*Special Issue Paper*

International Journal of HIGH PERFORMANCE COMPUTING APPLICATIONS

*The International Journal of High Performance Computing Applications*
2018, Vol. 32(4) 582–593

DOI: 10.1177/1094342016668241
journals.sagepub.com/home/hpc
SAGE

# MiniApps derived from production HPC applications using multiple programing models

**OE Bronson Messer, Ed D'Azevedo, Judy Hill, Wayne Joubert, Mark Berrill and Christopher Zimmer**

**Abstract**

We have developed a set of reduced, proxy applications ("MiniApps") based on large-scale application codes supported at the Oak Ridge Leadership Computing Facility (OLCF). The MiniApps are designed to encapsulate the details of the most important (i.e. the most time-consuming and/or unique) facets of the applications that run in production mode on the OLCF. In each case, we have produced or plan to produce individual versions of the MiniApps using different specific programing models (e.g., OpenACC, CUDA, OpenMP). We describe some of our initial observations regarding these different implementations along with estimates of how closely the MiniApps track the actual performance characteristics (in particular, the overall scalability) of the large-scale applications from which they are derived.

## 1. Introduction

There is a fundamental tension between the size and complexity of modern, large-scale scientific application codes and the design and fielding of new computational platforms on which these codes are run. The ultimate utility of current petascale and future exascale platforms will be judged based on the scientific productivity of their users. This means that modifying and improving the algorithms and implementations used in these large-scale codes to increase efficiency on these new platforms is of primary importance. It is equally clear that a thorough understanding of the requirements and limitations of current and planned software features is absolutely essential in system design. However, such optimization work is often arduous because of the sheer size of the code bases, the complexity of the codes themselves, and, in many cases, their associated build systems. This often means that accessible benchmarks are not available at any given instant in the software and hardware development processes—for vendors to optimize system design points against, for computer science and applied mathematics researchers to use as laboratories to test new algorithmic and implementation ideas, and even for application developers themselves to use as effective test mechanisms to investigate the impacts of new hardware and programing model developments. For these reasons and others, the use of reduced applications that share many of the performance and implementation features of large, fully-featured code bases ("MiniApps") has gained considerable traction in recent years, especially in the context of exascale planning exercises. The utility of MiniApps relies on the realization that many large scientific codes contain a small, countable number of "hot spots" that dominate their performance characteristics, and that other parts of these large codes contain recurring tropes, where, though the ultimate aim might be different for each routine, the performance characteristics are quite similar Heroux et al. (2009). The idea is to encapsulate these behaviors in a simpler, reduced version of the application, mimicking the important characteristics, but obviating the need for the user to be a developer/expert on the code itself. We have undertaken a project to create, analyze, and disseminate a collection of MiniApps, particular to the user community of the OLCF, that can be changed, revised, or even retired as architectures and software systems evolve. In this paper, we introduce the initial set of these OLCF MiniApps and analyze their performance, relating back to the performance of the progenitor application code wherever possible.

Oak Ridge National Laboratory, Oak Ridge, TN, USA

**Corresponding author:**
OE Bronson Messer, Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN Tennessee 37831, USA.
Email: bronson@ornl.gov

## 2. Background

An ever-expanding group of MiniApps has been developed in recent years. The term "MiniApps" is a fairly recent usage, and benchmark codes of much the same type had been developed long before this term was first used. As long ago as the turn of this century, the ASCI Purple benchmark suite Futral et al. () represented a remarkably similar set of condensed, easily analyzed codes. This work drew on experiences with even earlier codebases like the Livermore Loops collection (McMahon, 1986) (a collection of benchmarks still useful in its modern form as LCALS (Hornung and Keasler, 2013) and, of course, the well-known NAS benchmarks (Bailey et al., 1991). Other examples exist, perhaps with more limited use. Ultimately, the Mantevo suite (Heroux et al., 2009) was perhaps the first and largest set of benchmark codes described by the term MiniApps. Mantevo includes MiniApps designed to mimic the performance of finite-element codes (MiniFE), molecular dynamics codes (MiniMD), electrical circuit design codes (MiniXYCE), and others. It is important to note that the particular choice of MiniApps in the Mantevo suite is a direct result of the interests and expertise of the application community at Sandia National Laboratories, where the suite was developed. Other MiniApps have been produced at Los Alamos National Laboratory, e.g. CGPOP (Stone et al., 2011) and MCMini (Marcus, 2012), and at Argonne National Laboratory (see, e.g. the MiniApps list at the CESAR website, https://cesar.mcs.anl.gov/content/software).

It is interesting to note that MiniApps are often described via exclusion, i.e. by enumerating the things that are not MiniApps. Barrett et. al (2011) characterize MiniApps by noting that they are not to be confused with compact applications—wherein a particular implementation of physics is captured in isolation—nor are they so-called skeleton applications, designed to reproduce a particular pattern of inter-process communication (a goal so narrow that the computation performed between communication epochs is sometimes "faked"). Heroux et al. (2011) expand this list of what MiniApps are not to include Scalable Synthetic Compact Applications (SSCA), which were produced through the DARPA High Productivity Computing Systems (HPCS) program to evaluate the productivity of emerging HPC systems. The NNSA Exascale Applications Working Group noted (Springmeyer, 2011) that despite these attempts to constrain the definition of MiniApps, their role in cooperative R&D efforts (e.g. in so-called co-design efforts, where hardware and software designers engage directly with application programers to improve the design of both architectures and algorithms for future systems) requires them to be several things at once. They must be expansive enough to accurately model full application behavior, but small enough to be manageable and, in some sense, parseable, by researchers ranging from applied mathematicians to compiler writers. We consider a primary aim of the research started under this project to be an attempt to delineate what particular attributes maximize the usefulness of MiniApps for these multifarious purposes.

Related to this primary aim is the question of whether MiniApps can be effectively divorced from the underlying physics being modeled by a particular application code (see, e.g., the discussion of this point by Barrett et al., 2012). We posit that this notion is not a useful operating assumption based on our extensive past experience with transitioning large application codes across many orders of magnitude in scalability and performance. Indeed, we note that the realization that HPCCG (one of the original components of the Mantevo MiniApp suite (Heroux et al., 2009) could only, "provide a stronger tie to applications of interest," by realizing that, "the context in which the linear system is formed needed [to be] (sic) strengthened" (Barrett et al., 2012) suggests a stronger tie between performance and "intent" is often necessary to effectively understand full application performance. Nevertheless, the *additional* application of objective techniques to measure the degree to which a MiniApp models the progenitor application are likely useful as well. An example of this type of analysis is provided by the Byfl tool of Pakin and McCormick (2013).

We test several of these assumptions and ideas via our suite of MiniApps by comparing performance measurements made by varying MiniApp inputs and particular code implementations. We also attempt to understand the level of fidelity to the underlying physics required in order to accurately model full-application performance. We begin with a short discussion on formulating MiniApps from existing codes. In the following sections, we provide short descriptions of the first set of OLCF MiniApps, along with a set of initial scalability and performance results. Though many possible measures of performance are possible, we concentrate here on comparing the distributed-memory scalability of the MiniApps to their progenitor apps as a first measure of congruence. Such a comparison is probably a necessary but not sufficient condition to establish the quality of the MiniApps in reproducing the performance characteristics of the production applications. An exception to this general principle is our discussion of the the last MiniApp presented here, the XrayTrace MiniApp. This MiniApp is node-level only, obviating the possibility of a scalability measurement. However, the relative simplicity of this MiniApp also enables a large number of different programing model instantiations, enabling another kind of comparative analysis.

