Prologue

The story of the black hole begins with Schwarzschild's discovery [Sc] of the Schwarzschild solution in 1916, soon after Einstein's foundation of the general theory of relativity [Ei1] and his final formulation of the field equations of gravitation [Ei2], the *Einstein equations*, in 1915. The Schwarzschild solution is a solution of the vacuum Einstein equations which is spherically symmetric and depends on a positive parameter M, the *mass*. With r such that the area of the spheres, which are the orbits of the rotation group, is $4\pi r^2$, the solution in the coordinate system in which it was originally discovered had a singularity at $r = 2M$. For this reason only the part which corresponds to $r > 2M$ was originally thought to make sense. This part is static and represents the gravitational field outside a static, spherically symmetric body with surface area corresponding to some $r_0 > 2M$.

However, the understanding of Schwarzschild's solution gradually changed. First, in 1923 Birkoff [Bir] proved a theorem which shows that the Schwarzschild solution is the only spherically symmetric solution of the vacuum Einstein equations. One does not therefore need to assume that the solution is static. Thus, Schwarzschild's solution represents the gravitational field outside any spherically symmetric body, evolving in any manner whatever, for example undergoing gravitational collapse.

Eddington [Ed], in 1924, made a coordinate change which transformed the Schwarzschild metric into a form which is not singular at $r = 2M$, however he failed to take proper notice of this. Only in 1933, with Lemaître's work [L], was it realized that the singularity at $r = 2M$ is not a true singularity but rather a failure of the original coordinate system. Eddington's transformation was rediscovered by Finkelstein [Fi] in 1958, who realized that the hypersurface $r = 2M$ is an *event horizon*, the boundary of the region of spacetime which is causally connected to infinity, and recognized the dynamic nature of the region $r < 2M$. Now, Schwarzschild's solution is symmetric under time reversal, and one part of it, the one containing the *future event horizon*, the boundary of the region of spacetime which can send signals to infinity, is covered by one type of Eddington-Finkelstein coordinates, while the other part, the one containing the *past event horizon*, the boundary of the region of spacetime which can receive signals from infinity, is covered by the other type of Eddington-Finkelstein coordinates. Actually, only the first part is physically relevant, because only future event horizons can form dynamically, in gravitational collapse. Systems of coordinates that cover the complete analytic extension of the Schwarzschild solution had been provided earlier (in 1950) by Synge [Sy], and a single most convenient system that covers the complete analytic extension was discovered independently by Kruskal [Kr] and Szekeres [Sz] in 1960.

Meanwhile in 1939, Oppenheimer and Snyder had studied the gravitational collapse of a pressure-free fluid ball of uniform density, a uniform density "ball of dust". Even though this is a highly idealized model problem, their work was very significant, being the first work on relativistic gravitational collapse. As mentioned above, the spacetime geometry in the vacuum region outside the ball is given by the Schwarzschild metric. Oppenheimer and Snyder analyzed the causal structure of the solution. They considered in particular an observer on the surface of the dust ball sending signals to a faraway stationary observer at regularly spaced intervals as judged by his own clock. They discovered that the spacing between the arrival times of these signals to the faraway observer becomes progressively longer, tending to infinity as the radius r_0 corresponding to the surface of the ball approaches $2M$. This effect has since been called the *infinite redshift* effect. The observer on the surface of the dust ball may keep sending signals after r_0 has become less than $2M$, but these signals proceed to ever smaller values of r until, within a finite affine parameter interval, they reach a true singularity at $r = 0$. The observer on the surface of the ball reaches this singular state himself within a finite time interval as judged by his own clock. The concept of a future event horizon, and hence of a region of spacetime bounded by this horizon from which no signals can be sent which reach arbitrarily large distances, was thus already implicit in the Oppenheimer-Snyder work.

The 1964 work of Penrose [P1] introduced the concept of *null infinity*, which made possible the precise general definition of a *future event horizon* as *the boundary of the causal past of future null infinity*. A turning point was reached in 1965 with the introduction by Penrose of the concept of a closed *trapped surface* and his proof of the first *singularity theorem*, or, more precisely, *incompleteness theorem* [P2]. Penrose defined a trapped surface as being a spacelike surface in spacetime, such that an infinitesimal virtual displacement of the surface along either family of future-directed null geodesic normals to the surface leads to a pointwise decrease of the area element. On the basis of this concept, Penrose proved the following theorem:

A spacetime (M, g) *cannot be future null geodesically complete if*:

1. $\mathrm{Ric}(N, N) \geq 0$ *for all null vectors* N.
2. *There is a non-compact Cauchy hypersurface* H *in* M.

   and:
3. *There is a closed trapped surface* S *in* M.

Here Ric is the Ricci curvature of g and condition 1 is always satisfied by virtue of the Einstein equations and the physical positivity condition on the energy-momentum-stress tensor of matter.

Once the notions of null infinity and of a closed trapped surface were introduced, it did not take long to show that a spacetime with a complete future null infinity which contains a closed trapped surface must contain a future event horizon, the interior of which contains the trapped surface (see [H-E], Proposition 9.2.1). For the ideas and methods which go into Penrose's theorem the reader may consult, besides the monograph by Hawking and Ellis just mentioned, the article by Penrose in [P3] as well as his monograph [P4]. Further singularity theorems, which also cover cosmological situations, were

subsequently established by Hawking and Penrose (see [H-E]), but it is the original singularity theorem quoted above which is of interest in the present context, as it concerns gravitational collapse. We should also mention that the term *black hole* for the interior of the future event horizon was introduced by Wheeler in 1967 (see [Wh]).

Now, the 1952 work of Choquet-Bruhat [Cho1] (see also [Cho2] and [Cho3]) had shown that any initial data set $(H, \overline{g}, k)$, where H is a 3-dimensional manifold, $\overline{g}$ is a Riemannian metric on H and k is a symmetric 2-covariant tensorfield on H, such that the pair $(\overline{g}, k)$ satisfies the so-called "constraint equations", has a *future development* (M, g), namely a 4-dimensional manifold M endowed with a Lorentzian metric g satisfying the vacuum Einstein equations, such that H is the past boundary of M, $\overline{g}$ and k are the first and second fundamental forms of H relative to (M, g), and for each $p \in M$ each past-directed causal curve initiating at p terminates at a point of H. The constraint equations are the contracted Codazzi and twice contracted Gauss equations of the embedding of H in M. The subsequent 1969 work of Choquet-Bruhat and Geroch [C-G] then showed that each such an initial data set has a unique *maximal future development* M^*, namely a future development, in the above sense, which extends every other future development of the same initial data set. Geroch [Ge] subsequently showed that for any future development (M, g), M is diffeomorphic to $[0, \infty) \times H$. Moreover, the above theorems extend to the case where instead of vacuum we have suitable matter, such as a perfect fluid, or an electromagnetic field. In the light of the theorem of Choquet-Bruhat and Geroch, condition 2 in Penrose's theorem may be replaced by the statement that (M, g) is the maximal future development of initial data on a complete non-compact spacelike hypersurface.

In 1990 Rendall [R] solved in a very satisfactory manner the local *characteristic initial value problem* for the vacuum Einstein equations (earlier work had been done by Choquet-Bruhat [Cho4] and by Müller with Hagen and Seifert [M-S]). In this case we have, in the role of H, either two null hypersurfaces C and $\underline{C}$ intersecting in a spacelike surface S, S being the past boundary of both C and $\underline{C}$, or a future null geodesic cone C_o of a point o. The initial data on C and $\underline{C}$ are the conformal intrinsic geometry of these null hypersurfaces, together with the full intrinsic geometry of S, the initial rate of change of the area element of S under displacement along C and $\underline{C}$, and a certain 1-form on S (the torsion). The initial data on C_o are the conformal intrinsic geometry of C_o and certain regularity conditions at o. In contrast to the case where the initial data are given on a spacelike hypersurface, there are no constraints, and the initial data can be freely specified. The theorem of Rendall then shows that any such characteristic initial data has a future development (M, g), bounded in the past by a neighborhood of S in $C \bigcup \underline{C}$ and of o in C_o respectively. The theorem of Choquet-Bruhat and Geroch, which applies to future developments, then shows that there is a unique maximal future development (M^*, g) corresponding to the given characteristic initial data.

Now the proof of the theorem of Penrose is by showing that if M were complete, the boundary $\partial J^+(S)$ of the causal future $J^+(S)$ of the closed trapped surface S would be compact. The integral curves of any timelike vectorfield on M would define a continuous mapping of $\partial J^+(S)$ into H, M being a development of H, and this mapping would have to be a homeomorphism onto its image, $\partial J^+(S)$ being compact. This leads to a contradiction with the assumption that H is non-compact. We see that the proof makes no use

of the strictly spacelike nature of H other than through the assumption that M is a future development of H. We may therefore replace H by a complete future null geodesic cone and restate the theorem as follows. Here vacuum or suitable matter is assumed. We do not state the first condition of Penrose's because as we already mentioned, it is automatically satisfied by virtue of the physical positivity condition that the energy-momentum-stress tensor of matter satisfies.

Let us be given regular characteristic initial data on a complete null geodesic cone C_o of a point o. Let (M^*, g) be the maximal future development of the data on C_o. Suppose that M^* contains a closed trapped surface. Then (M^*, g) is future null geodesically incomplete.

An important remark at this point is that it is not *a priori* obvious that closed trapped surfaces are *evolutionary*. That is, it is not obvious whether closed trapped surfaces can form in evolution starting from initial conditions in which no such surfaces are present. What is more important, the physically interesting problem is the problem where the initial conditions are of arbitrarily *low compactness*, that is, arbitrarily far from already containing closed trapped surfaces, and we are asked to follow the long-time evolution and show that, under suitable circumstances, closed trapped surfaces eventually form. Only an analysis of the dynamics of gravitational collapse can achieve this aim.

Returning to our review of the historical development of the black hole concept, a very significant development took place in 1963, shortly before the work of Penrose. This was the discovery by Kerr [Ke] of a two-parameter family of axially symmetric solutions of the vacuum Einstein equations, with an event horizon, the exterior of which is a regular asymptotically flat region. The two parameters are the *mass* M, which is positive, and the *angular momentum* L about the axis of symmetry, which is subject to the restriction $|L| \leq M^2$. Kerr's solution reduces in the special case of vanishing angular momentum to Schwarzschild's solution. The Kerr solution possesses an additional Killing field, besides the generator of rotations about the axis, however this additional Killing field, in contrast to the case of the Schwarzschild solution, is timelike not on the entire exterior of the horizon, but only in the exterior of a non-spacelike hypersurface containing the horizon. So only in this exterior region is the solution stationary in a strict sense. At every point of the region between the two hypersurfaces, called the *ergosphere*, the additional Killing field is spacelike, but the plane which is the linear span of the two vectors at this point is timelike. On the horizon itself the plane becomes null and tangent to the horizon, and the null line generating this null plane defines the *angular velocity* of the horizon, a constant associated to the horizon. Kerr's solution is symmetric under time reversal if also the sign of the angular momentum is reversed, hence it possesses, besides the future event horizon, also an unphysical past event horizon, just like the Schwarzschild solution.

