Enabling Computational Science on the Cray XT3

Nick Nystrom, **Deborah Weisser**, **Junwoo Lim**,
Yang Wang, **Shawn T. Brown**, **Raghu Reddy**, *and*
Nathan T. Stone, *Pittsburgh Supercomputing Center;*
Paul Woodward *and* **David Porter**,
University of Minnesota,
Tiziana Di Matteo, *Carnegie Mellon University; and*
L. V. Kalé *and* **Gengbin Zheng**, *The University of Illinois at Urbana-Champaign*

ABSTRACT: *BigBen, the Pittsburgh Supercomputing Center's 2090-processor Cray XT3, entered production on October 1, 2005, bringing unprecedented capability to the NSF TeraGrid. Rapid deployment of production applications at scale was necessary to realize the full potential of the XT3. Aggressive efforts on precursor systems led to 4 full applications running on actual XT3 hardware by SC2004, and subsequent efforts have led to the majority of production applications running at full scale and with outstanding performance. These efforts have been highly collaborative, with users interacting closely with PSC scientists during a "friendly user" period, workshops, and ongoing, daily interaction. In this paper, we snapshot the state of applications on BigBen, highlight recent scientific and performance results, and describe several initiatives through which PSC and its collaborators are continuing to add value to the XT3.*

KEYWORDS: Cray XT3,

1. Introduction

BigBen, the Pittsburgh Supercomputing Center's 2090-processor Cray XT3, entered production[1] on October 1, 2005, bringing unprecedented capability to the National Science Foundation (NSF) TeraGrid[2]. As the first delivered Cray XT3, a commercial product based on Sandia National Laboratories' *Red Storm*[3] system, BigBen is, beyond its role as a production resource, an experiment in deploying a new tightly coupled parallel architecture for the national science and engineering community.

Funded by the National Science Foundation, BigBen is designed to provide capability computing across a very wide range of applications. PSC's users, distributed across the United States, represent diverse fields of science with particular emphasis on biology, chemistry, physics, geoscience, and engineering but also including social sciences, political science, economics, etc. This diversity of applications mandates a well-balanced architecture to deliver interconnect and memory bandwidth and latency commensurate with computational speed, together with high-performance I/O, large aggregate memory, and a scalable operating system.

BigBen contains 2,090 AMD Opteron[4] processors, of which 2,068 are typically configured as compute nodes with the remaining 22 serving as service nodes (serving login, filesystem, and networking needs). Each Model 150 2.4 GHz Opteron delivers 4.8 Gflop/s of peak performance (not considering SSE instructions), providing 10 Tflop/s theoretical peak aggregate. The Cray SeaStar interconnect[5] is configured on BigBen as a 3-dimensional torus. Link bandwidth is extremely high, 6.5 GB/s sustained, which together with the 3D torus topology leads to exceptional communications performance. The standard programming model is MPI[6]; however, certain applications developed at PSC and being developed elsewhere directly use the Portals[7] API, on which the XT3's MPI is implemented. The XT3 runs Cray's UNICOS/lc operating system, which consists of two pieces: the Catamount[8] microkernel, which provides a scalable, efficient operating system on compute nodes, and a full Linux™ distribution on service nodes to provide I/O, networking, login, and other system services.

Each BigBen compute node contains 1 GB of memory, of which approximately 900 MB is available to applications (assuming default environment settings). Primary storage is through the Lustre[9] parallel filesystem, which provides up to 200 TB of rotating storage and is interfaced to PSC's hierarchical storage manager, SLASH[10].

This paper is organized as follows. In Section 2, we review PSC's chronology in realizing rapid productivity on BigBen, which predated hardware delivery and proceeded to include a "friendly user" period, training, and ongoing consulting and development. Section 3 highlights several interesting scientific results and applications, representing an unfortunately small subset of users' work but illustrating different aspects of how the XT3 fits into PSC's workload. Finally, in Section 4 we survey, briefly, aspects of performance that have profound bearing on applications' efficiency, citing results from both standard benchmarks and actual production codes.