## 3. MiniApps generation and evaluation

The formulation of a new MiniApp can likely be described by a rather generic series of steps, but is dependent on the purpose and provenance of the MiniApp. We are concerned here with MiniApps that are directly derived from existing codes, as opposed to MiniApps that are generated *ab initio* to mimic a whole class of codes or algorithms. Therefore, some of the steps we outline below may not be applicable in every circumstance.

The first step we suggest is a complete performance profile of the progenitor code. Of primary importance in this analysis is determining which parts of the original code are responsible for determining the performance characteristics of the runtime. The most common set of questions in this analysis includes the following.

- Which code routines/functions perform the bulk of the floating-point operations (FLOPs)?
- Which code routines/functions perform the bulk of the memory operations?
- What are the relative computational intesities of these FLOP-rich and/or memory BW-dependent routines?
- How many of these FLOP-rich routines exist? Is the profile "flat" or concentrated in only a handful of routines?
- Are the FLOP-rich routines tightly coupled to one another—either logically or temporally?
- Do these same routines also dominate the total runtime, or do other routines (e.g., I/O or distributed-memory communication routines) dominate the total time-to-solution?

It then becomes important to decide which performance behaviors will be modeled with the MiniApp, as it may prove difficult or not as useful to model several of them in the same code. A straightforward example of such a separation of concerns could be the formulation of a I/O skeleton app **and** a FLOP-centric MiniApp from the same production code. The I/O portion of the original code might be under constant development and could change quite drastically on short development time scales, while the underlying algorithmic implementations could be far slower to change.

This possibility leads us to consider another step in our MiniApp generation recipe: Given that the MiniApp developers are also most often among the progenitor application developers (Heroux et al., 2009 suggest that this is, in fact, the preferred state), it is advantageous for the creation of the MiniApp to become part of the standard development cycle of the production application. Automating the generation (and re-generation) of associated MiniApps is an excellent way to make sure that the performance questions asked of the MiniApp at any epoch provide an accurate picture of the production application's performance at that point. In the absence of an automated way to perform the generation, some effort needs to be expended for a manual regeneration whenever major rewrites occur in the production code.

Finally, the build system and run system for the resultant MiniApps must be kept as simple as possible. Because the MiniApps need to be immediately accessible to a varied population of potential users, almost anything beyond a simple `makefile` and the specification of a handful of command-line arguments should be eschewed. Importantly, if more than this level of complexity seems to be required, it is possible that the resulting MiniApp itself is too complex to be human-parseable, reducing its usefulness.

## 4. AORSA MiniApp

AORSA (All Orders Spectral Algorithm) models the response of the plasma radio frequency (RF) waves in a fusion tokamak device (Jaeger et al., 2002). The plasma state is described by a distribution function. For RF applications, the fast-wave time scale leads to effective approximation of the electric field, magnetic field, and distribution function as a time-averaged equilibrium part and a rapidly oscillating time harmonic part. The time harmonic terms satisfy the generalized Helmholtz equation, relating the frequency of wave, the plasma current induced by the wave fields, and the plasma conductivity kernel. Fourier modes are used as basis functions to represent the electric field. Collocation on an $M \times M$ rectangular grid is used to construct a complex dense linear system of size $\mathrm{N} = 3 \times M \times M$. A reduced linear system can be constructed by transforming the linear system (using the fast Fourier transform) into the real physical space and considering only collocation points within the plasma region—typically about 30% of grid points are in the vacuum region. Each processor constructs complete block rows of the matrix to facilitate a two-dimensional fast Fourier transform to real space. The variables in the vacuum region are eliminated, the reduced rows are transformed back to a Fourier representation, and the data reshuffled into a two-dimensional block cyclic distribution that is compatible with ScaLAPACK. Finally, the dense linear system is solved by LU factorization; performance analysis tells us this is the most computationally intensive part of the calculation. AORSA is written using Fortran90 and MPI and, by default, makes calls to ScaLAPACK `PZGETRF/PZGETRS` for the LU factorization and the solution of the large dense complex system of linear equations. Because the dominant computational kernel is the dense LU factorization, an appropriate MiniApp for AORSA is the LU factorization of a dense matrix. With the advent of accelerators and co-processors in

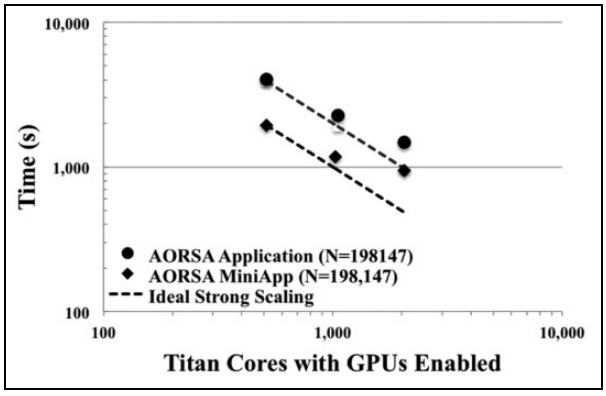


**Figure 1.** A comparison of the strong scaling for the full AORSA code and the AORSA MiniApp using Titan. The flop-heavy portions of the algorithm are computed on the GPU while the remainder of the algorithm uses the CPUs. The dashed lines indicate the ideal strong scaling behavior. The exhibited points are from the $32 \times 32$ processor grid runs in Table 1.

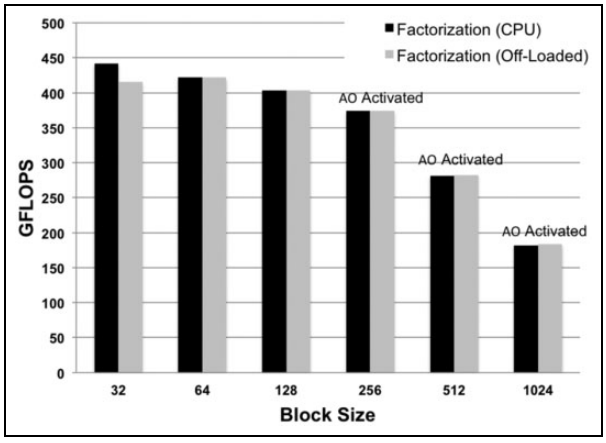


**Figure 2.** Comparison of the AORSA MiniApp performance when using just the CPU or when portions of the computation are automatically offloaded (AO), as determined by the system, to the Xeon Phi. Note that the automatic offloading is only activated at larger block sizes, as indicated by "AO Activated." For smaller block sizes, the system determines that AO is not advantageous, and the CPU performs all of the computations.

the HPC architecture spectrum and the relative limited memory available in each compared to traditional CPU architectures, additional care must be taken in memory management. Specific to AORSA, the amount of memory required to store the large, dense linear system far exceeds the available accelerator (GPU or MIC) memory, even on large distributed supercomputers such as Titan at the OLCF. Thus, out-of-core factorization methods, originally developed when CPU architectures were relatively memory-scarce (Azevedo, 2000; Brian et al., 2001) must be evaluated. In this work, we adapt the "left-looking" out-of-core algorithm for LU factorization described in Azevedo and Dongarra (2000) with a minor change that seeks to minimize the data transfer between the host and the device memory. Central to this algorithm is the necessity for an in-core parallel LU factorization method for a smaller sub matrix of the large linear system that operates primarily on the device with minimal communication between devices. For details of the out-of-core algorithm, see Azevedo (2000).

