Towards exascale supernova simulations with GenASiS

Christian Y. Cardall, Eirik Endeve, Reuben D. Budiardja, Pedro Marronetti, and Anthony Mezzacappa

Abstract. Core-collapse supernovae are worth simulating on leadership computing resources because they are the origin of many elements, are interesting targets for expensive observational programs, and provide valuable computational experience. The need to perform implicit neutrino transport addressing up to six dimensions of phase space makes simulations aimed at the explosion mechanism an exascale problem. The multiscale nature of the problem in space and time can be addressed with adaptive mesh refinement and implicit evolution. Several types of solvers—explicit and implicit, hyperbolic and possibly elliptic—must be deployed with operator splitting to address the multiphysics nature of the problem. While we have some ideas about how parallelism can be exploited, we look to applied mathematicians and computer scientists to work with us and provide libraries that deploy capabilities needed by our solvers on heterogeneous and communication-unfriendly computer architectures.

1. Why Simulate Supernovae?

Among all the applications of pressing practical importance facing the nation and the world, why should valuable leadership computing resources be spent on the simulation of supernovae? Practical questions are not the only ones that matter to us. Perhaps even more important—in the sense of having limited utility to anything else, and therefore having an intrinsic worth that answers to nothing ‘higher’—are answers to questions relating to our origins and place in the universe.

How the building blocks of our bodies and our environment, the chemical elements, came to be formed and spread through the universe is one such question. As the saying goes, we are stardust: many important elements, or nuclear species, are synthesized by the nuclear fusion that powers stars. But it is not enough to synthesize the elements; these must also be dispersed if they are to be incorporated into planets and people. For stars with masses between \( \sim 0.8 \, M_\odot \) (where \( M_\odot \) is the mass of the Sun) and \( \sim 8 \, M_\odot \), this dispersal is with a whimper as they puff their chemically-enriched envelopes into the interstellar medium. This can result in very beautiful structures, known historically, if semantically inaccurately, as planetary nebulae. In other cases the element dispersal can occur with a bang—one that includes further element formation (or ‘nucleosynthesis’). Such an explosion is called a ‘supernova,’ which is shorthand for an ‘exceptionally bright new star.’ The peak optical luminosity of a supernova is comparable to that of its host galaxy containing billions of stars. SN 1987A is the most recent nearby supernova; it went off in the Large Magellanic Cloud, a small satellite galaxy to the Milky Way.

There are two basic physical scenarios that lead to supernovae. A thermonuclear supernova occurs when a white dwarf (remnant of a star with mass \( M \lesssim 8 \, M_\odot \)) accretes too much mass from a binary
companion. Three thermonuclear supernovae have occurred in our Galaxy in historical times. A core-collapse supernova occurs when the core of a star with $M \gtrsim 8 M_\odot$ collapses at the end of several burning stages, releasing gravitational energy. There have also been three core-collapse supernovae in the Milky Way in historical times. Modern surveys confirm what the small-number statistics of our own recent Galactic history suggest: thermonuclear and core-collapse supernovae occur with comparable frequency.

On a slightly more practical level, but still within the realm of purely scientific interest, core-collapse supernovae are interesting targets for observational programs in which large public investments have been made. These include telescopes—both earthbound and spaceborne—that cover various ranges of the electromagnetic spectrum. Astronomical observations investigate such aspects as the launch of an explosion; neutron star mass, spin, magnetic field, and kick velocity; composition of the ejecta; and explosion morphology. Even more exciting would be the detection of nontraditional astronomical signals from a supernova in our galaxy. Neutrinos are ghostly particles that interact only very weakly with ordinary matter; they are emitted in a certain type of nuclear decay, and from hot and dense nuclear matter, including in copious amounts from core-collapse supernovae. Gravitational waves are tiny ripples in spacetime radiated from violent events in which high-density matter moves at close to the speed of light. Both of these types of signals are so weak that it takes truly heroic experimental efforts to detect them. But that same weakness of interaction means that they escape their sources comparatively easily, so that, unlike photons, they bring us observational information directly from the heart of the supernova.

On an even more practical level, accomplishing core-collapse supernova simulations enlarges community experience in addressing several computational challenges of interest in other applied contexts. These include magnetohydrodynamics; reaction kinetics; particle transport; and treating all these in a multidimensional, multiscale, and multiphysics system.

