Report on Symposium "Alternative Theories of Gravitation", **in**: *General Relativity and Gravitation, Proc. 11th Int. Conf. General Relativity and Gravitation* (GR11), Stockholm 1986, M.A.H. MacCallum (ed.), (Cambridge University Press, Cambridge).

Good, M.L. (1961): K_2^0 and the Equivalence Principle, *Phys. Rev.* **121**, 311.

Hasselbach, F., Nicklaus, M. (1988): An Electron Optical Sagnac Effect, *Physica* **B 151**, 230.

Hasselbach, F., Nicklaus, M. (1989): Observation of the rotational phaseshift of electron waves (Sagnac effect), presented at the Conference on Foundations of Quantum Mechanics to celebrate 30 Years of the Aharonov-Bohm effect, Columbia, South Carolina.

Hasselbach, F., Nicklaus, M. (1992): The Sagnac Effect with Electron Waves, *in preparation.*

Hecht, R. (1986): *private communication.*

Heer, C.V. (1961): Interference of Electromagnetic and Matter Waves in a Nonpermanent Gravitational Field, *Bull. Am. Phys. Soc.* **6**, 58.

Hehl, F. W. (1976): On the Energy-Momentum-Tensor of Spinning Massive Matter in Classical Field Theory and General Relativity, *Rep. Math. Phys.* **9**, 55.

Hehl, F. W. (1980): Four lectures on Poincaré gauge theory, **in**: Proceedings of the 6th Course of the School of Cosmology and Gravitation on *Spin, Torsion, Rotation, and Supergravity*, held at Erice, Italy, May 1979, P. G. Bergmann, V. de Sabbata, eds. (Plenum, New York), p.5.

Hehl, F.W. (1985): On the Kinematics of the Torsion of Space-Time, *Found. Phys.* **15**, 451.

Hehl, F. W., von der Heyde, P., Kerlick, G.D., Nester, J.M. (1976): General relativity with spin and torsion: Foundations and prospects, *Rev. Mod. Phys.* **48**, 393.

Hehl, F. W., Lemke, J., Mielke, E.W. (1991): Two lectures on fermions and gravity, in: *Geometry and Theoretical Physics*, Proc. of the Bad Honnef School, J. Debrus and A.C. Hirshfeld, eds. (Springer, Heidelberg).

Hehl, F.W., McCrea, J.D., Mielke, E.W., Ne'eman, Y. (1989): Progress in metric-affine gauge theories of gravity with local scale invariance, Found. Phys. **19**, 1075; see also by the same authors: Metric Affine Gauge Theory of Gravity, *Phys. Rep.*, to be published.

Hehl, F. W., Ni, W.-T. (1990): Inertial effects of a Dirac particle, *Phys. Rev.* **D42**, 2045.

Horne, M.A. (1986): Neutron interferometry in a gravity field, *Physica* **B137**, 260.

Huang, J. (1992): Dirac Particle, Gravity, and Inertial Effects, *Preprint*, Institute of Astronomy, Cambridge University.

Ivanenko, D, Sardanashvily, G. (1983): The gauge treatment of gravity, *Phys. Rep.* **94**, 1.

Kapitza, P.L., Dirac, P.A.M. (1933): The reflection of electrons from standing light rays, *Proc. Cambr. Phil. Soc.* **29**, 297.

Kasevich, M., Chu, S. (1991): Atomic Interferometry using stimulated Raman Transitions, *Phys. Rev. Lett.* **67**, 181.

Keith, D.W., Ekstrom, C.R., Turchette, Q.A., Pritchard, D.E. (1991): An Interferometer for Atoms, *Phys. Rev. Lett.* **66**, 2693.

Kibble, T.W.B. (1961): Lorentz invariance and the gravitational field, *J. Math. Phys.* **2**, 212.

Kibble, T.W.B., Stelle, K.S. (1986): Gauge Theories of Gravity and Supergravity, **in**: *Progress in Quantum Field Theory*. H. Ezawa and S. Kamefuchi, eds. (Elsevier, Amsterdam), p.57.