Figure 1 illustrates the effectiveness of the AORSA MiniApp in mimicking the scalability of the full AORSA application. Titan, the OLCF flagship supercomputer that employs a heterogeneous CPU/GPU architecture, was used for these simulations. Titan is a Cray XK7 composed of 200 cabinets. Titan is a hybrid architecture where each individual compute node uses both a conventional 16-core AMD 6274 Opteron CPU connected to 32 GB of 1600 MhZ DDR3 SDRAM and an NVIDIA K20x (Kepler) GPU with six GB of GDDR5 memory. Titan, with 18,688 of these hybrid compute nodes, has a theoretical peak computational performance of more than 27 PFLOPS. Each of the compute nodes is interconnected with Cray's high-performance, 3D-torus Gemini network. As expected, overall scalability of AORSA and its MiniApp are similar indicating that the LU factorization is the dominant computational kernel. Node-level performance results are presented in Table 1. There are still potential opportunities for performance optimization and tuning, such as individual tuning of matrix block sizes for ScaLAPACK on both CPU and GPU devices, exploiting asynchronous data transfer operations, and using look-ahead computation of the next panel to reduce the time spent on the critical path for LU factorization. However, the current gigaflop performance of the GPUs is approximately 20 times the performance of a single CPU core. Importantly, though this performance difference means that only just over half of the overall computational power on a node composed of a single GPU and a single multi-core CPU (typically with $\approx$16 cores) is delivered by the GPU, we expect the raw performance of near-future GPUs to eclipse CPUs quickly. Additionally, we expect node architectures to move towards multi-GPU configurations, also increasing the relative fraction of GPU performance per node. Therefore, pursuing efficient new algorithms implemented on GPUs offers a potential advantage in both cost and power efficiency.

Figure 2 illustrates our initial efforts to study the performance of the AORSA MiniApp on Intel Xeon Phi. These results were obtained using Beacon, a 48-node cluster operated by the National Institute of Computational Science (NICS) at the University of Tennessee. Each node is equipped with two eight-core Intel Xeon E5-2670 processors with 256 GB of memory and four Intel Xeon Phi coprocessors (5110P) with eight GB of memory. The results reported are for LU factorizations of matrix size N = 90,000 using a single Xeon Phi. While comparing the scalability to the full

**Table 1.** Strong scaling data from the AORSA MiniApp using the out-of-core GPU accelerated LU solver (compatible with ScaLAPACK `PZGETRF`.).

| Problem Size (N) | Nodes | Time (seconds) | GFLOPs per node | Total TFLOPs | Proc. Grid |
|---|---|---|---|---|---|
| 198,147 (257×257) 16 MPI tasks per node; | 36 | 3690 | 145.6 | 5.2 | 18 × 32 |
| OMP_NUM_THREADS=1 | 40 | 3379 | 143.1 | 5.7 | 20 × 32 |
| | 64 | 1992 | 151.7 | 9.7 | 32 × 32 |
| | 128 | 1324 | 114.2 | 14.6 | 64 × 64 |
| 198,147 (257×257) 8 MPI tasks per node; | 32 | 4049 | 149.3 | 4.8 | 16 × 16 |
| OMP_NUM_THREADS=2 | 66 | 2268 | 129.2 | 8.5 | 22 × 24 |
| | 128 | 1482 | 102.0 | 13.1 | 32 × 32 |
| 395,523 (513×257) 8 MPI tasks per node; | 98 | 10,946 | 143.4 | 14.1 | 28 × 28 |
| OMP_NUM_THREADS=2 | 128 | 8125 | 147.9 | 18.9 | 32 × 32 |
| | 253 | 4845 | 125.5 | 31.8 | 44 × 46 |
| 789,507 (513×513) 8 MPI tasks per node; | 378 | 24,157 | 134.1 | 50.7 | 54 × 56 |
| OMP_NUM_THREADS=2 | 561 | 17,240 | 126.5 | 71.0 | 66 × 68 |
| | 896 | 11,165 | 122.4 | 109.7 | 112 × 64 |

application is warranted, currently the small size of Beacon limits the realistic simulations that can be accomplished. However, in the single node case, automatic offloading—when the system determines that using the co-processor is advantageous—only activates for larger ScaLAPACK block sizes. Automatic offloading is currently available for a limited set of Intel Math Kernel Library (MKL) functions. It is, therefore, not surprising that the the automatic offloading activates only at larger block sizes when the flops are concentrated relative to communication and data transfer. For a larger block size, there is less opportunity for parallelism in each factorization. More idle time is inevitable, as at each step of the factorization, all processors must wait for the complete factorization of the block column panel before proceeding with the right-lookup update operations. The width of the block column panel is related to the block size (i.e. it is equal to the leading dimension of the block).

Significant performance optimizations remain before the Xeon Phi can become as performant as the current GPU implementation.

## 5. Ziz: A CHIMERA MiniApp

Next on our list of initial OLCF MiniApps is a proxy for the CHIMERA code that we have dubbed Ziz.[1] (Messer et al., 2008; Messer et al., 2013) is a multi-dimensional, multi-physics code designed to study core-collapse supernovae. The code is made up of three essentially independent parts: hydrodynamics, nuclear burning, and a neutrino transport solver combined within an operator-split approach. The hydrodynamics is directionally split, and the ray-by-ray transport and the thermonuclear kinetics solve occur after the radial sweep occurs, when all the necessary data for those modules is local to a processor. The directionally split hydrodynamics is implemented via MPI, while OpenMP is used to parallelize the local-to-each-MPI-rank work performed in the nuclear kinetics solver and the neutrino transport. All of the constituent parts of CHIMERA are written in FORTRAN 90. The combination of directionally split hydrodynamics and operator-split local physics provides the context for the communication and computation patterns found in CHIMERA. The neutrino transport in CHIMERA is performed only in the radial direction (as this step is the most computationally intensive, and would preclude realistic runtimes if treated in full generality). The result is a subcommunicator-local sparse linear solve at each timestep. The nuclear composition in regions that are not in nuclear statistical equilibrium is evolved via a completely local dense linear solve that is subcycled within each hydrodynamic timestep for every cell in the domain. Ziz will eventually be able to capture the behaviors of the directionally split hydrodynamics (i.e. the restricted data transposes), the operator-split local physics, and the dense and sparse linear solves used in the kinetics and transport solves. The resulting MiniApp should be useful as a proxy for a whole class of multi-physics codes. Individual modules for nuclear burning, sparse linear system solution, and hydrodynamics (specifically, flux reconstruction at finite-volume interfaces) will be produced to allow modular CPU-GPU MiniApps to be constructed. In an initial foray, we have used the directionally split hydro version of Ziz to quantify a handful of architectural impacts on Cray XK7 and XC30 platforms and have compared these impacts to results from a new Infiniband-based cluster at the OLCF.