2. Realizing Productivity

National computing resources are significant investments, for which maximum return is realized only if meaningful science is produced throughout the lifetime of the system.

2.1 Efforts Predating System Delivery

To ensure that full applications would be available immediately on delivery of Cray XT3 Serial Number 1, PSC staff began porting applications to Opteron processors under Linux™ and using the PGI compilers (which are standard on the XT3) in January 2004. There were no significant surprises, and this first step was valuable in establishing a firm foundation from which to proceed as hardware and software concurrently matured.

In mid-February, 2004, several PSC scientists were granted access to ASCI Red at Sandia National Laboratories. To gain experience with Cougar, the microkernel from which Catamount was derived, and with communications built on the Portals2 messaging protocol. In particular, we sought to determine early, prior to committing to the XT3, whether features (threads, forks, TCP, etc.) omitted from Cougar for the purpose of increased scalability would impede applications of interest. PSC successfully ported a variety of applications to this environment, corroborating our believe that the microkernels' performance gains would not be offset by limited applicability.

As XT3 software matured, so too did its hardware. In close cooperation with Cray, in March 2004, PSC began experimenting with a SeaStar simulator implemented on a pair of FPGAs. The purpose of this exercise was to move one step closer to running on an actual XT3, which did not yet exist, learning about any potential application issues and possibly providing feedback to Cray.

In April 2004, we began using a "look-alike system" at Cray built on Opteron workstations running Catamount, MPI, and a TCP implementation of Portals3. With the software APIs now much closer to those expected on the XT3, successful ports of important applications in materials science, molecular dynamics, weather, earthquake modeling, and cosmology built confidence that not only would applications run successfully when the XT3 was delivered, but that scientists' productivity would be high. Work on the look-alike system continued until late 2004, concurrently with other development activities, at which time we transitioned our efforts to the actual XT3 system.

On June 23, 2004, PSC had its first access to the SeaStar ASIC, together with Catamount and Portals3. As a shared resource time was limited, but by July, molecular dynamics and earthquake simulation codes were both running on the SeaStar.

PSC installed IA32-based look-alike systems running the Puma 2002 operating system and Portals in October, 2004, providing PSC developers with additional resources to port additional applications and otherwise expand the scope of their work.

These early efforts paid off in early November, 2004, when a Cray XT3 cabinet was delivered to PSC for display at SC2004 (Pittsburgh, November 6-12, 2004). On November 6, within a day of the XT3 being powered up in PSC's booth, four full applications were running: Quake (earthquake simulation), ARPS (storm forecasting), LSMS (materials science), and Gasoline (cosmology). This apparently immediate success was only possible because of the effort which preceded it, and which lay the groundwork for the many applications which would quickly follow.

2.2 The Friendly User Period

PSC took delivery of the first row (11 cabinets) of BigBen during the final week of December, 2004. After initial configuration and testing, December 30 marked the start of BigBen's friendly user period. Steven Gottlieb, a physicist at Indiana University, stepped up to the challenge and within a day was successfully running MILC, a code developed by himself and colleagues, to probe fundamental physics through very large simulations in lattice quantum chromodynamics.

PSC's friendly user period serves two purposes. First, it makes a valuable resource available to members of the scientific community prior to the official start of production. In return for their willingness to work in an evolving, occasionally unstable environment as the system is built and tuned, the friendly users can accomplish relatively ambitious simulations. Scientific results are thereby obtained even as installation progresses, ensuring that maximum value is obtained through the full lifetime of the resource. Second, the friendly user period provides valuable feed-

back to PSC through additional applications, requirements, and workflows that only real users can bring to bear. This feedback identifies potential issues as early as possible, facilitates porting and optimization of key applications, and provides a realistic workload on which realistic system configuration and tuning can occur.

44 research groups spanning NSF directorates and divisions and from across the country participated in BigBen's friendly user period. Some were invited based on their known resource requirements and roles in developing important applications, whereas others asked for and were granted access.