2. Why is the Core-Collapse Supernova Mechanism an Exascale Problem?
Stars spend most of their lives burning hydrogen to helium in their cores. Stars with masses greater than $8 M_\odot$ continue to burn up to oxygen, neon, and magnesium, and those heavier than $10 M_\odot$ burn all the way up to iron group elements near the top of the nuclear binding energy curve. In the absence of core burning, electron degeneracy becomes the main source of pressure support. But once the electrons become relativistic they have nothing more to give. This is the basic physics behind the Chandrasekhar limit, and when the mass of the core exceeds it, dynamical collapse ensues.

The collapsing interior divides into an inner core, in sonic contact with itself, and an outer core whose infall is supersonic. Collapse of the inner core halts when the nuclear equation of state stiffens above nuclear density. But the outer core has not heard the news that collapse has halted. A shock wave forms when supersonically falling material slams into the inner core. The shock moves out, heating the material through which it passes, and eventually will give rise to the optical emission we know as a supernova once it travels thousands of kilometers to the optically thin outermost stellar layers.

This does not happen right away, however. At around 150 or 200 km the shock stalls, due to a loss of pressure support, and becomes a ‘stationary accretion shock.’ This happens because of two enervating consequences of the high temperature of the shock-heated material: iron-group nuclei that fall through the shock are endothermically reduced to their constituent nucleons, and electron capture results in neutrino emission. The mechanism of shock revival—that is to say, the explosion mechanism—remains a subject of active investigation. But since the 1980s the delayed neutrino-driven explosion mechanism has been a primary paradigm: beyond the so-called ‘gain radius,’ heating by neutrino absorption outweighs cooling by electron capture, and on longer timescales (hundreds of milliseconds) may re-energize the shock [1].

Determining the heating and cooling rates that affect the fate of the shock requires neutrino transport. Deep inside the newly-born neutron star, where neutrinos are trapped and slowly diffuse outwards, their distribution is nearly isotropic. As they begin to decouple in the semitransparent
regime between the proto neutron star and the shock their angular distribution becomes more and more strongly forward-peaked. This transition happens differently for different neutrino energies. Ultimately, therefore, knowledge of the neutrino heating and cooling rates relies on knowledge of the neutrino distribution functions: at every instant in time and at every point in space, we would like to know how many neutrinos there are with a given energy moving in a given direction. This requires implicit solution of the Boltzmann equation or something equivalent.

All three spatial dimensions are important in core-collapse supernovae. Convection and related phenomena may occur in the proto neutron star, and there is convection in the region between the gain radius and the shock. Rotation and magnetic fields may also come into play. Aside from convection, rotation, and magnetic fields, it has been discovered in recent years that the stationary accretion shock is unstable. This stationary accretion shock instability, or SASI, may be responsible for phenomena previously attributed to rapid rotation of the progenitor star, including aspherical explosion morphology, pulsar spin, and magnetic field generation [2,3,4].

The high dimensionality of phase space—which, again, in its full glory consists of three spatial dimensions plus three momentum space dimensions—is, together with the need to perform implicit neutrino transport, most of what makes this an exascale problem. (Evolving a nuclear reaction network with many dozens of species at every spatial cell would also be overwhelming.) As an example, consider a very simpleminded estimate of how easily the inversion of dense blocks arising from momentum space coupling can exhaust exascale resources. Each dense block has \( N_p \) rows and columns, where \( N_p = N_{\nu}N_{E}N_{\theta}N_{\phi} \) is the number of momentum space bins arising from \( N_{\nu} \) neutrino species, \( N_{E} \) energy bins, and \( N_{\theta}N_{\phi} \) angle bins. Suppose the structure of the dense blocks can be exploited in such a way that their inversion can be solved in order \( N_p^3 \) operations rather than the \( N_p^2 \) operations that would be required by LU decomposition. There is one dense block for each of \( N_x \) spatial cells. Suppose this overall solve requires \( N_{it} \) iterations for each of \( N_t \) time steps. Then the total number of required flops goes like

\[
N_{\text{FLOP}} \sim N_{it} N_{it} N_{x} N_{p}^2.
\]

The wall time is

\[
T_{\text{wall}} = \frac{N_{\text{FLOP}}}{\epsilon_{\text{FLOP}} R_{\text{FLOP}}},
\]

where the efficiency \( \epsilon_{\text{FLOP}} \) is the achieved fraction of the theoretical maximum FLOP rate \( R_{\text{FLOP}} \). Plugging possible numbers, we arrive at the figure of merit

\[
T_{\text{wall}} \sim 7 \text{ weeks} \left( \frac{N_{t}}{10^6} \right) \left( \frac{N_{it}}{20} \right) \left( \frac{N_{x}}{10^6} \right) \left( \frac{N_{p}}{10^6} \right)^2 \left( \frac{R_{FLOP}}{10^{18} \text{s}^{-1}} \right)^{-1} \left( \frac{\epsilon_{\text{FLOP}}}{0.05} \right)^{-1}.
\]

We see that exascale resources can be exhausted for several weeks for a single simulation.