### 5.1. Experiments with Ziz

We have performed scaling studies for Ziz on three different platforms at OLCF: Titan, Eos, and Rhea. The architecture of Titan was described previously in Section 4.

Eos is a 744-node Cray XC30 cluster. The system has two external login nodes. The compute nodes are organized in blades. Each blade contains four nodes connected to a single Aries NIC. Every node has 64 GB of DDR3 SDRAM and two Intel Xeon E5-2670 CPUs with eight physical cores each. Intels Hyper-threading (HT) technology allows each physical core to work as two logical cores so each node can function as if it has 32 cores. Each of the two logical cores can store a program state, but they share most of their execution resources. In total, the Eos compute partition contains 11,904 traditional processor cores (23,808 logical cores with Intel HT enabled), and 47.6 TB of memory.

Rhea is a 196-node commodity-type Linux cluster. Each of Rhea's nodes contain two eight-core 2.0 GHz Intel Xeon E5-2650 processors with HT and 64 GB of main memory. The nodes are connected via Mellanox 4X FDR Infiniband MT27500 network controllers.

Initial experiments with Ziz have centered on investigating the degree to which the scaling performance of CHIMERA is mimicked by the MiniApp and how that scaling behavior changes when the interconnect is changed from Gemini to Aries to Infiniband. A comparison of CHIMERA weak scaling to Ziz "hydro-only" weak scaling on Titan is shown in Figure 3. The "hydro-only" qualification refers to a version of Ziz wherein only the hydrodynamics module is included, i.e. there is no active module for radiation transport or for nuclear kinetics. Because the hydro phase in Ziz uses `MPI_AllToAll` on dimensionally split subcommunicators (due to the way transport and hydro are handled in CHIMERA), collective communication cost (even on subcommunicators) quickly degrades the scalability of this version of MiniApp. The weak scaling behavior of this version of Ziz becomes less efficient much more quickly than does the more physics-laden full CHIMERA runs. Previous profiling of CHIMERA has shown that the hydrodynamics calculation (including the requisite `MPI_AllToAll` communications) represents roughly 33% of the total runtime in CHIMERA for runs at less than O(30K) cores. For those same runs, roughly 45% of the runtime is dedicated to the transport solve, and ≈20% is spent in the nuclear burning module, the balance being spread among other routines, including I/O. The transport and nuclear burning solves also significantly increase the payload sizes for the transpose `MPI_AllToAlls`, increasing the number of double-precision REALs per zone from six to 180 (160 for the transport and 14 for the nuclear burning). However, in this version of Ziz, the `MPI_AllToAlls` transpose only the six variables at every grid point required by the hydrodynamics solve, rendering the total runtime for the MiniApp a small fraction of the runtime for the production version of CHIMERA (hence the multiplicaiton of the Ziz runtime by 10,000 in Figure 3) Because of all of these caveats, the scalability

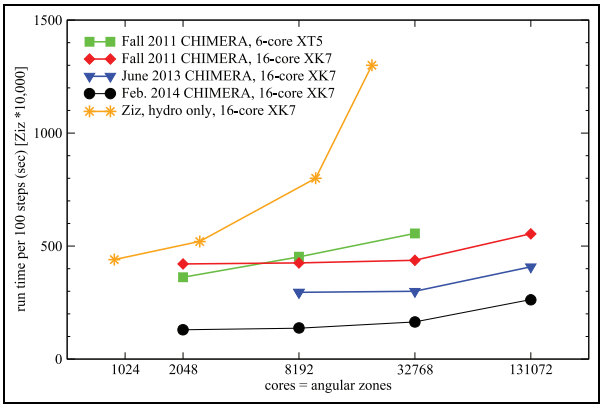


**Figure 3.** A comparison of hydro-only Ziz scaling and historical CHIMERA scaling experiments. The ordinate (i.e. time required for 100 time steps) is multiplied by 10,000 for the Ziz results: The time required to perform a hydro-only update is dwarfed by the transport and burning solves in the full CHIMERA code.

of this reduced Ziz version is in no way representative of CHIMERA. Nevertheless, it is important to understand the hydro-only scaling of the code, to better understand the turnover in the parallel efficiency that occurs above O(30K) cores.

Indeed, even this version of Ziz can be used to uncover architectural differences between machines. Comparing scaling results on different machines with the hydro-only version of Ziz can reveal specific information about relative node and network performance. Shown in Figure 4 are a subset of the same Ziz runs performed on Eos and on Rhea, along with the Titan scaling. The Eos and Rhea results are restricted in number because of the size of each of the machines and the particular modulo arithmetic that must be satisfied for each dimension in a Ziz (or CHIMERA) run. Perhaps unsurprisingly, Eos is (a) somewhat faster for a given number of MPI ranks and (b) weakly scales all the way out to 10,000 ranks with excellent efficiency. The runtime on Rhea is significantly longer than either of the Cray platforms, but the weak scaling is not quite as bad as for Titan at low rank counts. It should be noted that the Cray platform results were both obtained with code compiled with the CCE compiler suite (CCE 8.2.2 on Titan and CCE 8.1.9 on Eos), while the Rhea results came from code compiled with Intel IFORT (version 13.1.3). Neither the Eos nor the Rhea results allowed HyperThreading. Additional tests on Eos with Intel IFORT (version 13.1.3) produced results within 3% of the CCE results presented in Figures 3 and 4.

The architectural differences between the Cray XK7 and XC30 platforms include both a different processor and a different network technology. These two intertwined facets of each platform can impact the performance of even a reduced application like Ziz in unexpected ways. Layering complexity in the Ziz

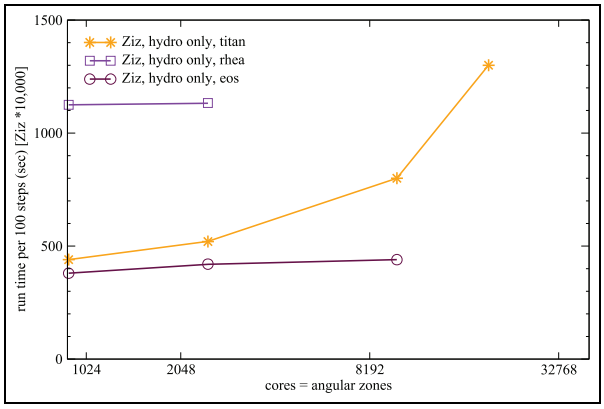


**Figure 4.** A comparison of hydro-only Ziz scaling on Titan, Eos, and Rhea. The ordinate (i.e. time required for 100 time steps) is multiplied by 10,000 for the Ziz results, just as in Figure 3.

MiniApp will allow each of the individual pieces of multi-physics modeled by the code to be measured and analyzed separately. In addition, the pairwise and higher-order interactions between the various modules can also be delineated in this manner.