2.3 Training

To ensure that users received in-depth technical training specific to the Cray XT3, PSC hosted workshops focusing on porting and optimizing users' applications in August and October, 2005. Each 4-day workshop consisted of approximately 1.5 days of instruction and approximately 2.5 days of time to work on applications. Cray provided instructors expert in optimization, tools, and compilers.

These workshops were well-attended. The first, during the "friendly user" period but open to all, was attended by 25 users representing 13 institutions. Highlights included resolution of an issue which was impeding NAMD and porting of Tcl to the XT3 compute nodes, identification and resolution of a latent performance issue in DNS, building Cactus, debugging ASH, and incorporating shmem into a Co-Array Fortran translator. 16 users from 5 institutions attended the second workshop. Illustrating the value of the application-oriented approach, one participant observed, "Yesterday I improved the code 30%, and today 5-10% more." Other highlights during the second workshop included debugging CPMD, an *ab initio* molecular dynamics code critical to certain very large allocations, and sustaining over 300 jobs per day for the last 3 days of the workshop, providing valuable empirical data for scaling XT3 resources to heavy production use.

2.4 Production

BigBen entered production service on October 1, 2005, as scheduled. The distinction was primarily one of accounting: allocations awarded on the basis of peer-reviewed proposals began to accrue time, but as a consequence of the successful "friendly user" period, the transition to production was seamless.

3. Performance

Before discussing applications, it is useful to understand aspects of the XT3's design which contribute to scalability, especially for bandwidth-intensive codes. Such codes become increasingly important at large scale because of vast amounts of information that must be exchanged. Single-processor efficiency is also critical, for which the Opteron's Hypertransport ameliorates the "memory wall" notorious for reducing commodity processors' realized efficiencies.

3.1 HPCC

The HPC Challenge benchmark[11] implements four local benchmarks (DGEMM, STREAM, RandomAccess, and FFT) and four global benchmarks (HPL, PTRANS, FFT, and RandomAccess) to quantify several critical dimensions of system performance including floating point execution rate, memory bandwidth, and interconnect latency and bandwidth.

Table 1. HPCC benchmark results for Cray XT3 and similarly sized systems.

		G-HPL		G-PTRANS	G-Random Access	G-FFTE	EP-STREAM Triad	EP-DGEMM	Random Ring BW	Random Ring Latency
	PEs	(TFlop/s)	(% of peak)	(GB/s)	(GUP/s)	(GFlop/s)	(GB/s)	(GFlop/s)	(GB/s)	(μs)
TCS baseline[a]	3000	4.215	70.2%	72.500	0.140	24.202	0.220	1.762	0.041	19.162
Cray XT3 baseline[a]	2048	7.713	78.5%	302.870	0.348	425.660	4.764	4.387	0.342	8.971
Cray XT3 optimized[a]	2048	8.113[b] 8.279[c]	82.5% 84.2%	302.870	0.355	575.799	5.789	4.387	0.342	8.971
Cray XT3 optimized[a]	5208[d]	20.409	81.6%	944.227	0.672	761.729	4.660	4.412	0.206	9.200

a. Measured by PSC, January-February, 2006.
b. On compute nodes configured with 1 GB of memory (bigben, PSC).
c. On compute nodes configured with 2 GB of memory (jaguar, ORNL).
d. Measured by PSC on jaguar (ORNL).

In Table 1, we present HPCC results for the Cray XT3. To address scaling to higher node counts (applicable to all HPCC global benchmarks) and larger amounts of memory (instrumental in realizing maximal values for the HPL benchmark), we also include timings from jaguar, Oak Ridge National Laboratory's Cray XT3. For comparison, HPCC results are also listed for TCS ("lemieux"), the 3000-processor HP AlphaServer SC system at PSC (1.0 GHz Alpha EV68 processors, 4 processors (PEs) per node, 2-rail Quadrics Elan3, bandwidth approximately 250 MB/s per rail, full fat tree topology). Baseline runs, representing builds of HPCC as distributed, are presented for all systems listed. We also present optimized results for the XT3, corresponding to system-specific optimizations and careful selection of problem sizes.