The fascinating properties of the Kerr solution were revealed in the decade following its discovery. In particular, Boyer and Lindquist [B-L] introduced a more convenient coordinate system and obtained the maximal analytic extension. One of these fascinating properties which concerns us here is that the hypersurfaces of constant Boyer-Lindquist coordinate t are complete asymptotically flat maximal spacelike hypersurfaces, their maximal future development contains closed trapped surfaces and, in accordance

with Penrose's theorem, is incomplete. Nevertheless the future boundary of the maximal development is *nowhere singular*, the solution extending as an analytic solution across this boundary. This future boundary is a regular null hypersurface, a *Cauchy horizon*, illustrating the fact that incompleteness of the maximal future development does not imply a singular future boundary.

Returning again to the question of whether closed trapped surfaces are evolutionary, one may at first hand say that the question was already settled in the affirmative by the Oppenheimer-Snyder analysis. This is because hyperbolic systems of partial differential equations, such as the Einstein-Euler equations describing a perfect fluid in general relativity, possess the property of continuous dependence of the solution on the initial conditions. This holds at a given non-singular solution, for a given finite time interval. Thus, since the initial condition of a homogeneous dust ball leads to a trapped sphere within a finite time interval, initial conditions which are sufficiently close to this special initial condition will also lead to the formation of closed trapped surfaces within the same time interval, the condition for a closed spacelike surface to be trapped being an open condition. However, as we remarked above, the case that one is really interested in is that for which the initial homogeneous dust ball is of low compactness, far from already containing trapped spheres, and it is only by contracting for a sufficiently long time that a trapped sphere eventually forms. In this case the closeness condition of the continuous dependence theorem may require the initial conditions to be so unreasonably close to those of a homogeneous dust ball that the result is devoid of physical significance.

With the above remarks in mind the author turned to the study of the gravitational collapse of an inhomogeneous dust ball [Chr1]. In this case, the initial state is still spherically symmetric, but the density is a function of the distance from the center of the ball. The corresponding spherically symmetric solution had already been obtained in closed form by Tolman in 1934 [T], in comoving coordinates, but its causal structure had not been investigated. This required integrating the equations for the radial null geodesics. A very different picture from the one found by Oppenheimer and Snyder emerged from this study. The initial density being assumed a decreasing function of the distance from the center, so that the central density is higher than the mean density, it was found that as long as the collapse proceeds from an initial state of low compactness, the central density becomes infinite before a black hole has a chance to form, thus invalidating the neglect of pressure and casting doubt on the predictions of the model from this point on, in particular on the prediction that a black hole eventually forms.

At this point the author turned to the spherically symmetric scalar field model [Chr2]. This is the next simplest material model after the dust model. The energy-momentum-stress tensor of matter is in this case that corresponding to a scalar field ϕ:

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi + \frac{1}{2}\sigma g_{\mu\nu}, \quad \sigma = -(g^{-1})^{\mu\nu}\partial_\mu\phi\partial_\nu\phi. \tag{1}$$

The integrability condition for Einstein's equations, namely that $T_{\mu\nu}$ is divergence-free, is then equivalent to the wave equation for ϕ relative to the metric g. The problem had been given to the author by his teacher, John Archibald Wheeler, in 1968 (see [Chr3]), as a model problem through which insight into the dynamics of gravitational collapse would

be gained. In the case of the dust model, there is no force opposing the gravitational attraction , so there is no alternative to collapse. This is not the case for the scalar field model and indeed in [Chr2] it was shown that, if the initial data are suitably small, we obtain a complete regular solution dispersing to infinity in the infinite future. So, for the scalar field model there is a *threshold* for gravitational collapse. In this paper and the papers on the scalar field that followed, the initial data were given on a complete future null geodesic cone C_o extending to infinity. The initial data on C_o consist of the function $\alpha_0 = \partial(r\phi)/\partial s|_{C_o}$, s being the affine parameter along the generators of C_o.

The next paper [Chr4] on the scalar field problem addressed the general case, when the initial data were no longer restricted by a smallness condition. The aim of this work was to prove the existence of a solution with a complete *domain of outer communications*, that is, a development possessing a complete future null infinity, the domain of outer communications being defined as the causal past of future null infinity. This was tantamount to proving the *weak cosmic censorship conjecture* of Penrose [P5] (called "asymptotic future predictability" in [H-E]) in the context of the spherically symmetric scalar field model. The aim was not reached in this paper. What was established instead was the existence, for all regular asymptotically flat initial data, of a *generalized solution* corresponding to a complete domain of outer communications. A generalized solution had enough regularity to permit the study of the asymptotic behavior in the domain of outer communications in the next paper, however no uniqueness could be claimed for these generalized solutions, so the conjecture of Penrose was left open.

In [Chr5] it was shown that when the final Bondi mass, that is, the infimum of the Bondi mass at future null infinity, is different from zero, a black hole forms of mass equal to the final Bondi mass, surrounded by vacuum. The rate of growth of the redshift of light seen by faraway observers was determined and the asymptotic wave behavior at future null infinity and along the event horizon was analyzed. However, the question of whether there exist initial conditions which lead to a non-zero final Bondi mass was not addressed in this paper.

The next paper [Chr6] was a turning point in the study of the spherically symmetric scalar field problem. Because of the fact that it has provided a stepping stone for the present monograph, I quote its main theorem. Here C_o denotes the initial future null geodesic cone.

Consider on C_o an annular region bounded by two spheres $S_{1,0}$ and $S_{2,0}$ with $S_{2,0}$ in the exterior of $S_{1,0}$. Let δ_0 and η_0 be the dimensionless size and the dimensionless mass content of the region, defined by

$$\delta_0 = \frac{r_{2,0}}{r_{1,0}} - 1, \qquad \eta_0 = \frac{2(m_{2,0} - m_{1,0})}{r_{2,0}},$$

$r_{1,0}, r_{2,0}$ and $m_{1,0}, m_{2,0}$ being the area radii and mass contents of $S_{1,0}$, $S_{2,0}$ respectively. Let $\underline{C}_1$ and $\underline{C}_2$ be incoming null hypersurfaces through $S_{1,0}$ and $S_{2,0}$ and consider the spheres S_1 and S_2 at which $\underline{C}_1$ and $\underline{C}_2$ intersect future null geodesic cones C with vertices on the central timelike geodesic Γ_0. There are positive constants c_0 and c_1 such that, if

$\delta_0 \leq c_0$ and

$$\eta_0 > c_1 \delta_0 \log\left(\frac{1}{\delta_0}\right),$$

then S_2 becomes trapped before S_1 reduces to a point on Γ_0. There is a future null geodesic cone C^* with vertex on Γ_0 such that S_2^* is a maximal sphere in C^* while $r_1^* > 0$.

It was further shown that the region of trapped spheres, the *trapped region*, terminates at a strictly spacelike singular boundary, and contains spheres whose mass content is bounded from below by a positive constant depending only on $r_{1,0}, r_{2,0}$, a fact which implies that the final Bondi mass is positive, thus connecting with the previous work.

An important remark concerning the proof of the above theorem needs to be made here. The proof does not consider at all the region interior to the incoming null hypersurface $\underline{C}_1$. However, the implicit assumption is made that no singularities form on Γ_0 up to the vertex of C^*. If a smallness condition is imposed on the restriction of the initial data to the interior of $S_{1,0}$, the argument of [Chr2] shows that this assumption indeed holds. Also, by virtue of the way in which the theorem was later applied in [Chr9], the assumption in question was a priori known to hold.

Since the spherical dust model had been disqualified by [Chr1] as establishing the dynamical formation of trapped spheres, the work [Chr6] was the first to establish the dynamical formation of closed trapped spheres in gravitational collapse, although, of course, severely limited by the restriction to spherical symmetry and by the fact that it concerned an idealized matter model.

Solutions with initial data of bounded variation were considered in [Chr7] and a sharp sufficient condition on the initial data was found for the avoidance of singularities, namely that the total variation be sufficiently small, greatly improving the result of [Chr2]. Moreover, a sharp extension criterion for solutions was established, namely that if the ratio of the mass content to the radius of spheres tends to zero as we approach a point on Γ_0 from its causal past, then the solution extends as a regular solution to include a full neighborhood of the point. The structure of solutions of bounded variation was studied and it was shown that at each point of Γ_0 the solutions are locally scale invariant. Finally, the behavior of the solutions at the singular boundary was analyzed.

In [Chr8] the author constructed examples of solutions corresponding to regular asymptotically flat initial data which develop singularities that are not preceded by a trapped region but have future null geodesic cones expanding to infinity. It was thus established for the first time that *naked singularities* do, in fact, occur in the gravitational collapse of a scalar field. Also, other examples were constructed which contain *singular future null geodesic cones* that have collapsed to lines and again are not preceded by trapped regions.

The work on the spherically symmetric scalar field model culminated in [Chr9]. Taking the space of initial data to be the space $\mathcal{A}$ of absolutely continuous functions on the non-negative real line, the theorem proved in [Chr9] was the following.

Let us denote by $\mathcal{R}$ the subset of $\mathcal{A}$ consisting of those initial data which lead to a complete maximal future development, and by $\mathcal{S}$ its complement in $\mathcal{A}$. Let also $\mathcal{G} \subset \mathcal{S}$ be the subset consisting of those initial data which lead to a maximal development possessing a

complete future null infinity and a strictly spacelike singular future boundary. Then $\mathcal{E} = \mathcal{S} \setminus \mathcal{G}$ *has the following property. For each initial data* $\alpha_0 \in \mathcal{E}$ *there is a function* $f \in \mathcal{A}$*, depending on* α_0*, such that the line* $\mathcal{L}_{\alpha_0} = \{\alpha_0 + cf \ : \ c \in \Re\}$ *in* $\mathcal{A}$ *is contained in* $\mathcal{G}$*, except for* α_0 *itself. Moreover, the lines* $\mathcal{L}_{\alpha_{0,1}}$, $\mathcal{L}_{\alpha_{0,2}}$ *corresponding to distinct* $\alpha_{0,1}, \alpha_{0,2} \in \mathcal{E}$ *do not intersect.*

The exceptional set $\mathcal{E}$ being, according to this theorem, of codimension at least 1, the theorem established, within the spherically symmetric scalar field model, the validity not only of the weak cosmic censorship conjecture of Penrose, but also of his *strong cosmic censorship conjecture*, formulated in [P6]. This states, roughly speaking, that generic asymptotically flat initial data have a maximal development which is either complete or terminates in a totally singular future boundary. The general notion of causal boundary of a spacetime manifold was defined in [G-K-P]. The relationship between the two cosmic censorship conjectures is discussed in [Chr10]. In the case of the spherically symmetric scalar field model, there is no system of local coordinates in which the metric extends as a Lorentzian metric through any point of the singular future boundary.