### 5.2. Initial work on an OpenACC Ziz version

We also have a beta OpenACC version of Ziz under active development. This port is based on a version of MVH3 that was used as the "code lab" for an early 2012 OLCF GPU programing tutorial. Aside from a handful of loop reorderings that proved necessary for the port, little additional coding is required. Because Ziz relies heavily on the use of global variables declared and managed in Fortran modules (as does CHIMERA), care must be taken to properly scope `private` and `copyin` variables for `acc parallel` loops. Aside from these modest code changes, the only additional code present in the OpenACC version of Ziz is the addition of `acc parallel loop` directives around the single loops containing calls to the main PPM function ( `ppmlr`) in each directional sweep routine (`sweepx1`, `sweepx2`, `sweepy`, and `sweepz`). `ppmlr` contains the most computationally intensive hydro routine (`parabola`), which is already highly vectorizable. The result is a 94% increase in the overall per-rank performance on Titan

```
OpenAcc Ziz---speed = 887.4 kz/s/pe
Ziz---speed = 457.1 kz/s/pe.
```

The next step for the OpenACC port of Ziz will be to add accelerated versions of the LU decompositions performed in the transport and burning modules. Because these changes will, for the most part, be confined to formation of the individual Jacobians and the use of accelerated libraries, the expected code changes for these additions are expected to be as modest as those performed for the hydro module.

### 5.3 Using Ziz to investigate process placement on Titan

Node allocation in HPC systems can significantly impact application performance. This is due to the implications of node mapping upon the physical network. These impacts are particularly strong in 3D torus networks (Peña et al., 2013) which seek to reduce network costs but with some performance trade-offs. One such network is the Gemini Alverson et al. (2010) network used in the Titan supercomputer.

Titan's network is an anisotropic 3D torus connecting 19,200 compute and service nodes through 9600 Gemini routers. Each Gemini router connects two nodes to the network. Link materials and length result in directionally asymmetric performance. The X-dimension links are eight lanes of cable resulting in 9.4 GB/s (18.8 GB/s bi-directional). The Y-dimension alternates among four lanes of mezzanine and cable, performance achieving 9.4 GB/s and 4.7 GB/s, respectively. Finally, the Z-dimension is composed of eight lanes of backplane and achieves 15 GB/s. Latency is generally ordered mezzanine, backplane, and cable (from lowest to highest latency).

Given the heterogeneous performance characteristics of the Gemini network, we seek to understand the node allocation mechanism on Titan and determine if the choices made are best suit a diversified workload. Titan uses the Application Level Placement Scheduler (ALPS) for generating allocation ordering within the supercomputer. ALPS lays out the machine using the notion of space-filling curves within basic building blocks (Albing et al., 2011). The space-filling curves used in Titan are $2 \times 2 \times 2$ Hilbert curves. This provides single hop guarantees for 16 nodes within the grouping. Each of these smaller blocks is composed into a larger block of $4 \times 2 \times 4$ Gemini routers, these larger blocks are called basic building blocks. From the network perspective, a $4 \times 2 \times 4$ block spans four-wide in the X and Z dimensions and two-wide in the Y-dimension. This is due to the interleaved cable and mezzanine within the Y-dimension and this constrains intra-block communication to the more performant mezzanine links. The full layout of the machine uses basic building blocks chained together. When chaining blocks together, ALPS prioritizes dimensional placement as Z, Y, X. More specifically on Titan, which is $25 \times 16 \times 24$, we start at coordinate 0, 0, 0 and place six blocks in the Z-dimension before growing in the Y-dimension and placing six more blocks in the Z-dimension.

It is within the context of block chaining our study saw a potential performance improvement for larger scale applications. The block prioritization of Z, Y, X leads to reduced intra-job bandwidth for any jobs exceeding 384 compute nodes. This is because each block growth in the Y-dimension will introduce 4.7 GB/s Y-cable links to the underlying routes available to that application. While this approach works well for low-bandwidth 2D nearest neighbor communication many applications use more advanced techniques such as 3D stencils, collectives, or IO-aggregation that may suffer due to the increased frequency of communication traversals over slower links. To address this, we devised two new block layout strategies to be incorporated into ALPS that may better suit our general workload. One layout strategy was a simple reprioritization of the layout to Z, X, Y, as in Enos et al. (2014). This layout essentially prioritizes intra-job bandwidth above all other network factors. The second layout was a more balanced approach that alternates between Y and X building priorities to balance between latency and bandwidth.

A preliminary evaluation was performed using a full-system test-shot on Titan. During this evaluation, we ran a workload representative of the general applications on Titan at sizes ranging from 512 to 2048 nodes. Ziz was included amongst these applications at both a 1024 and 2048 node sample. In both cases, Ziz was impacted by the changes to the layout indicating non-trivial communication requirements. Run over 1024 nodes, Ziz benefited from the balanced allocation strategy, seeing a 1.5% overall performance improvement. This improvement increased to 7% in a larger 2048 node run. Both evaluations saw performance degradation in the bandwidth-centric layouts. In the case of the balanced allocations, Ziz benefited from improved network characteristics and reduced intra-node hop-counts. Typical default and balanced placements are shown in Figure 5 and Figure 6, respectively. We note that the subcommunicator-spanning ALLTOALLs central to the performance of Ziz map to these placement layouts in an intuitive way.

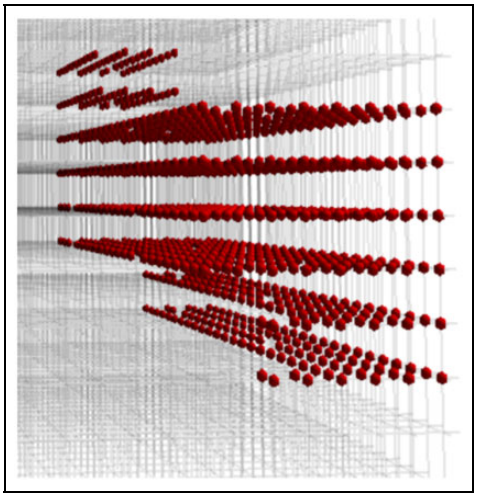

**Figure 5.** Default process placement via ALPS for 1024-node Ziz run on Titan. Shown is the layout with the Z-dimension normal to the page.

## 6 The Sweep MiniApp: a proxy for Denovo

The Sweep MiniApp is designed to be a performance proxy for the Denovo radiation transport code. Denovo solves the linear Boltzmann transport equation using the discrete ordinates method and has uses in nuclear technology applications (Evans et al., 2010). It is part of the SCALE nuclear safety analysis code package (SCALE, 2009) and is the primary deterministic radiation transport code employed by the Consortium for Advanced Simulation of Light Water Reactors (CASL) headquartered at Oak Ridge National Laboratory. The MiniApp and the the algorithms it is designed to capture are quite similar to the SNAP MiniApp developed at Los Alamos National Laboratory (Zerr and Baker, 2013).

The majority of execution time for a Denovo run (typically 80–99%) is generally spent in a 3D-sweep wavefront calculation. The Denovo sweep kernel has been successfully ported to GPUs for the ORNL Titan system (Baker et al., 2012). The Sweep MiniApp considered here is a reimplementation of the Denovo sweeper to target multiple current and future architectures under a wide range of alternative programing models.