Klein, A.G., Opat, G.I., Cimmino, A., Zeilinger, A., Treimer, W., Gähler, R. (1981): Neutron propagation in moving matter: The Fizeau Experiment with Massive Particles, *Phys. Rev. Lett.* **24**, 1551.

Kopczyński, W. (1990): Variational Principles for Gravity and Fields, *Ann. Phys.* (N.Y.) **203**, 308.

Lämmerzahl, C. (1990): The Geometry of Matter Fields, **in**: Audretsch, J., deSabbata, V. (eds.): *Quantum Mechanics in Curved Space-Time* (Plenum Press, New York).

Lämmerzahl, C. (1991): A Hamilton Operator of the Dirac Field for Rotating and Accelerating Observers, *Preprint*, University of Konstanz.

Lämmerzahl, C. (1992): A Hamilton Operator for Matter Fields and Matter Wave Interferometry. *Preprint*, University of Konstanz.

Lord, E.A., Goswami, P. (1986): Gauge theory of a group of diffeomorphisms. I. General principles, *J. Math. Phys.* **27**, 2415.

Martin, P.J., Oldaker, B.G., Miklich, A.H., Prichard, D.E. (1988): Bragg Scatttering of Atoms from a Standing Light Wave, *Phys. Rev. Lett.* **60**, 515.

Marton, L., Simpson, J.A., Suddeth, J.A. (1953): Electron Beam Interferometer, *Phys. Rev.* **90**. 490.

Mashhoon, B. (1988): Neutron Interferometry in a Rotating Frame of Reference, *Phys. Rev. Lett.* **61**. 2639.

McCrea, J.D. (1987): Poincaré gauge theory of gravitation: foundations, exact solutions and computer algebra, **in**: *Proceedings of the 14th Int. Conference on Differential Geometric Methods in Mathematical Physics, Salamanca 1985*, P.L. García and A. Pérez-Rendón, eds., Lecture Notes in Mathmatics, Vol. **1251** (Springer, Berlin), p.222.

McCrea, J.D. (1989): Beyond General Relativity: Theories of Gravitation with Dynamic Torsion. Unpublished manuscript, University College Dublin.

Meyer, H. (1982), Møller's Tetrad theory of gravitation as a special case of Poincaré gauge theory – a coincidence? *Gen. Rel. Grav.* **14**, 531.

Michelson, A.A., Gale, H.G. (1925): The Effect of the Earth's Rotation on the Velocity of Light, *Nature* **115**, 566.

Mielke, E.W. (1987): *Geometrodynamics of Gauge Fields* – On the geometry of Yang-Mills and gravitational gauge theories (Akademie-Verlag, Berlin).

Miniatura, Ch., Perales, F., Vassilev, G., Reinhardt, J., Robert, J., Baudon, J. (1991): A longitudinal Stern-Gerlach interferometer: the "beaded" atom, *J. Phys. II* (France) **1** 425.

Möllenstedt, G., Bayh, W. (1961): Elektronen-Biprisma-Interferenzen mit weit getrennten kohärenten Teilbündeln, *Naturwissenschaften* **48**, 400.

Möllenstedt, G., Düker, H. (1954): Fresnelscher Interferenzversuch mit einem Biprisma für Elektronenwellen, *Naturwissenschaften* **42**, 41.

Möllenstedt, G., Jönsson, C. (1959): Elektronen-Mehrfachinterferenzen an regelmäßig hergestellten Feinspalten, *Z. Phys.* **155**, 472.

Ne'eman, Y. (1980): Gravity, groups and gauges, **in**: *General Relativity and Gravitation. One Hundred Years after the Birth of Albert Einstein*, A. Held ed.(Plenum Press, New York), Vol. 1, Chap. 10.

Ne'eman, Y., Regge, T. (1978): Gravity and supergravity as gauge theories on a group manifold, *Phys. Lett.* **B74**, 54; Gauge theory of gravity and supergravity on a group manifold *Riv. Nuovo Cimento* **1** N5, Ser. 3.