The proof of the above theorem is along the following lines. It is first shown that if Γ_0 is complete, the maximal future development is also complete. Thus one can assume that Γ_0 has a singular end point e. We then consider $\underline{C}_e$, the boundary of the causal past of e. This intersects the initial future null geodesic cone C_o in a sphere $S_{0,e}$. Given then any sphere $S_{0,1}$ exterior to $S_{0,e}$ on C_o, but as close as we wish to $S_{0,e}$, we consider the incoming null hypersurface $\underline{C}_1$ through $S_{0,1}$. Allowing a suitable modification of the initial data as in the statement of the theorem, with f a function vanishing in the interior of $S_{0,e}$ on C_o, it is then shown that there exists a point p_0 on Γ_0, earlier than e, such that the annular region on the future null geodesic cone C_{p_0}, with vertex at p_0, bounded by the intersections with $\underline{C}_e$ and $\underline{C}_1$ satisfies the hypotheses of the theorem of [Chr6]. It is in this part of the proof that the singular nature of the point e is used. Application of the theorem of [Chr6] then shows that if we consider future null geodesic cones C_p with vertices p on the segment of Γ_0 between p_0 and e, and the corresponding intersections with $\underline{C}_e$ and $\underline{C}_1$, then for some p^* in this segment earlier than e, $C_{p^*} \bigcap \underline{C}_1$ is a maximal sphere in C_{p^*}, and the part of $\underline{C}_1$ to the future of this sphere lies in a trapped region. We see therefore the essential role played by the formation of trapped spheres theorem of [Chr6] in the proof of the cosmic censorship conjectures in [Chr9] in the framework of the spherically symmetric scalar field model.

A model, closely related to the scalar field model but with surprising new features, was studied by Dafermos in [D1], [D2]. In this model we have in addition to the scalar field an electromagnetic field. The two fields are only indirectly coupled, through their interaction with the gravitational field, the energy-momentum-stress tensor of matter being the sum of (1) with the Maxwell energy-momentum-stress tensor for the electromagnetic field. By the imposition of spherical symmetry, the electromagnetic field is simply the Coulomb field corresponding to a constant charge Q. This is non-vanishing by virtue of the fact that the topology of the manifold is $\Re^2 \times S^2$, like the manifold of the Schwarzschild solution, so there are spheres which are not homologous to zero. Dafermos showed that in this case part of the boundary of the maximal development is a Cauchy horizon, through which the metric can be continued in a C^0 manner, but at which, generically,

the mass function blows up. As a consequence, generically, there is no local coordinate system in any neighborhood of any point on the Cauchy horizon in which the connection coefficients (Christoffel symbols) are square-integrable. This means that the solution ceases to make sense even as a weak solution of the Einstein-Maxwell-scalar field equations if we attempt to include the boundary. The work of Dafermos illustrates how much care is needed in formulating the strong cosmic censorship conjecture. In particular, the formulation given in [Chr10] according to which C^0 extensions through the boundary of the maximal development are generically excluded, turned out to be incorrect. Only if the condition is added that there be no extension *as a solution, even in a weak sense*, to include any part of the boundary, is the counterexample avoided.

Before the work on the scalar field model was completed, the author introduced and studied a model which was designed to capture some of the features of actual stellar gravitational collapse, while capitalizing to a maximum extent on the knowledge gained in the study of the dust and scalar field models. This was the two-phase model, introduced in [Chr11] and studied further in [Chr12] and [Chr13]. Let us recall here that a perfect fluid model is in general defined by specifying a function $e(n, s)$, the *energy per particle* as a function of n, the *number of particles per unit volume*, and s, the *entropy per particle*. This is called the *equation of state*. Then the *mass-energy density* ρ, the *pressure* p and the *temperature* θ are given by:

$$\rho = ne, \quad p = n^2 \frac{\partial e}{\partial n}, \quad \theta = \frac{\partial e}{\partial s}. \tag{2}$$

The mechanics of a perfect fluid are governed by the differential conservation laws

$$\nabla_\nu T^{\mu\nu} = 0, \qquad \nabla_\mu I^\mu = 0 \tag{3}$$

where $T^{\mu\nu}$ is the energy-momentum-stress tensor

$$T^{\mu\nu} = \rho u^\mu u^\nu + p((g^{-1})^{\mu\nu} + u^\mu u^\nu), \tag{4}$$

u^μ being the *fluid velocity*, and

$$I^\mu = nu^\mu \tag{5}$$

is the *particle current*. In the case of the two-phase model, p is a function of ρ alone. For such fluids, called *barotropic*, p and ρ are functions of the single variable

$$\mu = nm(s) \tag{6}$$

where $m(s)$ is a positive increasing function of s. In the two-phase model, if ρ is less than a critical value, which by proper choice of units we may set equal to 1, the matter is as soft as possible, the sound speed being equal to 0, while if ρ is greater than 1, the matter is as hard as possible, the sound speed being equal to 1, that is, to the speed of light in vacuum. Let us recall here that the *sound speed* η is in general given by

$$\eta^2 = \left(\frac{dp}{d\rho}\right)_s. \tag{7}$$

The pressure in the two-phase model is then given by

$$p = \begin{cases} 0 & : \text{if } \rho \leq 1 \\ \rho - 1 & : \text{if } \rho > 1 \end{cases}. \tag{8}$$

The condition of spherical symmetry being imposed, the flow is irrotational. The soft phase of the two-phase model coincides with the dust model while the hard phase coincides with the scalar field model with the restriction that $-(g^{-1})^{\mu\nu}\partial_\nu\phi$ be a future-directed timelike vectorfield. With

$$\sigma = -(g^{-1})^{\mu\nu}\partial_\mu\phi\partial_\nu\phi, \tag{9}$$

the density of mass-energy ρ and the fluid velocity u^μ are given by

$$\rho = \frac{1}{2}(\sigma + 1), \quad u^\mu = -\frac{(g^{-1})^{\mu\nu}\partial_\nu\phi}{\sqrt{\sigma}}. \tag{10}$$

The energy-momentum-stress tensor in the hard phase is

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi + \frac{1}{2}(\sigma - 1)g_{\mu\nu}, \tag{11}$$

so it differs from the standard one for a scalar field (1) by the term $-(1/2)g_{\mu\nu}$, the divergence-free condition on $T_{\mu\nu}$ being again equivalent to the wave equation for ϕ in the metric g.

Each of the two phases is by itself incomplete, the soft phase being limited by the condition $\rho \leq 1$ and the hard phase being limited by the condition $\sigma \geq 1$. The soft phase turns in contraction into the hard phase, while the hard phase turns upon expansion into the soft phase. Only the two phases taken together constitute a complete model. The hypersurface which forms the interface between the two phases has both spacelike and timelike components. Across a spacelike component, the thermodynamic variables n, s or ρ, p and the fluid velocity u^μ are continuous, the final values of one phase providing the initial values for the next phase. However, across a timelike component the thermodynamic variables and the fluid velocity suffer discontinuities, determined by the integral form of the conservation laws. These are of an irreversible character, each point of a timelike component which is crossed by a flow line being a point of increase of the entropy. A timelike component of the phase boundary is therefore a *shock*, and the development of these shocks is a free-boundary problem, which was studied in [Chr12] and [Chr13]. With initial condition an inhomogeneous dust ball at zero entropy, these papers showed that the core of the ball turns continuously into the hard phase, however at a certain sphere a shock forms which propagates outwards, absorbing the exterior part of the original dust ball. Behind this shock we have the hard phase at positive entropy, but the analysis was not carried further to investigate under what initial conditions a black hole will eventually form. We should also mention here that the two-phase model admits a one-parameter family of static solutions, balls of the hard phase, surrounded by vacuum.

The goal of the effort in the field of relativistic gravitational collapse is the study of the formation of black holes and singularities for general asymptotically flat initial conditions, that is, when no symmetry conditions are imposed.

In this connection, an interesting theorem was established by Schoen and Yau [S-Y1], as an outgrowth of their proof of the general case of the positive mass theorem [S-Y2] (their earlier work [S-Y1] covered the case of a maximal spacelike hypersurface of vanishing linear momentum; a different proof of the general theorem was subsequently given by Witten [Wi]). In [S-Y1] it is shown that if the energy density minus the magnitude of the momentum density of matter on a spacelike hypersurface is everywhere bounded from below by a positive constant b in a region which is large enough in a suitable sense that roughly corresponds to linear dimensions of at least $b^{-1/2}$, then the spacelike hypersurface must contain a closed trapped surface diffeomorphic to S^2. Although this work does not address the problem of evolution, the constraint equations alone entering the proof, it is nevertheless relevant for the problem of evolution, in so far as it reduces the problem of the dynamical formation of a closed trapped surface to the problem of showing that, under suitable circumstances, the required material energy concentration eventually occurs.

As far as the problem of evolution itself, let us first discuss the case where the material model is a perfect fluid. Then, as we have seen in the spherically symmetric case, before closed trapped surfaces form, shock waves already form. Now, the general problem of shock formation in a relativistic fluid, in the physical case of three spatial dimensions, has recently been studied by the author in the monograph [Chr14]. This work is in the framework of special relativity. We should remark here that the only previous result in relation to shock formation in fluids in three spatial dimensions was the result of Sideris [Si] which considers the non-relativistic problem of a classical ideal gas with adiabatic index $\gamma > 1$. Moreover, in that work it is only shown that the solutions cannot remain C^1 for all time, no information being given as to the nature of the breakdown.

However, more detailed results on breakdown for certain quasilinear wave equations in more than one spatial dimension had in the meantime been obtained by Alinhac [A1, A2]. The theorems proved in the monograph [Chr14] give a detailed picture of shock formation. In particular a detailed description is given of the geometry of the boundary of the maximal development of the initial data and of the behavior of the solution at this boundary. The notion of maximal development in this context is not that relative to the background Minkowski metric $g_{\mu\nu}$, but rather the one relative to the *acoustical metric*

$$h_{\mu\nu} = g_{\mu\nu} + (1 - \eta^2)u_\mu u_\nu, \quad u_\mu = g_{\mu\nu}u^\nu, \tag{12}$$

a Lorentzian metric, the null cones of which are the sound cones. It is not appropriate to give here a complete summary of the results of [Chr14]. Instead, the following short discussion should suffice to give the reader a feeling for the present status of shock wave theory in the physical case of three spatial dimensions. In [Chr14] it is shown that the boundary of the maximal development in the above "acoustical" sense consists of a regular part and a singular part. Each component of the regular part $\underline{C}$ is an incoming characteristic (relative to h) hypersurface which has a singular past boundary. The singular part of the boundary is the locus of points where the density of foliations by outgoing characteristic (relative to h) hypersurfaces blows up. It is the union $\partial_- B \bigcup B$, where each component of $\partial_- B$ is a smooth embedded surface in Minkowski spacetime, the tangent plane to which at each point is contained in the exterior of the sound cone at that point. On

the other hand, each component of B is a smooth embedded hypersurface in Minkowski spacetime, the tangent hyperplane to which at each point is contained in the exterior of the sound cone at that point, with the exception of a single generator of the sound cone, which lies on the hyperplane itself. The past boundary of a component of B is the corresponding component of $\partial_- B$. The latter is at the same time the past boundary of a component of $\underline{C}$. This is the surface where a shock begins to form. Now the maximal development in the acoustical sense, or "maximal classical solution", is the physical solution of the problem up to $\underline{C} \bigcup \partial_- B$, but not up to B. In the last part of [Chr14] the problem of the physical continuation of the solution is set up as the *shock development problem*. This is a free-boundary problem associated to each component of $\partial_- B$. In this problem one is required to construct a hypersurface of discontinuity K, the shock, lying in the past of the corresponding component of B but having the same past boundary as the latter, namely the given component of $\partial_- B$, the tangent hyperplanes to K and B coinciding along $\partial_- B$. Moreover, one is required to construct a solution of the differential conservation laws in the domain in Minkowski spacetime bounded in the past by $\underline{C} \bigcup K$, agreeing with the maximal classical solution on $\underline{C} \bigcup \partial_- B$, while having jumps across K relative to the data induced on K by the maximal classical solution, these jumps satisfying the jump conditions which follow from the integral form of the conservation laws. Finally, K is required to be spacelike relative to the acoustical metric induced by the maximal classical solution, which holds in the past of K, and timelike relative to the new solution, which holds in the future of K. The maximal classical solution thus provides the boundary conditions on $\underline{C} \bigcup \partial_- B$, as well as a barrier at B.