In its current version, the Sweep MiniApp is comprised of roughly 6000 lines of source code, compared to Denovo's line count in excess of 200,000 lines. To allow for maximum portability to programing models such as OpenCL, which requires kernels to be implemented in C, the Sweep MiniApp is written in the C language using an unobtrusive object-like programing style. The code currently supports MPI, CUDA, OpenMP and Intel Xeon Phi coprocessor directives; ports to OpenCL, OpenACC, and OpenMP 4 are currently in progress.

As in previous sections, we concentrate first on the distributed memory scalability of the Sweep MiniApp versus that of Denovo. Figure 7 shows the comparative performance of the Denovo GPU sweeper and the Sweep MiniApp for a series of test cases in a weak scaling regime. In both cases the codes exhibit good weak scaling behavior, with slight efficiency loss at high node counts, which is to be expected due to the nature of the sweep algorithm. The MiniApp runs slightly slower than the Denovo sweeper because the latter code is more mature and more heavily tuned to the GPU architecture of Titan. The excellent agreement in the scaling behavior between the two codes validates the suitability

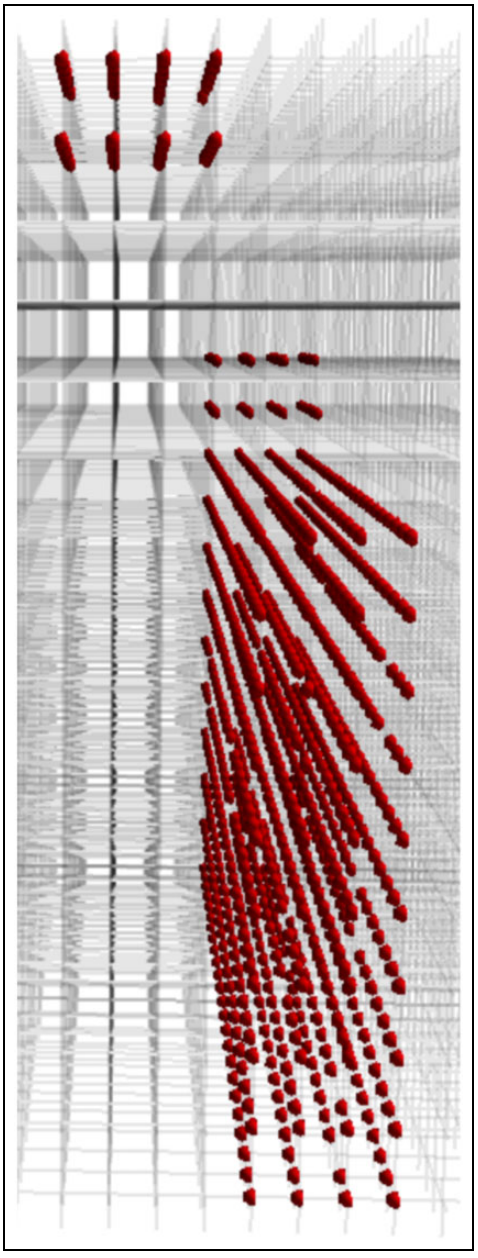

**Figure 6.** Balanced process placement via ALPS for 1024-node Ziz run on Titan. Shown is the layout with the Z-dimension normal to the page.

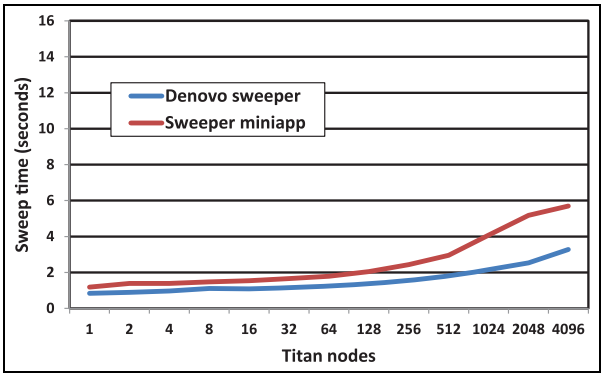


**Figure 7.** A comparison of weak scaling for the Denovo sweep kernel and the Sweep MiniApp.

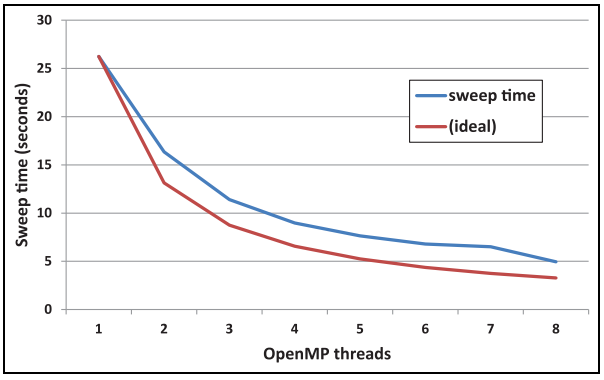


**Figure 8.** OpenMP strong scaling performance for the Sweep MiniApp.

of the Sweep MiniApp as a performance proxy for the Denovo sweep kernel.

The convergence of pre-exascale node architectures to a generally uniform configuration of cores, threads, and vectors makes it conceivable to support diverse architectures in a single code base. The Sweep MiniApp is written to support many threads per compute node, whether they be CUDA or OpenMP threads. Figure 8 shows a strong scaling study on one Titan processor socket for a test case using OpenMP threading. Though codes which execute sweeps for radiation transport are generally run in a weak scaling regime, as demonstrated by the Titan GPU and Beacon Intel Phi results, this figure gives a more stringent test to illustrate efficiency of on-node threading.

Figure 9 demonstrates weak scaling performance for hybrid MPI+OpenMP execution on the NICS Beacon system using a single Intel Xeon Phi processor. These runs use four OpenMP threads per core and scale across processor cores using MPI. The code is parallelized using OpenMP and Intel SIMD directives and is run on the coprocessor in native mode. Good weak scaling performance is attained. The fraction of processor floating point peak performance achieved on the Intel Phi is similar to what is obtained on a Titan GPU, roughly 5–10% of peak; though the sweep operation is flop-rich, the large amount of indexing required limits the number of registers that can be used. This indexing also limits the flop rate. Both these effects mean the kernel behaves in a memory-bound fashion.