Management and analysis of supernova simulation output is expected represent its own set of challenges on exascale machines. Long time integrations of large-scale, high-resolution simulations describing physics on wide range of spatial and temporal scales result in massive amounts of data being written to disk at frequent time intervals for analysis and restart. (Our current suite of high-resolution MHD simulations has produced hundreds of terabytes of simulation data, and is already posing challenges for post-simulation storage and data mining.) The additional physics required for a credible supernova model (i.e., relativistic gravity, neutrino momentum space, and nuclear kinetics) is expected to dramatically increase simulation data output. Our traditional (linear) operating mode of simulation execution followed by data analysis will likely have to be revised. Much of the expensive interaction between memory and disk may be reduced by carrying out basic visualization and analysis while simulation data is in memory, but initial simulations will explore new territory in supernova theory, and significant amounts of data is expected to be written and retained for post processing as well.
Given these overwhelming requirements, ways have been found by several groups (see for instance [5,6,7,8,9,10] and references therein) to perform more tractable computations by reducing the dimensionality of phase space in various ways. Given the different possible ways of reducing the dimensionality of both space and momentum space, not to mention different physics implementations and numerical methods, there has not been complete convergence of simulation outcomes over time.

The Oak Ridge/Florida Atlantic/North Carolina State collaboration, with which we are associated, has two major projects. One of them, CHIMERA [5], has recently seen explosions in axisymmetry with energy-dependent neutrino transport. A corresponding three-dimensional full multiphysics run with this code has also been started. The other major project of our collaboration, GenASiS, is ultimately aimed at addressing the full dimensionality of the neutrino transport problem, with reasonable intermediate steps.

3. How Can the Multiscale Nature of the Problem be Addressed?

Several things make a core-collapse supernova a multiscale problem. Perhaps the first and most obvious is that gravitational collapse changes the length scale by roughly a factor of 100. There are also critical surfaces that should be well resolved: the large density gradient at the proto neutron star surface; the shock; and interfaces between regions of different compositions.

Then there is the matter of turbulent cascades to small scales, which can make a difference. Small-scale turbulence may increase the dwell time of fluid elements in the ‘gain region,’ increasing their heating by neutrino emission. Moreover, the outcome of simulated magnetic field generation depends strongly on resolution. In the case of the magneto-rotational instability this is because the linear growth rate of unstable modes is inversely proportional to their wavelength. Another amplification mechanism operates in simulations we have performed with GenASiS [4]: turbulence tangles and stretches magnetic flux tubes until their thickness is comparable to the spatial resolution. Figure 1 shows slices through two 3D MHD simulations that differ by a factor of two in resolution; the difference in the amount of yellow visually betrays the difference in magnetic field strength. The dependence of average magnetic field strength on resolution is illustrated more quantitatively in Figure 2, which shows the behavior expected of this simple geometric effect in 2D and 3D. These simulations are still unresolved even at $1024^3$ on over 32,768 cores. (As an aside, we note that GenASiS magnetohydrodynamics scaling has now been pushed to $1280^3$ on 64,000 cores, as illustrated in Figure 3.)

![Figure 1. Magnetic field magnitude at $t = 1500$ ms for models with cell width $\Delta l \approx 2.34$ km (left panel) and $\Delta l \approx 1.17$ km (right panel). The orientation of the plots is chosen so that the normal vector of the slicing plane is parallel to the total angular momentum vector of the flow between the PNS and the shock surface.](image-url)
Figure 2. Time-averaged rms magnetic field strength \( (B_{\text{rms}}) \), versus time-averaged magnetic rms scale \( (\lambda_{\text{rms}}) \), in the ‘saturated’ nonlinear stage. Results are shown (in blue) for the axisymmetric 2D models and (in black) for the 3D models. (The time-average extends from \( t = 500 \) ms to the end of each model.) The dotted blue and dashed black reference lines are proportional to \( (\lambda_{\text{rms}}) \) and \( (\lambda_{\text{rms}})^2 \), respectively.