The shock development problem is only set up, not solved, in [Chr14]. The author plans to address this problem in the near future. One final result of [Chr14] needs to be mentioned here however. In the context of [Chr14], the solution is irrotational up to K. At the end of that monograph a formula is derived for the jump in vorticity across K of a solution of the shock development problem. This formula shows that while the flow is irrotational ahead of the shock, it acquires vorticity immediately behind.

This brings us to another problem. This is the problem of the long-time behavior of the vorticity along the fluid flow lines. By reason of the result just quoted, this problem must also be solved to achieve an understanding of the dynamics, even when the initial conditions are restricted to be irrotational. Now, even in the non-relativistic case, and even in the case that the compressibility of the fluid is neglected, this is a very difficult problem. Indeed, the problem, in the context of the incompressible Euler equations, of whether or not the vorticity blows up in finite time along some flow lines, is one of the great unsolved problems of mathematics (see [Co]).

In conclusion, it is clear that the above basic fluid mechanical problems must be solved first, before any attempt is made to address the problem of the general non-spherical gravitational collapse of a perfect fluid in general relativity.

However, once the restriction to spherical symmetry is removed, the dynamical degrees of freedom of the gravitational field itself come into play, and the thought strikes one that we may not need matter at all to form black holes. Even in vacuum, closed trapped surfaces could perhaps be formed by the focusing of sufficiently strong incoming gravitational waves. It is in fact this problem which John Wheeler related to the author back in

1968: *the formation of black holes in pure general relativity, by the focusing of incoming gravitational waves*. And it is this problem the complete solution of which is found in the present monograph. Because of the absence of spherically symmetric solutions of the vacuum Einstein equations other than the Schwarzschild solution, the problem in question was far out of reach at that time, and for this reason John Wheeler advised the author to consider instead the spherically symmetric scalar field problem as a model problem, by solving which, insights would be gained that would prepare us to attack the original problem. Indeed there is some analogy between scalar waves and gravitational waves, but whereas a scalar field is a fiction introduced only for pedagogical reasons, gravitational waves are a fundamental aspect of physical reality. We should remember here the remarks of Einstein in regard to the two sides of his equations. The right-hand side, which involves the energy-momentum-stress tensor of matter, he called "wood", while the left-hand side, the Ricci curvature, he called "marble", recalling, perhaps, the simplicity of an ancient Greek temple.

We shall now state the simplest version of the theorem on the formation of closed trapped surfaces in pure general relativity which this monograph establishes. This is the limiting version, where we have an asymptotic characteristic initial value problem with initial data at past null infinity. Denoting by $\underline{u}$ the "advanced time", it is assumed that the initial data are trivial for $\underline{u} \leq 0$. Our methods allow us to replace this assumption by a suitable falloff condition in $|\underline{u}|$ for $\underline{u} \leq 0$, thereby extending the theorem. This would introduce no new difficulties of principle, but would require more technical work, which would have considerably lengthened the monograph, obscuring the main new ideas.

Let k, l be positive constants, $k > 1$, $l < 1$. Let us be given smooth asymptotic initial data at past null infinity which is trivial for advanced time $\underline{u} \leq 0$. Suppose that the incoming energy per unit solid angle in each direction in the advanced time interval $[0, \delta]$ is not less than $k/8\pi$. Then if δ is suitably small, the maximal development of the data contains a closed trapped surface S which is diffeomorphic to S^2 and has area

$$\text{Area}(S) \geq 4\pi l^2.$$

The form of the smallness assumption on δ is specified in the precise form of the theorem, stated in Chapter 17. We remark that, by virtue of the scale invariance of the vacuum Einstein equations, the theorem holds with k, l, and δ, replaced by ak, al, and $a\delta$, respectively, for any positive constant a.

The above theorem is obtained through a theorem in which the initial data is given on a complete future null geodesic cone C_o. The generators of the cone are parametrized by an affine parameter s measured from the vertex o and defined so that the corresponding null geodesic vectorfield has projection T at o along a fixed unit future-directed timelike vector T at o. It is assumed that the initial data are trivial for $s \leq r_0$, for some $r_0 > 1$. The boundary of this trivial region is then a round sphere of radius r_0. The advanced time $\underline{u}$ is then defined along C_o by

$$\underline{u} = s - r_0. \tag{13}$$

The formation of the closed trapped surfaces theorem is similar in this case, the only difference being that the "incoming energy per unit solid angle in each direction in the

advanced time interval $[0, \delta]$", a notion defined only at past null infinity, is replaced by the integral

$$\frac{r_0^2}{8\pi} \int_0^\delta e d\underline{u} \tag{14}$$

on the affine parameter segment $[r_0, r_0 + \delta]$ of each generator of C_o. The function e is an invariant of the conformal intrinsic geometry of C_o, given by

$$e = \frac{1}{2} |\hat{\chi}|^2_{\not g}, \tag{15}$$

where $\not g$ is the induced metric on the sections of C_o corresponding to constant values of the affine parameter, and $\hat{\chi}$ is the *shear* of these sections, the trace-free part of their 2nd fundamental form relative to C_o. The theorem for a cone C_o is established for any $r_0 > 1$ and the smallness condition on δ is independent of r_0. The domain of dependence, in the maximal development, of the trivial region in C_o is a domain in Minkowski spacetime bounded in the past by the trivial part of C_o and in the future by $\underline{C}_e$, the past null geodesic cone of a point e at arc length $2r_0$ along the timelike geodesic Γ_0 from o with tangent vector T at o. Considering then the corresponding complete timelike geodesic in Minkowski spacetime, fixing the origin on this geodesic to be the point e, the limiting form of the theorem is obtained by letting $r_0 \to \infty$, keeping the origin fixed, so that o tends to the infinite past along the timelike geodesic.

The theorem on the formation of closed trapped surfaces in this monograph may be compared to the corresponding theorem in [Chr6] for the spherically symmetric scalar field problem quoted above. In sharp contrast to that theorem however, here almost all the work goes into establishing an *existence theorem* for a development of the initial data which extends far enough into the future so that trapped spheres have eventually a chance to form within this development. This theorem is first stated as Theorem 12.1 in the way in which it is actually proved, and then restated as Theorem 16.1, in the way in which it can most readily be applied, after the proof is completed. So all chapters of this monograph, with the exception of the last (and shortest) chapter, are devoted to the proof of the existence theorem. On the other hand, there is a wealth of information in Theorem 16.1, which gives us full knowledge of the geometry of spacetime when closed trapped surfaces begin to form. The theorems established in this monograph thus constitute the first foray into the long-time dynamics of general relativity in the large, that is, when the initial data are no longer confined to a suitably small neighborhood of Minkowskian data. However, the existence theorem which we establish does not cover the whole of the maximal development, and for this reason the question regarding the nature of the future boundary of the maximal development is left unanswered.

We shall now give a brief discussion of the mathematical methods employed in this monograph, for, as is generally acknowledged, the methods in a mathematical work are often more important than the results. This monograph relies on three methods, two of which stem from the author's work with Klainerman [C-K] on the stability of the Minkowski spacetime, and the third method is new. We shall first summarize the first two methods.

The work [C-K] which established the global nonlinear stability of the Minkowski spacetime of special relativity within the framework of general relativity, was a work within pure general relativity, concerned, like the present one, with the "marble side" of Einstein's equations, the "wood" side having been set equal to zero. Both of the main mathematical methods employed were new at the time when the work was composed. The first method was peculiar to Einstein's equations, while the second had wider application, and could, in principle, be extended to all Euler-Lagrange systems of partial differential equations of hyperbolic type.

The first method was a way of looking at Einstein's equations which allowed estimates for the spacetime curvature to be obtained. A full exposition of this method is given in Chapter 12, which is also self-contained, except for Propositions 12.1, 12.5 and 12.6, which are quoted directly from [C-K]. Only the barest outline of the chief features will be given here. The method applies also in the presence of matter, to obtain the required estimates for the spacetime curvature. Its present form is dependent on the 4-dimensional nature of the spacetime manifold, although a generalization to higher dimensions can be found.

Instead of considering the Einstein equations themselves, we considered the Bianchi identities in the form which they assume by virtue of the Einstein equations. We then introduced the general concept of a *Weyl field* W on a 4-dimensional Lorentzian manifold (M, g) to be a 4-covariant tensorfield with the algebraic properties of the Weyl or *conformal* curvature tensor. Given a Weyl field W one can define a left dual ${}^*W$ as well as a right dual W^*, but as a consequence of the algebraic properties of a Weyl field, the two duals coincide. Moreover, ${}^*W = W^*$ is also a Weyl field. A Weyl field is subject to equations which are analogues of Maxwell's equations for the electromagnetic field. These are linear equations, in general inhomogeneous, which we call *Bianchi equations*. They are of the form

$$\nabla^\alpha W_{\alpha\beta\gamma\delta} = J_{\beta\gamma\delta}, \tag{16}$$

the right-hand side J, or more generally any 3-covariant tensorfield with the algebraic properties of the right-hand side, we call a *Weyl current*. These equations seem at first sight to be analogues of only half of Maxwell's equations, but it turns out that they are equivalent to the equations

$$\nabla_{[\alpha} W_{\beta\gamma]\delta\epsilon} = \epsilon_{\mu\alpha\beta\gamma} J^{*\mu}{}_{\delta\epsilon}, \quad J^*_{\beta\gamma\delta} = \frac{1}{2} J_\beta{}^{\mu\nu} \epsilon_{\mu\nu\gamma\delta} \tag{17}$$

which are analogues of the other half of Maxwell's equations. Here ϵ is the volume 4-form of (M, g). The fundamental Weyl field is the Riemann curvature tensor of (M, g), (M, g) being a solution of the vacuum Einstein equations, and in this case the corresponding Weyl current vanishes, the Bianchi equations reducing to the Bianchi identities.