## 7 The XRayTrace MiniApp

The XRayTrace MiniApp represents the primary computational component for a 3D coupled atomic-

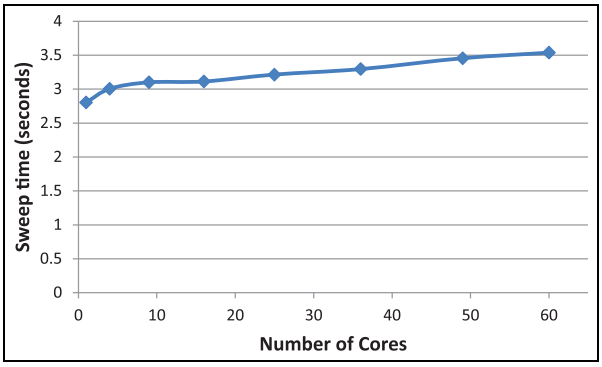


**Figure 9.** Weak scaling of the Sweep MiniApp on Intel Xeon Phi.

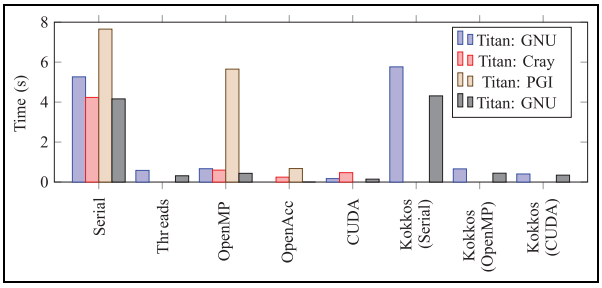


**Figure 10.** XRayTrace MiniApp performance for the small ASE problem.

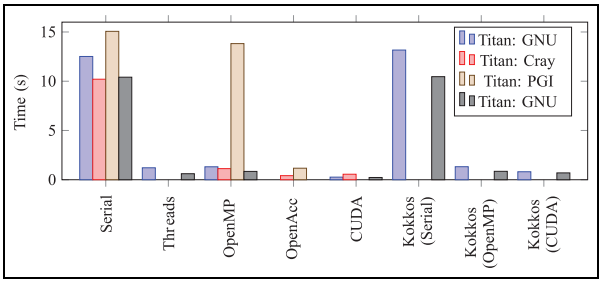


**Figure 11.** XRayTrace MiniApp performance for the medium ASE problem.

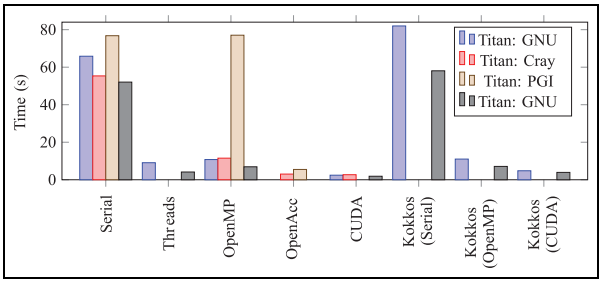


**Figure 12.** XRayTrace MiniApp performance for the small seed problem.

physics/ray-propagation code used to simulate ASE (Amplified Spontaneous Emission) and seeded X-ray lasers (Berrill, 2010; Wang et al., 2008). The MiniApp was developed to test new ideas and programing models in a lightweight application that perfectly mirrors the behavior of the ray propagation kernel in the full application. It solves the 3D ray propagation and amplification equations (1), and is auto-generated from the application source code to remain consistent with the main application. The MiniApp solves many independent rays in parallel aggregating the results to form an image that is used to couple the atomic physics in the main application and to generate the output images in the full application. The MiniApp does not utilize MPI since the problem decomposition and communication is handled by the main application and there is no global communication within the ray propagation kernel. The equations for amplification and ray propagation are

$$\begin{aligned} \frac{dI(\nu,\vec{\mathbf{r}},t)}{d\mathbf{s}} &= j(\nu,\vec{\mathbf{r}},t) + g(\nu,\vec{\mathbf{r}},t)I(\nu,\vec{\mathbf{r}},t) \\ \frac{d}{ds}\left(n\frac{\vec{\mathbf{r}}}{ds}\right) &= \nabla n \end{aligned} \tag{1}$$

where $I$ is the ray intensity, $j$ and $g$ are the emissivity and gain, $n$ is the index of refraction, and $\vec{\mathbf{r}}$ is the position vector.

The MiniApp includes several predefined problems of different sizes representing the various sizes that are used within a single node. There are two sizes (small and medium/large) for seeded and ASE lasers. Larger problems within a node are not typically run with the production code due to the resulting walltime required: More computational nodes are typically used instead. Figures 10, 11, 12 and 13 show the times required by the MiniApp for each problem utilizing different methods of parallelization/offloading. The machines utilized were a single node of Titan which consists of a AMD Opteron 6274 processor and a NVIDIA Kepler GPU, while Emmet is a workstation containing two Intel Xeon E5-2630 v3 processors and four NVIDIA KTesla K40 GPUs (only one GPU was utilized for these tests). The times reported are the times for a single iteration within the MiniApp. The full application will typically include 10,000–100,000 calls to the computational kernel within a single simulation.

Examining Figure 13 as typical, we observe several general trends. The MiniApp achieves good—but not perfect—speedup via the application of CPU threads. CUDA speedups are maximal, while the Cray implementation of OpenACC is very competitive with the CUDA version. The PGI OpenACC implementation achieves speedups much lower than those seen when using the Cray compiler. More striking is the seeming lack of speedup seen with the PGI OpenMP implementation on the CPU. We stress that in all of

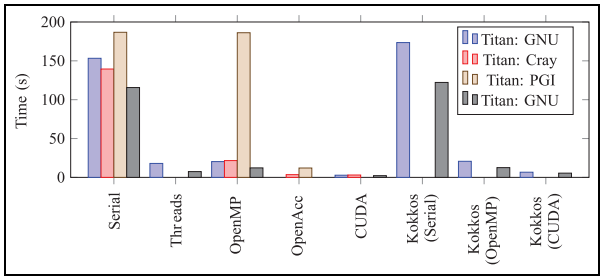


**Figure 13.** XRayTrace MiniApp performance for the medium seed problem.

Figures 10–13, we report the wallclock time for completion, and we also note that the ASE problems do exhibit mild speedup using the PGI OpenMP implementation. Given this, the lack of substantive speedup for the PGI implementation remains unexplained.

## 8 Summary and future work

Our experience with performance benchmarking and modeling for leadership-class supercomputers has convinced us of the necessity of keeping benchmark codes up-to-date with respect to their progenitor production codes. It is this notion that has motivated the empirical tests of performance behavior between MiniApps and full applications described in this paper. This same philosophy is reflected in our plans to produce and support versions of these (and other) MiniApps in the future. We plan to release public versions of the MiniApps to various communities, including vendors, computer scientists, developers, etc., using all of the programing models presented here. These codes will have to be updated and modified as new programing models come to the fore and older systems mature and change.

### Acknowledgements

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan). We thank Scott Atchley and Saurabh Gupta for their contributions to the Titan process placement work.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This research used resources of the Beacon project at the University of Tennessee, which is supported by the National Science Foundation under contract OCI-1137097.

### References

Scale (2009) SCALE: A modular code system for performing standardized computer analyses for licensing evaluations. Technical Report ORNL/TM-2005/39, Oak Ridge National Laboratory, Oak Ridge, TN, USA.

Albing C, Troullier N, Whalen S, et al. (2011) Scalable node allocation for improved performance in regular and anisotropic 3D torus supercomputers. In: *Proceedings of the 18th European MPI users' group conference on recent advances in the message passing interface*, EuroMPI'11, 2011, pp.61–70. Berlin, Heidelberg: Springer-Verlag.

Alverson R, Roweth D and Kaplan L (2010) The gemini system interconnect. In: *High performance interconnects (HOTI), 2010 IEEE 18th annual symposium on*, Mountain View, CA, 2010, pp.83–87.