Figure 3. Weak scaling behavior of GenASiS magnetohydrodynamics.
The existence of disparate spatial scales, coupled with the high cost per spatial cell, motivates more efficient methods of crossing spatial scales than simply increasing unigrid resolution. Adaptive mesh refinement (AMR) is one primary means. One possibility is block-structured AMR [11]. In the context of explicit evolution this has the virtue of being able to deploy existing solvers on individual regular cell blocks. There are also efficiencies associated with the use of a predictable basic structure. Another possibility is cell-by-cell AMR with a fully-threaded tree [12]. This offers more fine-grained control over cell division and placement, which is a virtue when the cost per spatial cell is very high. It is not conceptualized around local explicit solvers, so that it might be more amenable to elliptic and implicit solvers which often are needed in a multiphysics context. We are interested in cell-by-cell AMR for GenASiS. Beyond AMR, we also would be interested in subgrid models of turbulence, and in particular turbulent MHD.

Not being aware of a publicly available cell-by-cell refinement library, we undertook to develop this capability on our own. We have found it heavy going for a small application science team, but have made some progress. Figures 4 and 5 illustrate some of our initial efforts at cell-by-cell AMR with GenASiS.

Figure 4. The mesh and its partitioning based on density gradient for a uniform-density sphere; 3D view (upper left) and a 2D slice (upper right). Also shown are the gravitational potential as computed with our multigrid Poisson solver (lower left) and its error relative to the analytic solution (lower right).
Figure 5. Two snapshots from a 2D shock tube problem illustrate the load rebalancing of our cell-by-cell AMR in a hydrodynamics context. Both the mesh and its partitioning (left panels) and density (right panels) are shown.

We have been discussing spatial scales, but it must also be pointed out that the core-collapse supernova phenomenon is multiscale in time as well. This is handled by implicit evolution of neutrino transport and nuclear reaction kinetics.

4. How Can the Multiphysics Nature of the Problem be Addressed?
From the foregoing sketch of the supernova phenomenon we can see that ideally we would like simulations to cover a wide range of physics, which can be grouped under three main headings: the tangent bundle, the magnetofluid, and the neutrino distributions. The term ‘tangent bundle’ is a fancy way of saying ‘spacetime plus momentum space.’ Ideally we would like our representation of spacetime to include all three space dimensions, with good resolution on a wide range of length and time scales; we would like momentum space to include all three dimensions, with good resolution of energies and angles; and ideally self-gravity should be treated with general relativity. The treatment of magnetohydrodynamics must be able to handle shocks. It would be desirable to track nuclear composition using a wide range of species. An equation of state for nuclear matter at finite temperature in neutron-rich conditions is needed. Neutrino transport must be computed in diffusive, decoupling, and free-streaming regimes. Neutrino interactions with all fluid components, and with other neutrinos and antineutrinos, must be included. Neutrino flavor mixing should be included.
Not surprisingly, these different physics pieces involve different types of equations, and correspondingly different kinds of solvers.

In the case of gravity there are some options. If one stays with Newtonian gravity, which is not completely realistic, there is a single linear elliptic equation for the gravitational potential. One can stay with approximate relativity, which captures the enhanced nonlinear strength of relativistic gravity, by solving a set of nonlinear elliptic equations for a reduced set of metric components describing the curvature of spacetime [13]. The in-situ calculation of gravitational waves, however, requires full relativity; the most common formulation these days is as a set of nonlinear first-order hyperbolic equations with continuous solutions satisfying elliptic constraints [14,15]. If elliptic solvers are deployed, multigrid-style approaches seem to be an attractive option. For GenASiS we have worked on one such solver that uses a distributed FFT for the coarsest level and global sparse solves on the finer levels, as illustrated in Figure 4. If only the hyperbolic equations of full relativity are solved a number of methods can be used. High order solvers have been common because of the absence of discontinuities in the gravity variables and the presence of numerical instabilities, but our initial work on full general relativity in GenASiS has been only second order due to the expected availability of only a single layer of ghost cells in our cell-by-cell AMR scheme.

A few different things are required by the magnetohydrodynamics. The main thing is a set of first-order hyperbolic conservation laws—or technically, balance equations, since we have source terms from interactions of the fluid with gravity and neutrinos. The solenoidal constraint on the magnetic field must be respected, in addition to an algebraic constraint known as the equation of state (pressure as a function of density, temperature, and composition). Determination of the equation of state for finite temperature nuclear matter is a significant computational problem in and of itself. A reaction network for time evolution of nuclear abundances is also needed. Many methods have been developed to solve conservation laws. In GenASiS our initial choice is an explicit second-order finite-volume ‘central scheme’, combined with the ‘constrained transport’ technique to automatically enforce the solenoidal constraint in the evolution of the magnetic field [16,17]. We have worked with tabulated equations of state for nuclear matter. Local nuclear reaction networks (one for each spatial cell), which must be implicit, have been deployed by our colleagues in CHIMERA but have not yet been put into GenASiS.