G-HPL (global High Performance LINPACK) results for the XT3 demonstrate an impressive fraction of the theoretical peak, i.e. 78.5% (7.713 Tflop/s) for the unoptimized baseline case on 2048 nodes of bigben and 82.5% of theoretical peak (8.113 Tflop/s) for the optimized case. Exploiting the larger memory, 2 GB/node, of jaguar, fastidious specification of input parameters increased the optimized rate to 84.2% of theoretical peak (8.279 Tflop/s) on 2048 nodes. The optimized case benefits from a shmem-based implementation of all-to-all (courtesy of John Levesque, Cray, Inc.), which is an important operation in HPL. Exploiting the larger memory increases efficiency by maximizing the amount of data that can be streamed through cache into the floating-point units. The lower value of only 81.6% at 5208 nodes is due to settling for less well optimized input parameters given the extremely large size of the commensurate runs. That these values are all considerably higher than those for TCS reflects both communications performance (broadcast and allgather, in particular) and the importance of memory bandwidth in keeping the floating-point pipelines busy.

G-PTRANS, the global parallel transpose, depends primarily on interconnect bandwidth. When installed, TCS had the highest-bandwidth interconnect available, a 2-rail Quadrics fat tree, which attains 72.500 GB/s on G-PTRANS. The XT3's SeaStar shines for G-PTRANS, achieving 256.383 GB/s (baseline) and 261.472 GB/s (optimized) using 2048 nodes. Here, the baseline and optimized cases are essentially the same, with the difference attributed to job placement and other system activity[12]. At 5208 nodes, G-PTRANS records an immense 944.227 GB/s on the XT3, and the XT3 easily leads current HPCC results for the bandwidth-intensive G-PTRANS benchmark.

We see a similar pattern for G-FFTE, the parallel 3D Fast Fourier Transform benchmark. Again, the XT3 gets excellent marks, achieving 425.660 Gflop/s for the baseline case on 2048 nodes relative to only 24.202 Gflop/s on TCS. Like G-PTRANS, G-FFTE depends strongly on interconnect bandwidth. Also as with G-PTRANS, the XT3 leads reported HPCC G-FFTE results for similarly sized systems. In the case of G-PTRANS, the shmem-based implementation of all-to-all dramatically improves efficiency by 35% to 575.799 Gflop/s. Many applications in CFD, electronic structure, cosmology, and other fields have 3D FFTs as important kernels, imbuing G-PTRANS with very tangible relevance.

G-RandomAccess, which measures the rate of random updates of memory, depends on both interconnect and memory latency. The XT3 performs well at 0.348 GUP/s (baseline) on 2048 nodes and 0.672 GUP/s (optimized HPCC, but no direct optimization to G-RandomAccess) on 5208 nodes.

The local benchmarks EP-STREAM Triad and EP-DGEMM demonstrate the processor's memory bandwidth and floating-point capacity, respectively. The XT3 does very well for each: EP-STREAM benchmark results are 4.764 GB/s for the baseline case and an impressive 5.789 GB/s for the optimized case, which uses an assembly-coded triad kernel (also courtesy of John Levesque, Cray, Inc.). The EP-DGEMM benchmark, which computes a matrix-matrix multiply on each processor, runs at 4.387 Gflop/s, i.e. 91.4% of theoretical peak.

Random Ring bandwidth and latency measure the average bandwidth and latency in randomly ordered rings. Measured under UNICOS/lc Release 1.3.09, we measured a random ring bandwidth of 0.251 GB/s and a random ring latency of approximately 10 μs. (This is distinct from ping-pong latency, which we current measure at 5.70 μs.)

The HPCC results are directly comparable against those for other systems; however, it is often difficult to translate the results of synthetic benchmarks into actual application performance. While the full gamut of application performance transcends the space available here, it is nonetheless useful to consider two particular case studies.