Given a vectorfield Y and a Weyl field W or Weyl current J there is a "variation" of W and J with respect to Y, a modified Lie derivative $\tilde{\mathcal{L}}_Y W$, $\tilde{\mathcal{L}}_Y J$, which is also a Weyl field or Weyl current respectively. The modified Lie derivative commutes with duality. The Bianchi equations have certain conformal covariance properties which imply the following. If J is the Weyl current associated to the Weyl field W according to the

Bianchi equations, then the Weyl current associated to $\tilde{\mathcal{L}}_Y W$ is the sum of $\tilde{\mathcal{L}}_Y J$ and a bilinear expression which is on one hand linear in ${}^{(Y)}\tilde{\pi}$ and its first covariant derivative and on the other hand also linear in W and its first covariant derivative (see Proposition 12.1). Here we denote by ${}^{(Y)}\tilde{\pi}$ the *deformation tensor* of Y, namely the trace-free part of the Lie derivative of the metric g with respect to Y. This measures the rate of change of the conformal geometry of (M, g) under the flow generated by Y. From the fundamental Weyl field, the Riemann curvature tensor of (M, g), and a set of vector fields $Y_1, \ldots, Y_n$ which we call *commutation fields*, derived Weyl fields of up to any given order m are generated by the repeated application of the operators $\tilde{\mathcal{L}}_{Y_i} : i = 1, \ldots, n$. A basic requirement on the set of commutation fields is that it spans the tangent space to M at each point. The Weyl currents associated to these derived Weyl fields are then determined by the deformation tensors of the commutation fields.

Given a Weyl field W there is a 4-covariant tensorfield $Q(W)$ associated to W, which is symmetric and trace-free in any pair of indices. It is a quadratic expression in W, analogous to the Maxwell energy-momentum-stress tensor for the electromagnetic field. We call $Q(W)$ the *Bel-Robinson tensor* associated to W, because it was discovered by Bel and Robinson [Be] in the case of the fundamental Weyl field, the Riemann curvature tensor of a solution of the vacuum Einstein equations. The Bel-Robinson tensor has a remarkable positivity property: $Q(W)(X_1, X_2, X_3, X_4)$ is non-negative for any tetrad X_1, X_2, X_3, X_4 of future-directed non-spacelike vectors at a point. Moreover, the divergence of $Q(W)$ is a bilinear expression which is linear in W and in the associated Weyl current J (see Proposition 12.6). Given a Weyl field W and a triplet of future directed non-spacelike vectorfields X_1, X_2, X_3, which we call *multiplier fields*, we define the *energy-momentum density* vectorfield $P(W; X_1, X_2, X_3)$ associated to W and to the triplet X_1, X_2, X_3 by:

$$P(W; X_1, X_2, X_3)^\alpha = -Q(W)^\alpha_{\beta\gamma\delta} X_1^\beta X_2^\gamma X_3^\delta. \tag{18}$$

Then the divergence of $P(W; X_1, X_2, X_3)$ is the sum of $-(\mathrm{div} Q(W))(X_1, X_2, X_3)$ and a bilinear expression which is linear in $Q(W)$ and in the deformation tensors of X_1, X_2, X_3. The divergence theorem in spacetime, applied to a domain which is a development of part of the initial hypersurface, then expresses the integral of the 3-form dual to $P(W; X_1, X_2, X_3)$ on the future boundary of this domain, in terms of the integral of the same 3-form on the past boundary of the domain, namely on a part of the initial hypersurface, and the spacetime integral of the divergence. The boundaries being *achronal* (that is, no pair of points on each boundary can be joined by a timelike curve) the integrals are integrals of non-negative functions, by virtue of the positivity property of $Q(W)$. For the set of Weyl fields of order up to m which are derived from the fundamental Weyl field, the Riemann curvature tensor of (M, g), the divergences are determined by the deformation tensors of the commutation fields and their derivatives up to order m, and from the deformation tensors of the multiplier fields. And the integrals on the future boundary give control of all the derivatives of the curvature up to order m. This is how estimates for the spacetime curvature are obtained, once a suitable set of multiplier fields and a suitable set of commutation fields have been provided.

This is precisely where the second method comes in. This method constructs the required sets of vectorfields by using the geometry of the two-parameter foliation of spacetime by the level sets of two functions. These two functions, in the first realization of this method in [C-K], where the *time function* t, the level sets of which are maximal spacelike hypersurfaces H_t of vanishing total momentum, and the *optical function* u, which we may think of as "retarded time", the level sets of which are outgoing null hypersurfaces C_u. These are chosen so that density of the foliation of each H_t by the traces of the C_u, that is, by the surfaces of intersection $S_{t,u} = H_t \bigcap C_u$, which are diffeomorphic to S^2, tends to 1 as $t \to \infty$. In other words, the $S_{t,u}$ on each H_t become evenly spaced in the limit $t \to \infty$. It was already clear at the time of composition of the work [C-K] that the two functions did not enter the problem on equal footing. The optical function u played a much more important role. This is due to the fact that the problem involved outgoing waves reaching future null infinity, and it is the outgoing family of null hypersurfaces C_u which follows these waves. The role of the family of maximal spacelike hypersurfaces H_t was to obtain a suitable family of sections of each C_u, the family $S_{t,u}$ corresponding to a given u, to provide the future boundary, or part of the future boundary, of domains where the divergence theorem is applied, and also to serve as a means by which, in the proof of the existence theorem, the method of continuity can be applied. The geometric entities describing the two-parameter foliation of spacetime by the $S_{t,u}$ are estimated in terms of the spacetime curvature. This yields estimates for the deformation tensors of the multiplier fields and the commutation fields in terms of the spacetime curvature, thus connecting with the first method.

Another realization of this method is found in [Chr14]. There in the role of the time function we have the Minkowskian time coordinate t which vanishes on the initial hyperplane. The level sets of this function are then a family H_t of parallel spacelike hyperplanes in Minkowski spacetime. In the role of the optical function we have the *acoustical function* u, the level sets C_u of which are outgoing characteristic hypersurfaces relative to the acoustical metric h. In this case however, these are defined by their traces $S_{0,u}$ on the initial hyperplane H_0, which are diffeomorphic to S^2. The density of the foliation of each H_t by the traces of the C_u, that is, by the surfaces of intersection $S_{t,u} = H_t \bigcap C_u$, in fact blows up in finite time $t^*(u, \vartheta)$ for (u, ϑ) in an open subset of $\Re \times S^2$, $\vartheta \in S^2$ labelling the generators of each C_u, and this defines the singular boundary B, whose past boundary $\partial_- B$ is the surface, not necessarily connected, from which shocks begin to form. The relative roles of the two functions are even clearer in this work, because the blowup of the density of foliations by outgoing characteristic hypersurfaces is what characterizes shock formation.

Returning to general relativity, a variant of the method is obtained if we place in the role of the time function t another *optical function* $\underline{u}$, which we may think of as "advanced time", the level sets of which are incoming null hypersurfaces. This approach had its origin in the author's effort to understand the so-called "memory effect" of gravitational waves [Chr15]. This effect is a manifestation of the nonlinear nature of the asymptotic gravitational laws at future null infinity. Now future null infinity is an ideal incoming null hypersurface at infinity, so the analysis required consideration of a family of incoming null hypersurfaces, the interiors of the traces of which on the initial spacelike hypersurface

H_0 give an exhaustion of H_0. A two-parameter family of surfaces diffeomorphic to S^2, the "wave fronts", was then obtained, namely the intersections of this incoming family with the outgoing family of null hypersurfaces. A set of notes [Chr16] was then written up where the basic structure equations of such a "double null" foliation were derived, including the propagation equations for the mass aspect functions (see below).

A double null foliation was subsequently employed by Klainerman and Nicolò in [K-N] (where the aforementioned notes are gratefully acknowledged) to provide a simpler variant of the exterior part of the proof of the stability of Minkowski spacetime, namely that part which considers the domain of dependence of the exterior of a compact set in the initial asymptotically flat spacelike hypersurface. The developments stemming from the original work [C-K] include the work of Zipser [Z], which extended the original theorem to the Einstein-Maxwell equations, and most recently the work of Bieri [Bie], which extended the theorem in vacuum by requiring a smallness condition only up to the 1st derivatives of the Ricci curvature of $\overline{g}_0$, the induced metric on H_0, and up to the 2nd derivatives of k_0, the 2nd fundamental form of H_0, instead of up to the 2nd and up to the 3rd derivatives respectively, as in the original theorem; moreover the respective weights depend on the distance from an origin on $(H_0, \overline{g}_0)$, which are reduced by one power of this distance, relative to the weights assumed in [C-K].

In the present work, the roles of the two optical functions are reversed, because we are considering incoming rather than outgoing waves, and it is the incoming null hypersurfaces $\underline{C}_{\underline{u}}$, the level sets of $\underline{u}$, which follow these waves. However, in the present work, taking the other function to be the conjugate optical function u is not merely a matter of convenience, but it is essential for what we wish to achieve. This is because the C_u, the level sets of u, are here, like the initial hypersurface C_o itself, future null geodesic cones with vertices on the timelike geodesic Γ_0, and the trapped spheres which eventually form are sections $S_{\underline{u},u} = \underline{C}_{\underline{u}} \bigcap C_u$ of "late" C_u, everywhere along which those C_u have negative expansion.

We now come to the new method. This method is a method of treating the focusing of incoming waves, and like the second method it is of wider application. A suitable name for this method is *short pulse method*. Its point of departure resembles that of the short wavelength or geometric optics approximation, in so far as it depends on the presence of a small length, but thereafter the two approaches diverge. The short pulse method is a method which, in the context of Euler-Lagrange systems of partial differential equations of hyperbolic type, allows us to establish an existence theorem for a development of the initial data which is large enough so that interesting things have a chance to occur within this development, if a nonlinear system is involved. One may ask at this point: what does it mean for a length to be small in the context of the vacuum Einstein equations? For, the equations are scale invariant. Here *small* means *by comparison to the area radius of the trapped sphere to be formed*.