For neutrino transport we have first-order hyperbolic integro-differential equations. In conservative form these feature a spacetime divergence, a momentum space divergence, and source terms with integrals over momentum space. Because of the wide range of time scales addressed by transport solvers that range from near-equilibrium diffusive to semi-transparent to nearly free streaming regimes, implicit evolution is a necessity. Past work in our group has used Newton-Raphson to address the nonlinear aspect. Inside each Newton-Raphson iteration is a linear solve. As alluded to previously, this involves dense blocks from momentum space couplings (arising from the integrals over momentum space). It also involves nearest-neighbor spatial couplings, which can be conceptualized as separate sparse matrices, one for every momentum space bin. Our initial strategy in GenASiS is to obtain a large amount of parallelism by separately inverting the dense momentum space blocks and sparse spatial matrices, and iterating these separate solves to fixed-point convergence.

5. How Can the Expected Features of an Exascale Machine be Utilized?
Some expected features of exascale machines that have penetrated the consciousness of application scientists like ourselves include distributed memory (at least partially), multicore nodes, and heterogeneous processors (use of GPUs and so forth). The first two have been somewhat adapted to using MPI and OpenMP, but our use of these may have to change significantly as communication becomes increasingly expensive. The use of hardware accelerators such as GPUs is something of a brave new world. Nevertheless our colleagues working on CHIMERA have begun experimenting with harnessing GPUs in their implicit solves of nuclear reaction networks.

At some level, awareness of and responsibility for these issues has to percolate all the way up to the application scientists. However we have to rely heavily on, and work together with, applied
mathematicians and computer scientists to help us understand what might be possible and to build libraries that can handle many of these issues in often-needed solvers and kernels. Can what is needed be done at the exascale? Here is a list, with heavy emphasis on expected issues related to data non-locality, of some needs of the several different physics solvers described above:

- Domain decomposition, level-by-level; empirical determination of workload per cell for load balancing? (sparse point-to-point communication for prolongations and restrictions)
- Nearest-neighbor data for explicit hyperbolic equation updates (sparse point-to-point communication overlapping with work)
- Distributed FFT on 3D spatial domain (either parallel FFT, or transposes within slabs)
- Distributed sparse solves over spatial domain (parallel solver libraries)
- Fast inversion of dense blocks (local; use GPUs)
- Fast local sparse spatial solves (local after transpose from space decomposition to momentum space decomposition; use GPUs)
- Residuals, time step determination, global sums (reductions)
- Synchronized evolution of the entire domain in time

On this last point, we are aware that ‘synchronization’ is a bad word. If one were doing only explicit solves it would be possible to contemplate allowance for asynchrony outside causality cones, but with elliptic and especially implicit solves we have difficulty imagining any way around stepping the entire system forward in time. As far as we can tell, any ‘out of order execution’ will have to be confined to the solves within a single time step.

As is already known from current experience (but can only be expected to get much worse), collectives—transposes and reductions, in terms of the items on the foregoing list—are the biggest challenge in terms of communication. It may be that non-blocking collectives could open up important new possibilities for overlapping work and communication. To take a very simple example, time step determination could be lagged one or two or a few steps, in order for evolution to continue a bit while a reduction proceeds in the background. Another example particular to radiation transport treating the full momentum space: exposure of an important opportunity for parallelism, namely separate local inversions of a sparse spatial coupling matrix for each momentum space bin, depends on transposing spatial decomposition into momentum space decompositions. Given non-blocking collectives, such a transpose (via All-to-Alls) could occur in the background while the naturally local dense block inversions are performed.

Acknowledgements
We acknowledge computational resources provided through the INCITE program at the Leadership Computing Facility at the National Center for Computational Sciences at Oak Ridge National Laboratory. C. Y. C. and A. M. acknowledge support from the Office of Nuclear Physics, DOE. C. Y. C., E. E., and A. M. acknowledge support from the Office of Advanced Scientific Computing Research, DOE. Oak Ridge National Laboratory is managed by UT-Battelle for the DOE. E. E., R. B., and P. M. acknowledge support from NSF-OCI-0749204. P. M. and A. M. acknowledge support from NASA 07-ATFP07-0011.

References