Bailey DH, Barszcz E, Barton JT, et al. (1991) The NAS parallel benchmarks. *The International Journal of Supercomputer Applications* 5(3): 63–73.

Baker C, Davidson G, Evans T, et al. (2012) High performance radiation transport simulations: Preparing for TITAN. In: *Proceedings of the international conference for high performance computing, networking, storage and analysis (SC12)*, Salt Lake City, UT, 10–16 November 2012, pp.1–10.

Barrett RF, Crozier PS, Doerfler DW, et al. (2012) Summary of work for ASC L2 milestone 4465: Characterize the role of the mini-application in predicting key performance characteristics of real applications. Technical Report SAND2012-4667, Sandia National Laboratories.

Barrett v, Heroux MA, Lin PT, et al. (2011) Poster: Mini-applications: Vehicles for co-design. In: *Proceedings of the 2011 companion on high performance computing networking, storage and analysis companion (SC 2011 Companion)*, New York, NY, USA, pp.1–2, New York: ACM.

Berrill M (2010) *Modeling of laser-created plasmas and soft x-ray lasers*. PhD Thesis, Colorado State University, USA.

D'Azevedo E and Dongarra J (2000) The design and implementation of the parallel out-of-core ScaLAPACK LU, QR, and Cholesky factorization routines. *Concurrency and Computation: Practice and Experience* 12(15):1481–1493.

Enos J, Bauer G, Brunner R, et al. (2014) Topology-aware job scheduling strategies for torus networks. In: *Proceedings of Cray user group meeting-CUG-2014*, Lugano, Switzerland, May 2014.

Evans TM, Stafford AS, Slaybaugh RN, et al. (2010) Denovo: A new three-dimensional parallel discrete ordinates code in scale. *Nuclear Technology* 171(2): 171–200.

Futral S, et al. (2003) The ASCI purple benchmark codes. Available at: https://asc.llnl.gov/computing_resources/purple/archive/benchmarks/ (accessed 3 August 2016).

Gunter BC, Reiley WC and Van de Geijn RA (2001) Parallel out-of-core Cholesky and qr factorization with pooclapack. In: *Proceedings of the 15th international parallel & distributed processing symposium*, IPDPS '01, Washington, DC, USA, pp.179–, IEEE Computer Society.

Heroux MA (2011) Miniapplications: Vehicles for co-design. In: *Presentation at 15th Workshop on Distributed Supercomputing (SOS15)*, 14 March 2011, Engelberg, Switzerland.

Heroux MA, Doerfler DW, Crozier PS, et al. (2009) Improving performance via mini-applications. Technical Report SAND2009-5574, Sandia National Laboratories.

Hornung RD and Keasler JA (2013) A case for improved C++ compiler support to enable performance portability in large physics simulation codes. Technical Report LLNL-TR-635681, Lawrence Livermore National Laboratory.

Jaeger EF, Berry LA, D'Azevedo E, et al. (2002) Advances in full-wave modeling of radio frequency heated, multidimensional plasmas. *Physics of Plasmas* 9(5): 1873–1881.

McMahon FH (1986) The Livermore Fortran kernels: A computer test of the numerical performance range. Technical Report UCRL-53745, Lawrence Livermore National Laboratory.

Marcus R (2012) MCMini: Monte Carlo on GPGPU. Technical Report LA-UR-12-23206, Los Alamos National Laboratory.

Messer OEB, Bruenn SW, Blondin JM, et al. (2008) Multidimensional, multiphysics simulations of core-collapse supernovae. *Journal of Physics Conference Series* 125: 012010.

Messer OEB, Harris JA, Parete-Koon ST, et al. (2013) Multicore and accelerator development for a leadership-class stellar astrophysics code (*Lecture Notes in Computer Science*, 7782), p.92.

Pakin S and McCormick P. (2013) Hardware-independent application characterization. In: *2013 IEEE international symposium on workload characterization (IISWC)*. Institute of Electrical & Electronics Engineers (IEEE).

Peña AJ, Carvalho RGC, Dinan J, et al. (2013) Analysis of topology-dependent MPI performance on Gemini networks. In: *Proceedings of the 20th European MPI users' group meeting, EuroMPI 2013*, New York, NY, pp.61–66. ACM.

Springmeyer R, Still C, Schulz M, et al. (2011) From petascale to exascale: Eight focus areas of R&D challenges for HPC simulation environments. Technical Report LLNL-TR-474731, Lawrence Livermore National Laboratory.

Stone A, Dennis JM, Strout MM, et al. (2011) The CGPOP miniapp, version 1.0. Technical Report CS-11-103, Colorado State University.

Wang Y, Granados E, Pedaci F, et al. (2008) Phase-coherent, injection-seeded, table-top soft-x-ray lasers at 18.9 nm and 13.9 nm. *Nature Photonics* 2(2): 94–98.

Zerr RJ and Baker RS (2013) SNAP: SN (discrete ordinates) application proxy: Description. Technical Report LA-UR-13-21070, Los Alamos National Laboratory, Los Alamos, NM, USA.

## Author biographies

*OE Bronson Messer* received his PhD in physics from the University of Tennessee in 2000. He was a postdoctoral fellow at Oak Ridge National Laboratory and The University of Chicago before becoming a staff member in the Scientific Computing Group at ORNL in 2005. His research interests include supernovae, stellar evolution, radiative transfer, astrophysical combustion, high performance computing, and numerical relativity.

*Ed D'Azevedo* is a staff scientist in the Computational Mathematics Group at the Oak Ridge National Laboratory. He obtained his PhD in 1989 from the Department of Computer Science in the Faculty of Mathematics at the University of Waterloo, Ontario, Canada. D'Azevedo's current research interests include developing highly scalable parallel solvers that can take advantage of acceleration on GPUs and in enhancing the parallel efficiency of scientific applications running on leadership supercomputers.

*Judith Hill* received a PhD in Computational Science and Engineering in 2004 from Carnegie Mellon University. She is currently Liaison Task Leader in the Scientific Computing Group at ORNL. Previously, from 2005 to 2008, she was a member of the Applied Math and Applications department at Sandia National Laboratories. Dr. Hill's research interests are in the area of computational fluid dynamics and PDE-constrained optimization, numerical methods, and large-scale computing.

*Wayne Joubert* received his PhD in mathematics from The University of Texas in 1990. He worked at Los Alamos National Laboratory in the 1990s, developing solver algorithms and software for high performance systems. Since 2008 he has been a member of the Scientific Computing Group at ORNL, where he specializes in performance analysis of applications and porting of algorithms and software to advanced architectures.

*Mark Berrill* received his PhD in Electrical Engineering from Colorado State University in 2010. He was a wigner fellow at Oak Ridge National Laboratory from 2010–2012 and was a DOE CSGF fellow from 2006–2010. His research interests are computational infrastructures, laser created plasmas, X-ray lasers, and large-scale parallel computing.

*Christopher Zimmer* is a HPC Systems Engineer in the National Center for Computational Sciences at Oak Ridge National Laboratory. He received his PhD. from North Carolina State University in Computer Science in 2012. His research interests include understanding the impacts to performance and predictability associated with system topology and scheduling in supercomputers.