With initial data on a complete future null geodesic cone C_o, as explained above, which are trivial for $s \leq r_0$, we consider the restriction of the initial data to $s \leq r_0 + \delta$. In terms of the advanced time $\underline{u}$, we restrict attention to the interval $[0, \delta]$, the data being trivial for $\underline{u} \leq 0$. The retarded time u is set equal to $u_0 = -r_0$ at o and therefore on C_o, which is then also denoted C_{u_0}. Also, $u - u_0$ is defined along Γ_0 to be one-half the

arc length from o. This determines u everywhere. The development whose existence we want to establish is that bounded in the future by the spacelike hypersurface H_{-1} where $\underline{u} + u = -1$ and by the incoming null hypersurface $\underline{C}_\delta$. We denote this development by M_{-1}. We define L and $\underline{L}$ to be the future-directed null vectorfields, the integral curves of which are the generators of the C_u and $\underline{C}_{\underline{u}}$, parametrized by $\underline{u}$ and u respectively, so that

$$L u = \underline{L}\underline{u} = 0, \quad L\underline{u} = \underline{L} u = 1. \tag{19}$$

The flow Φ_τ generated by L defines a diffeomorphism of $S_{\underline{u},u}$ onto $S_{\underline{u}+\tau,u}$, while the flow $\underline{\Phi}_\tau$ generated by $\underline{L}$ defines a diffeomorphism of $S_{\underline{u},u}$ onto $S_{\underline{u},u+\tau}$. The positive function Ω defined by

$$g(L, \underline{L}) = -2\Omega^2 \tag{20}$$

may be thought of as the inverse density of the double null foliation. We denote by $\hat{L}$ and $\underline{\hat{L}}$ the corresponding normalized future-directed null vectorfields

$$\hat{L} = \Omega^{-1} L, \quad \underline{\hat{L}} = \Omega^{-1} \underline{L}, \quad \text{so that } g(\hat{L}, \underline{\hat{L}}) = -2. \tag{21}$$

The first step is the analysis of the equations along the initial hypersurface C_{u_0}. This analysis is performed in Chapter 2, and it is particularly clear and simple because of the fact that C_{u_0} is a null hypersurface, so we are dealing with the characteristic initial value problem and there is a way of formulating the problem in terms of free data which are not subject to any constraints. The full set of data, which includes all the curvature components and their transversal derivatives, up to any given order along C_{u_0}, is then determined by integrating ordinary differential equations along the generators of C_{u_0}. As we shall see in Chapter 2, the free data may be described as a 2-covariant symmetric positive definite tensor density m, of weight -1 and unit determinant, on S^2, depending on $\underline{u}$. This is of the form:

$$m = \exp \psi \tag{22}$$

where ψ is a 2-dimensional symmetric trace-free matrix-valued "function" on S^2, depending on $\underline{u} \in [0, \delta]$, and transforming under change of charts on S^2 in such a way as to make m a 2-covariant tensor density of weight -1. The transformation rule is particularly simple if stereographic charts on S^2 are used. Then there is a function O defined on the intersection of the domains of the north and south polar stereographic charts on S^2, with values in the 2-dimensional symmetric orthogonal matrices of determinant -1 such that in going from the north polar chart to the south polar chart or vice-versa, $\psi \mapsto \tilde{O} \psi O$ and $m \mapsto \tilde{O} m O$. The crucial ansatz of the short pulse method is the following. We consider an arbitrary smooth 2-dimensional symmetric trace-free matrix-valued "function" ψ_0 on S^2, depending on $s \in [0, 1]$, which extends smoothly by 0 to $s \leq 0$, and we set

$$\psi(\underline{u}, \vartheta) = \frac{\delta^{1/2}}{|u_0|} \psi_0 \left(\frac{\underline{u}}{\delta}, \vartheta \right), \quad (\underline{u}, \vartheta) \in [0, \delta] \times S^2. \tag{23}$$

The analysis of the equations along C_{u_0} then gives, for the components of the spacetime curvature along C_{u_0}:

$$\sup_{C_{u_0}} |\alpha| \leq O_2(\delta^{-3/2} |u_0|^{-1}),$$

$$\begin{aligned} \sup_{C_{u_0}} |\beta| &\leq O_2(\delta^{-1/2}|u_0|^{-2}), \\ \sup_{C_{u_0}} |\rho|, \sup_{C_{u_0}} |\sigma| &\leq O_3(|u_0|^{-3}), \\ \sup_{C_{u_0}} |\underline{\beta}| &\leq O_4(\delta|u_0|^{-4}), \\ \sup_{C_{u_0}} |\underline{\alpha}| &\leq O_5(\delta^{3/2}|u_0|^{-5}). \end{aligned} \tag{24}$$

Here α, $\underline{\alpha}$ are the trace-free symmetric 2-covariant tensorfields on each $S_{\underline{u},u}$ defined by

$$\alpha(X,Y) = R(X,\hat{L},Y,\hat{L}), \quad \underline{\alpha}(X,\underline{\hat{L}},Y,\underline{\hat{L}}) = R(X,\underline{\hat{L}},Y,\underline{\hat{L}}) \tag{25}$$

for any pair of vectors X, Y tangent to $S_{\underline{u},u}$ at a point; β, $\underline{\beta}$ are the 1-forms on each $S_{\underline{u},u}$ defined by

$$\beta(X) = \frac{1}{2}R(X,\hat{L},\underline{\hat{L}},\hat{L}), \quad \underline{\beta}(X) = \frac{1}{2}R(X,\underline{\hat{L}},\underline{\hat{L}},\hat{L}), \tag{26}$$

and ρ, σ are the functions on each $S_{\underline{u},u}$ defined by

$$\rho = \frac{1}{4}R(\underline{\hat{L}},\hat{L},\underline{\hat{L}},\hat{L}), \quad \frac{1}{2}R(X,Y,\underline{\hat{L}},\hat{L}) = \sigma \not{\epsilon}(X,Y) \tag{27}$$

for any pair of vectors X, Y tangent to $S_{\underline{u},u}$ at a point, $\not{\epsilon}$ being the area form of $S_{\underline{u},u}$. The symbol $O_k(\delta^p|u_0|^r)$ means the product of $\delta^p|u_0|^r$ with a non-negative non-decreasing continuous function of the C^k norm of ψ_0 on $[0,1] \times S^2$. The pointwise magnitudes of tensors on $S_{\underline{u},u}$ are with respect to the induced metric $\not{g}$, which is positive definite, the surfaces being spacelike. The precise estimates are given in Chapter 2. We should emphasize here that the role of the ansatz (23) is to obtain estimates of the form (24), that is with the same dependence on δ and $|u_0|$, and analogous estimates for the L^4 norms on the $S_{\underline{u},u_0}$, $\underline{u} \in [0,\delta]$, of the 1st derivatives of the curvature components, and for the L^2 norms on C_{u_0} of the 2nd derivatives of the curvature components (with the exception of the 2nd transversal derivative of $\underline{\alpha}$). If the quantities

$$\begin{gathered} \sup_{C_{u_0}} \left(\delta^{3/2}|u_0||\alpha|\right), \\ \sup_{C_{u_0}} \left(\delta^{1/2}|u_0|^2|\beta|\right), \\ \sup_{C_{u_0}} \left(|u_0|^3|\rho|\right), \sup_{C_{u_0}} \left(|u_0|^3|\sigma|\right), \\ \sup_{C_{u_0}} \left(\delta^{-1}|u_0|^4|\underline{\beta}|\right), \\ \sup_{C_{u_0}} \left(\delta^{-3/2}|u_0|^{9/2}|\underline{\alpha}|\right), \end{gathered} \tag{28}$$

and analogous quantities for the 1st and 2nd derivatives, are assumed to have bounds which are independent of $|u_0|$ or δ, the ansatz (23) can be dispensed with, and indeed the chapters following Chapter 2 make no reference to it, until, at the end of Chapter 16, the existence theorem is restated, after it has been proven, as Theorem 16.1. However the

ansatz (23) is the simplest way to ensure that the required bounds hold, and there is no loss of generality involved, ψ_0 being an arbitrary "function" on $[0, 1] \times S^2$ with values in the 2-dimensional symmetric trace-free matrices. Note here that, since $|u_0| > 1$, what is required of the last of (28) is weaker than what is provided by the last of (24). A last remark before we proceed to the main point is that the last three of the estimates (24) require more than two derivatives of ψ_0, so there is an apparent loss of derivatives from what would be expected of curvature components. This loss of derivatives is intrinsic to the characteristic initial value problem and occurs even for the wave equation in Minkowski spacetime (see [M]). It is due to the fact that one expresses the full data, which includes transversal derivatives of any order, in terms of the free data. No such loss of derivatives is present in our spacetime estimates, which are sharp, and depend only on the L^2 norm of up to the 2nd derivatives of the curvature components on the initial hypersurface (with the exception of the 2nd transversal derivative of $\underline{\alpha}$), precisely as in [C-K]. Nevertheless the initial data are assumed to be C^∞ in this work, and the solutions which we construct are also C^∞.

To come to the main point, the reader should focus on the dependence on δ of the right-hand sides of (24). This displays what we may call the *short pulse hierarchy*. And this hierarchy is *nonlinear*. For, if only the linearized form of the equations was considered, a different hierarchy would be obtained: the exponents of δ in the first two of (24) would be the same, but the exponents of δ in the last three of (24) would instead be $1/2, 3/2, 5/2$, respectively.

A question that immediately comes up when one ponders the ansatz (24), is why is the "amplitude" of the pulse proportional to the square root of the "length" of the pulse? (the factor $|u_0|^{-1}$ is the standard decay factor in three spatial dimensions, the square root of the area of the wave fronts). Where does this relationship come from? Obviously, there is no such relationship in a linear theory. The answer is that it comes from our desire to form trapped surfaces in the development M_{-1}. If a problem involving the focusing of incoming waves in a different context was the problem under study, for example the formation of electromagnetic shocks by the focusing of incoming electromagnetic waves in a nonlinear medium, the relationship between length and amplitude would be dictated by the desire to form such shocks within our development.

Another remark concerning different applications of the short pulse method, in particular applications to problems of shock formation, is that it is more natural in these problems to use, in the role of the retarded time u, the time function t whose level sets are parallel spacelike hyperplanes of the background Minkowski metric, as in [Chr14]. However the analysis of the equations along an outgoing characteristic hypersurface is indispensable as a crucial step of the short pulse method, because, once the correct relationship between length and amplitude has been guessed, it is this analysis which yields the short pulse hierarchy.

The short pulse hierarchy is the key to the existence theorem as well as to the trapped surface formation theorem. We must still outline however in what way we establish that the short pulse hierarchy is preserved in evolution. This is of course the main step of the short pulse method. What we do is to reconsider the first two methods previously outlined in the light of the short pulse hierarchy.

Let us revisit the first method. We take as multiplier fields the vectorfields L and K, where

$$K = u^2 \underline{L}. \tag{29}$$

In this monograph, as already mentioned above, we take the initial data to be trivial for $\underline{u} \leq 0$ and as a consequence the spacetime region corresponding to $\underline{u} \leq 0$ is a domain in Minkowski spacetime. We may thus confine attention to the nontrivial region $\underline{u} \geq 0$. We denote by M'_{-1} this non-trivial region in M_{-1}. To extend the theorem to the case where the data is non-trivial for $\underline{u} \leq 0$ but satisfy a suitable falloff condition in $|\underline{u}|$, in the region $\underline{u} \leq 0$ we replace L as a multiplier field by

$$\underline{L} + L = 2T \tag{30}$$

and redefine K to be

$$K = u^2 \underline{L} + \underline{u}^2 L. \tag{31}$$

Since, in any case, a smallness condition can be imposed on the part of the data corresponding to $\underline{u} \leq 0$, we already know from the work on the stability of Minkowski spacetime that in the associated domain of dependence, that is, in the spacetime region $\underline{u} \leq 0$, the solution will satisfy a corresponding smallness condition. In particular the said smallness condition will be satisfied along $\underline{C}_0$, and this suffices for us to proceed with our estimates in the region $\underline{u} \geq 0$ with the multiplier fields L and K, with K as in (29). So all the difficulty lies in the region M'_{-1} where the pulse travels.