Nester, J. (1984): Gravity, Torsion and Gauge Theory, **in**: *Introduction to Kaluza-Klein theories*, H.C. Lee, ed. (World Scientific, Singapore), p. 83.

O'Raifeartaigh, L. (1979): Hidden Gauge Symmetry, *Rep. Prog. Phys.* **42**, 159.

Overhauser, A.W., Colella, R. (1974): Experimental Test of Gravitational Induced Quantum Interference, *Phys. Rev. Lett.* **33**, 1237.

Page, L.A. (1975): Effect of Earth's Rotation in Neutron Interferometry, *Phys. Rev. Lett.* **35**, 543.

Rauch, H., Treimer, W., Bonse, U. (1974): Test of single crystal neutron interferometer, *Phys. Lett.* **A47**, 369.

Physics Today (1989): October, p.20.

Riehle, F., Kisters, Th., Witte, A., Helmcke, J., Bordé, Ch.J. (1991): Optical Ramsey Spectroscopy in a Rotating Frame: Sagnac Effect in a Matter Wave Interferometer, *Phys. Rev. Lett.* **67**, 177.

Schmutzer, E. (1973): Maxwell-Theorie (in Medien) und Quantentheorie in einem rotierenden Bezugssystem, *Ann. Phys.* (Leipzig) **29**, 75.

Sciama, D.W. (1962): On the analogy between charge and spin in general relativity, **in**: *Recent developments in general relativity* (Pergamon, Oxford), p.415.

Sexl, R.U., Urbantke, H.K. (1982): *Relativität, Gruppen, Teilchen*, 2nd ed. (Springer, Wien).

Sexl, R.U., Urbantke, H.K. (1983): *Gravitation und Kosmologie*, 2nd ed. (Bibliographisches Institut, Mannheim 1983).

Shimizu, F., Shimizu, K., Takuma, H. (1992): Double slit interference with ultracold metastable neon atoms, *preprint*, University of Tokyo.

Shimony, A. (1979): Proposed neutron interferometer test of some nonlinear variants of wave mechanics, *Phys. Rev.* **A20**, 394.

Shull, C.G., Atwood, D.K., Arthur, J., Horne, M.A. (1980): Search for a non-linear variant of the Schrödinger equation, *Phys. Rev. Lett.* **44**, 765.

Thirring, W. (1986): *Classical Field Theory*, A Course in Mathematical Physics 2, 2nd ed.(Springer, New York).

Trautman, A. (1973): On the structure of the Einstein-Cartan equations, **in**: *Differential Geometry*, Symposia Matematica Vol.12 (Academic Press, London) p.139.

Trautman, A. (1980): Fiber bundles, gauge fields and gravitation, **in**: *General Relativity and Gravitation. One Hundred Years after the Birth of Albert Einstein*, A. Held ed.(Plenum, New York), Vol.1, Chap.9, pp.287-308.

Utiyama, R. (1956): Invariant theoretical interpretation of interaction, *Phys. Rev.* **101**, 1597.

Utiyama, R. (1980): Introduction to the theory of general gauge fields, *Prog. Theor. Phys.* **64**, 2207.

Weinberg, S. (1989): Testing Quantum Mechanics, *Ann. Phys.*(N.Y.) **194**, 336.

Werner, S.A., Kaiser, H. (1990): Neutron Interferometry – Macroscopic Manifestations of Quantum Mechanics, in: Audretsch, J., deSabbata, V.: *Quantum Mechanics in Curved Space-Time* (Plenum Press, New York), pp 1-21.

Werner, S.A., Staudenmann, J.-L., Colella, R. (1979): Effect of Earth's Rotation on Quantum Mechanical Phase of the Neutron, *Phys. Rev. Lett.* **42**, 1103.

Yang, C.N. and Mills, R.L. (1954), Conservation of isotopic spin and isotopic gauge invariance, *Phys. Rev.* **96**, 191.

Yasskin P.B., Stoeger, W.R. (1980): Propagation equations for test bodies with spin and rotation in theories of gravity with torsion, *Phys. Rev.* **D21**, 2081.

Author Index