For each of the Weyl fields to be specified below, we define the energy-momentum density vectorfields

$$\overset{(n)}{P}(W) \ : \ n = 0, 1, 2, 3 \tag{32}$$

where

$$\begin{aligned}
&\overset{(0)}{P}(W) = P(W; L, L, L),\\
&\overset{(1)}{P}(W) = P(W; K, L, L),\\
&\overset{(2)}{P}(W) = P(W; K, K, L),\\
&\overset{(3)}{P}(W) = P(W; K, K, K).
\end{aligned} \tag{33}$$

As commutation fields we take L, S, defined by

$$S = u\underline{L} + \underline{u}L, \tag{34}$$

and the three rotation fields $O_i \ : i = 1, 2, 3$. The latter are defined according to the second method as follows. In the Minkowskian region we introduce rectangular coordinates $x^\mu \ : \mu = 0, 1, 2, 3$, taking the x^0 axis to be the timelike geodesic Γ_0. In the Minkowskian region, in particular on the sphere S_{0,u_0}, the O_i are the generators of rotations about the $x^i \ : i = 1, 2, 3$ spatial coordinate axes. The O_i are then first defined on C_{u_0} by conjugation with the flow of L and then in spacetime by conjugation with the flow of $\underline{L}$. The

Weyl fields which we consider are, besides the fundamental Weyl field R, the Riemann curvature tensor, the following derived Weyl fields:

$$\begin{aligned}&\text{1st-order: } \tilde{\mathcal{L}}_L R,\ \tilde{\mathcal{L}}_{O_i} R : i = 1, 2, 3,\ \tilde{\mathcal{L}}_S R.\\ &\text{2nd-order: } \tilde{\mathcal{L}}_L \tilde{\mathcal{L}}_L R,\ \tilde{\mathcal{L}}_{O_i} \tilde{\mathcal{L}}_L R : i = 1, 2, 3,\ \tilde{\mathcal{L}}_{O_j} \tilde{\mathcal{L}}_{O_i} R : i, j = 1, 2, 3,\\ &\qquad \tilde{\mathcal{L}}_{O_i} \tilde{\mathcal{L}}_S R : i = 1, 2, 3,\ \tilde{\mathcal{L}}_S \tilde{\mathcal{L}}_S R. \end{aligned} \tag{35}$$

We assign to each Weyl field the index l according to the number of $\tilde{\mathcal{L}}_L$ operators in the definition of W in terms of R. We then define total 2nd-order energy-momentum densities

$$\overset{(n)}{P}_2 : n = 0, 1, 2, 3 \tag{36}$$

as the sum of $\delta^{2l} \overset{(n)}{P}(W)$ over all the above Weyl fields in the case $n = 3$, all the above Weyl fields except those whose definition involves the operator $\tilde{\mathcal{L}}_S$ in the cases $n = 0, 1, 2$. We then define the total 2nd-order energies $\overset{(n)}{E}_2(u)$ as the integrals on the C_u and the total 2nd-order fluxes $\overset{(n)}{F}_2(\underline{u})$ as the integrals on the $\underline{C}_{\underline{u}}$, of the 3-forms dual to the $\overset{(n)}{P}_2$. Of the fluxes, only $\overset{(3)}{F}_2(\underline{u})$ plays a role in the problem. Finally, with the exponents $q_n : n = 0, 1, 2, 3$ defined by

$$q_0 = 1,\ \ q_1 = 0,\ \ q_2 = -\frac{1}{2},\ \ q_3 = -\frac{3}{2}, \tag{37}$$

according to the short pulse hierarchy, we define the quantities

$$\overset{(n)}{\mathcal{E}}_2 = \sup_u \left(\delta^{2q_n} \overset{(n)}{E}_2(u) \right) : n = 0, 1, 2, 3, \qquad \overset{(3)}{\mathcal{F}}_2 = \sup_{\underline{u}} \left(\delta^{2q_3} \overset{(3)}{F}_2(\underline{u}) \right). \tag{38}$$

The objective then is to obtain bounds for these quantities in terms of the initial data.

This requires properly estimating the deformation tensor of K, as well as the deformation tensors of L, S and the $O_i : i = 1, 2, 3$ and their derivatives of up to 2nd-order. In doing this, the short pulse method meshes with the second method previously described. This is the content of Chapters 3–9 and shall be very briefly described in the outline of the contents of each chapter which follows.

The estimates of the error integrals, namely the integrals of the absolute values of the divergences of the $\overset{(n)}{P}_2$, which is the content of Chapters 13–15, then yield inequalities for the quantities (38). These inequalities contain, besides the initial data terms

$$\overset{(n)}{D} = \delta^{2q_n} \overset{(n)}{E}_2(u_0) : n = 0, 1, 2, 3, \tag{39}$$

terms of $O(\delta^p)$ for some $p > 0$, which are innocuous, as they can be made less than or equal to 1 by subjecting δ to a suitable smallness condition, *but they also contain terms which are nonlinear in the quantities* (38). From such a nonlinear system of inequalities,

no bounds can in general be deduced, because here, in contrast with [C-K], the initial data quantities are allowed to be arbitrarily large. However a miracle occurs: our system of inequalities is *reductive*. That is, the inequalities, taken in proper sequence, reduce to a sequence of sublinear inequalities, thus allowing us to obtain the bounds we sought.

We remark that although the first two methods on which the present work is based stem from the work [C-K], it is only in the present work, in conjunction with the new method, that the full power of the original methods is revealed.

In applying the short pulse method to problems in other areas of the field of partial differential equations of hyperbolic type, an analogue of the first method is needed. This is supplied in the context of Euler-Lagrange systems, that is, systems of partial differential equations derivable from an action principle, by the structures studied in [Chr17]. The analogue of the concept of a Weyl field is the general concept of *variation*, or variation through solutions. The analogue of the Bel-Robinson tensor is the *canonical stress* associated to such variations. In the area of continuum mechanics or the electrodynamics of continuous media, the fundamental variation is that with respect to a subgroup of the Poincaré group of the underlying Minkowski spacetime, while the higher-order variations are generated by the commutation fields, as in general relativity (see [Chr14]). A particularly interesting problem that may be approached on the basis of the methods which we have discussed, in conjunction with ideas from [Chr14], is the formation of electromagnetic shocks by the focusing of incoming electromagnetic waves in isotropic nonlinear media, that is, media with a nonlinear relationship between the electromagnetic field and the electromagnetic displacement. In this problem, unlike the problem of shock formation by outgoing compression waves in fluid mechanics, there is a *threshold* for shock formation, as there is a threshold for closed trapped sphere formation in the present monograph.

We shall now give a brief outline of the contents of the different chapters of this monograph and of their logical connections. The basic geometric construction, the structure equations of the double null foliation called the "optical structure equations", and the Bianchi identities, are presented in the introductory Chapter 1. The Einstein equations are contained in the optical structure equations. The basic geometric entities associated to the double null foliation are the inverse density function Ω, the metric $\not{g}$ induced on the surfaces $S_{\underline{u},u}$ and its Gauss curvature K, the second fundamental forms χ and $\underline{\chi}$ of $S_{\underline{u},u}$ relative to C_u and $\underline{C}_{\underline{u}}$ respectively, the torsion forms η and $\underline{\eta}$ of $S_{\underline{u},u}$ relative to C_u and $\underline{C}_{\underline{u}}$ respectively, and the functions ω and $\underline{\omega}$, the derivatives of $\log\Omega$ with respect to L and $\underline{L}$ respectively. The torsion forms are given by

$$\eta = \zeta + \not{d}\log\Omega, \qquad \underline{\eta} = -\zeta + \not{d}\log\Omega, \tag{40}$$

where ζ may be called *the* torsion. It is the obstruction to integrability of the distribution of planes orthogonal to the tangent planes to the $S_{\underline{u},u}$. In (40) $\not{d}$ denotes the differential of the restriction of a function to any given $S_{\underline{u},u}$. The optical entities χ, $\underline{\chi}$, η, $\underline{\eta}$, ω, $\underline{\omega}$ are called *connection coefficients* in the succeeding chapters, to emphasize their differential order, intermediate between the metric entities Ω and $\not{g}$, and the curvature entities K and the spacetime curvature components. "Canonical coordinates" are defined in the last section of Chapter 1 and play a basic role in this monograph.

The subject of Chapter 2 is the characteristic initial data and the derivation of the estimates for the full data in terms of the free data. This is where the ansatz (23) is introduced and the short pulse hierarchy first appears. Thus Chapter 2 is fundamental to the whole work.

Chapters 3–7 form a unity. The subject of these chapters is the derivation of estimates for the connection coefficients in terms of certain quantities defined by the spacetime curvature. These chapters are in logical sequence, which extends to Chapters 8 and 9, however the place of the whole sequence of Chapters 3–9 in the logic of the proof of the existence theorem, Theorem 12.1, is *after* Chapters 10 and 11. This is because the assumptions on which Chapters 3–9 rely, namely the boundedness of the quantities defined by the spacetime curvature, is established, in the course of the proof of Theorem 12.1, through the comparison lemmas, Lemmas 12.5 and 12.6, which make use of the results of Chapters 10 and 11. Thus, Chapter 10 represents a *new beginning*. The chapters following Chapter 12 are again in logical sequence.

Chapters 3–7 are divided by Chapter 5 into the two pairs of chapters, on one hand Chapters 3 and 4, and on the other hand Chapters 6 and 7, each of these two pairs forming a tighter unity. The first pair considers only the *propagation equations* among the optical structure equations. These are ordinary differential equations for the connection coefficients along the generators of the C_u and the $\underline{C}_{\underline{u}}$. The second pair considers *coupled systems*, *ordinary differential equations* along the generators of the C_u or the $\underline{C}_{\underline{u}}$ *coupled to elliptic systems* on their $S_{\underline{u},u}$ sections. This allows us to obtain estimates for the connection coefficients which are of one order higher than those obtained through the propagation equations, and are optimal from the point of view of differentiability. There is however a loss of a factor of $\delta^{1/2}$ in behavior with respect to δ, in comparison to the estimates obtained through the propagation equations, in the case of the entities η, $\underline{\eta}$ and ω. What is crucial is that there is no such loss in the case of the entities χ, $\underline{\chi}$, and $\underline{\omega}$, but the proof of this fact again uses the former estimates.

In Chapter 3 the basic L^∞ estimates for the connection coefficients are obtained. The last section of Chapter 3 explains the nature of smallness conditions on δ throughout the monograph. Chapter 4 derives L^4 on the surfaces $S_{\underline{u},u}$ for the 1st derivatives of the connection coefficients.

Chapter 5 is concerned with the isoperimetric and Sobolev inequalities on the surfaces $S_{\underline{u},u}$, and with L^p elliptic theory on these surfaces for $2 < p < \infty$. The main part of the chapter is concerned with the proof of the uniformization theorem for a 2-dimensional Riemannian manifold $(S, \not{g})$ with S diffeomorphic to S^2, when only an L^2 bound on the Gauss curvature K is assumed. The reason why this is required is that although the Gauss equation gives us L^∞ control on K, the estimate is not suitable for our purposes because it involves the loss of a factor of $\delta^{1/2}$ in behavior with respect to δ. Thus one can only rely on the estimate obtained by integrating a propagation equation, which although optimal from the point of view of behavior with respect to δ, only gives us L^2 control on K.

The L^p elliptic theory on the $S_{\underline{u},u}$ is applied in Chapter 6, in the case $p = 4$, to the elliptic systems mentioned above, to obtain L^4 estimates for the 2nd derivatives of the connection coefficients on the surfaces $S_{\underline{u},u}$. What makes possible the gain of one degree

of differentiability by considering systems of ordinary differential equations along the generators of the C_u or the $\underline{C}_{\underline{u}}$ coupled to elliptic systems on the $S_{\underline{u},u}$ sections, is the fact that the principal terms in the propagation equations for certain optical entities vanish, by virtue of the Einstein equations. In the case of the coupled system pertaining to χ and $\underline{\chi}$, these entities are simply the traces $\mathrm{tr}\chi$ and $\mathrm{tr}\underline{\chi}$, and the Codazzi equations constitute the elliptic systems for the trace-free parts $\hat{\chi}$ and $\hat{\underline{\chi}}$ respectively. In the case of the coupled systems pertaining to η and $\underline{\eta}$, the entities are found at one order of differentiation higher. They are the *mass aspect functions* μ and $\underline{\mu}$, called by this name because of the fact, shown in [Chr15], that with r being the area radius of the $S_{\underline{u},u}$, the limits of the functions $r^3\mu/8\pi$ and $r^3\underline{\mu}/8\pi$ at past and future null infinity respectively, represent mass-energy per unit solid angle in a given direction and at a given advanced or retarded time respectively. The elliptic systems are Hodge systems, constituted by one of the structure equations and by the definition of μ and $\underline{\mu}$ in terms of η and $\underline{\eta}$ respectively. Moreover, the two sets of coupled systems, that for η on the C_u and that for $\underline{\eta}$ on the $\underline{C}_{\underline{u}}$, are themselves coupled. (The propagation equations for η and $\underline{\eta}$ studied in Chapters 3 and 4 are similarly coupled.) In the case of the coupled systems pertaining to ω and $\underline{\omega}$, the entities which satisfy propagation equations in which the principal terms vanish are found at one order of differentiation still higher. They are the functions ϕ and $\underline{\phi}$ and the elliptic equations are simply the definitions of these functions in terms of ω and $\underline{\omega}$ respectively.

In the case of the χ system we have, besides what has already been described, also a coupling with the propagation equation for the Gauss curvature K, through the elliptic theory of Chapter 5 applied to the Codazzi elliptic system for $\hat{\chi}$.

Chapter 7 applies L^2 elliptic theory on the $S_{\underline{u},u}$ to the same coupled systems to obtain L^2 estimates on the C_u for the third derivatives of the connection coefficients, the top order needed to obtain a closed system of inequalities in the proof of the existence theorem.

One general remark concerning the contents of Chapters 3–7 is that, although some of the general structure was already encountered in the work on the stability of Minkowski spacetime, the estimates and their derivation are here quite different, and for two reasons. One is the obvious reason that some of the geometric properties are here very different, in view of the fact that we are no longer confined to a suitably small neighborhood of Minkowski spacetime and closed trapped surfaces eventually form. The second is the fact, in connection with the short pulse method, that behavior with respect to δ is here all-important.

In Chapter 8 the multiplier fields and the commutation fields are defined and L^∞ estimates for their deformation tensors are obtained. In Chapter 9, L^4 estimates on the $S_{\underline{u},u}$ for the 1st derivatives of these deformation tensors and L^2 estimates on the C_u for their 2nd derivatives are obtained.

In Chapters 3–9 the symbol $O(\delta^p|u|^r)$ denotes the product of $\delta^p|u|^r$ with a non-negative, non-decreasing continuous function of certain initial data and spacetime curvature quantities $q_1, \ldots, q_n$. The set of quantities $\{q_1, \ldots, q_n\}$ is gradually enlarged as we proceed through the sequence of chapters. The set of quantities is replaced in the seventh section of Chapter 12 by a set which includes only initial data quantities, and it is in this

new sense that the symbol $O(\delta^p|u|^r)$ is meant throughout the proof of Theorem 12.1, which occupies the four succeeding chapters.

As we mention above, Chapter 10 represents a new beginning. The point is the following. Chapters 3–7 derive estimates for the connection coefficients in terms of quantities involving the L^∞ norms on the $S_{\underline{u},u}$ of the curvature components, the L^4 norms on the $S_{\underline{u},u}$ of the 1st derivatives of the curvature components, and the L^2 norms on the C_u of the 2nd derivatives of the curvature components (with the exception of those involving $\underline{\alpha}$ and the 2nd transversal derivatives of $\underline{\beta}$). The first two are to be estimated in terms of the last through Sobolev inequalities on the C_u (except for the quantities involving $\underline{\alpha}$, which are estimated in terms of the L^2 norm on the $\underline{C}_{\underline{u}}$ of up to the 2nd derivatives of that component through a Sobolev inequality on the $\underline{C}_{\underline{u}}$), but in establishing these Sobolev inequalities one cannot rely on the results of the preceding chapters, otherwise the reasoning would be circular. So, the Sobolev inequalities on the C_u and the $\underline{C}_{\underline{u}}$ are instead established on the basis of certain *bootstrap* assumptions. The sharp form of the Sobolev inequality on the C_u given by Proposition 10.1 fits perfectly with the short pulse method and is essential to its success.

The subject of Chapter 11 is the *coercivity* properties of the operators $\not{\mathcal{L}}_{O_i} : i = 1, 2, 3$, the Lie derivatives of covariant tensorfields on the $S_{\underline{u},u}$ with respect to the rotation fields $O_i : 1 = 1, 2, 3$. These inequalities show that, for $m = 1, 2$ we can bound the sum of the squares of the L^2 norms on $S_{\underline{u},u}$ of up to the mth intrinsic to $S_{\underline{u},u}$ covariant derivatives of these tensorfields in terms of the sum of the squares of the L^2 norms on the $S_{\underline{u},u}$ of their Lie derivatives of up to mth order with respect to the set of rotation fields. This is important because only these rotational Lie derivatives of the curvature components (and the Lie derivatives of the curvature components with respect to L and $\underline{L}$), not their covariant derivatives intrinsic to the $S_{\underline{u},u}$, are directly controlled by the energies and fluxes. To establish the coercivity inequalities, additional bootstrap assumptions are introduced.

Chapter 12 is the central chapter of the monograph. This chapter lays out the first method and defines the energies and fluxes according to the short pulse method as discussed above. These definitions are followed by the comparison lemmas, Lemmas 12.5 and 12.6 which show that the quantity $\mathcal{Q}_2'$, which bounds all the curvature quantities that enter the estimates for the connection coefficients and the deformation tensors of Chapters 3–9, is itself bounded in terms of the quantity

$$\mathcal{P}_2 = \max\{\overset{(0)}{\mathcal{E}}_2, \overset{(1)}{\mathcal{E}}_2, \overset{(2)}{\mathcal{E}}_2, \overset{(3)}{\mathcal{E}}_2; \overset{(3)}{\mathcal{F}}_2\}. \tag{41}$$

To establish the comparison lemmas, additional bootstrap assumptions are introduced. The last section of Chapter 12 gives the statement of Theorem 12.1 in the way it is actually proved, and then gives an outline of the first and most important part of the continuity argument, that which concludes with the derivation of the reductive system of inequalities for the quantities $\overset{(0)}{\mathcal{E}}_2, \overset{(1)}{\mathcal{E}}_2, \overset{(2)}{\mathcal{E}}_2, \overset{(3)}{\mathcal{E}}_2$ and $\overset{(3)}{\mathcal{F}}_2$ in the first section of Chapter 16.

Chapters 13–15 deal with the error estimates, namely the estimates for the *error integrals*, the spacetime integrals of the absolute values of the divergences of the energy-momentum density vectorfields $\overset{(n)}{P}_2$. There are two kinds of error integrals: the error inte-

grals arising from the deformation tensors of the multiplier fields and those arising from the Weyl currents generated by the commutation fields. The first are treated in Chapter 13 and the second in Chapters 14 and 15. Because of the delicacy of the final estimates, all the error terms are treated in a systematic fashion. All error integrals are estimated using Lemma 13.1. The concepts of *integrability index* and *excess index* are introduced. The integrability index s being negative allows us to apply Lemma 13.1. The excess index e then gives the exponent of δ contributed by the error term under consideration to the final system of inequalities for the quantities $\overset{(0)}{\mathcal{E}}_2$, $\overset{(1)}{\mathcal{E}}_2$, $\overset{(2)}{\mathcal{E}}_2$, $\overset{(3)}{\mathcal{E}}_2$ and $\overset{(3)}{\mathcal{F}}_2$. All error terms turn out to have a negative integrability index and a non-negative excess index. The terms with a positive excess index contribute the innocuous terms $O(\delta^e)$ mentioned above. To the terms with zero excess index are associated *borderline error integrals* . These contribute the *nonlinear terms* to the final system of inequalities mentioned above.

Chapter 16 completes the proof of the existence theorem. The reductive system of inequalities for the quantities $\overset{(0)}{\mathcal{E}}_2$, $\overset{(1)}{\mathcal{E}}_2$, $\overset{(2)}{\mathcal{E}}_2$, $\overset{(3)}{\mathcal{E}}_2$ and $\overset{(3)}{\mathcal{F}}_2$ is obtained in the first section of Chapter 16, and the required bounds for these quantities are deduced. The second section of Chapter 16 deduces the higher-order bounds for the spacetime curvature components and the connection coefficients. The higher-order estimates are of linear nature and are needed to show that the solution extends as a smooth solution. Only the roughest bounds are needed. The continuity argument is completed in the third section of Chapter 16. In this section the work of Choquet-Bruhat ([Cho1], [Cho2], [Cho3]) and that of Rendall [R] are used to obtain a smooth local extension of the solution in "harmonic" (also called "wave") coordinates . This is followed by an argument showing that, in a suitably small extension contained in the extension just mentioned, canonical null coordinates can be set up and the coordinate transformation from harmonic coordinates to canonical null coordinates is a smooth transformation with a smooth inverse, hence the metric extends smoothly also in canonical null coordinates. The proof of the existence theorem is then concluded. In the last section, the existence theorem is restated in the way in which it can most readily be applied.

The last chapter, Chapter 17, establishes the theorem on the formation of closed trapped surfaces, achieving the aim of this monograph.

The present monograph is of course a work in mathematics. However, by virtue of the fact that Einstein's theory is a physical theory, describing a fundamental aspect of nature, this work is also of physical significance. For those mathematicians who, by reading the present monograph, become interested in the physical basis of general relativity, we recommend the excellent book [M-T-W] where not only is the physical basis of the theory explained, but also a wealth of information is given which illustrates how the theory is applied to describe natural phenomena.

As this work was being completed the author learned that his old teacher, John Wheeler, passed away. This monograph testifies to John Wheeler's enduring legacy in the scientific community.