3.2 The criticality of bandwidth: Understanding the dynamics of pollutant particles

PSCC, the Parallel Spectral Channel Code, is a highly scalable parallel flow solver developed by Junwoo Lim and colleagues to enable more realistic simulations of turbulent boundary layer flows, including tracking the traces of massless particles under stably stratified conditions. This project is sponsored by NSF and is a part of the international collaboration research program between PSC and KISTI (Korea Institute of Science and Technology Information) Supercomputing center. PSCC simulations will improve our understanding of the dynamics of pollutant particles in atmospheric boundary layers, which is a pressing problem in environmental science.

PSCC depends heavily on FFTs and discrete cosine transforms and is therefore highly bandwidth-intensive. To gauge the effectiveness of the SeaStar interconnect,

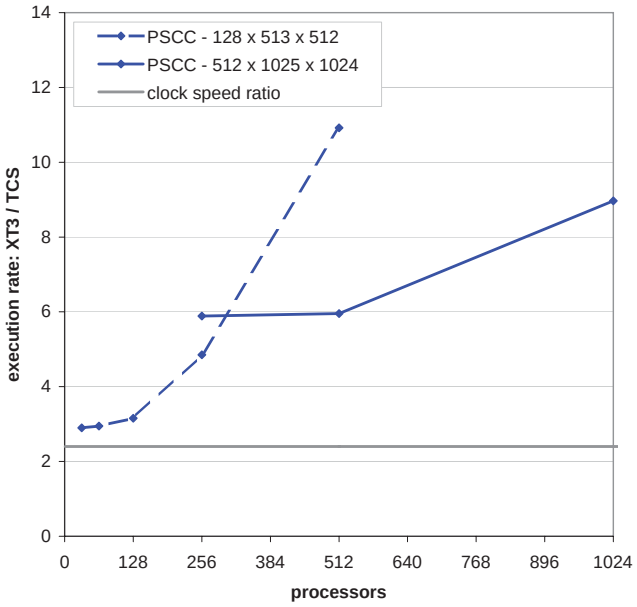


Figure 1. Performance of PSCC, a Parallel Spectral Channel Code to study stratified turbulent boundary layers, on BigBen (Cray XT3) relative to TCS (HP AlphaServer SC, 2-rail Quadrics Elan3). The ratio of clock speeds is 2.4; however, PSCC runs up to 10.9 times faster on the XT3 due to the SeaStar sustaining high bandwidth. PSCC quickly saturates the Quadrics fabric on TCS, leading to increasingly strong relative performance on the XT3 as node count increases.

we consider the performance of PSCC on bigben (Cray XT3) relative to TCS (described in Section 3.1). A naive comparison of the two architectures solely on the basis of clock speeds (XT3: 2.4 GHz, TCS: 1.0 GHz; both feature dual-issue floating point) would suggest that the XT3 is approximately 2.4 times as powerful as TCS, processor-for-processor. However, as seen in Figure 1, the XT3 does considerably better than floating-point rates alone would suggest. For the smaller problem considered, a $128 \times 513 \times 512$ grid, XT3 performance begins at 32 processors is 2.9× TCS performance and then rises steeply and progressively with processor count, achieving 10.9× TCS performance by 512 processors. The larger problem, a $512 \times 1025 \times 1024$ grid, begins at 5.9× TCS performance on 256 processors (the minimum required to provide the necessary memory, on either system) and increases to 9.0× TCS performance on 1024 processors.

The primary reason for these stunning performance gains is the exceptional link bandwidth (6.5 GB/s) of the SeaStar interconnect, which affords ample capacity to support the also-high injection bandwidth (1.1 GB/s).

3.3 LSMS

LSMS[13], which implements the Locally Self-consistent Multiple Scattering method, facilitates understanding magnetic domain walls from first principles. Historically, this is one of the greatest challenges to *ab initio* electronic structure calculations. The LSMS method is a first-principles $O(N)$ scaling technique particularly well-suited to studying magnetic properties of alloys (Figure 2). It was the first application to sustain 1 TFlop/s (on the Cray T3E), and it sustains 4.65 TFlop/s on TCS (theoretical peak: 6 Tflop/s). Developed by G. M. Stocks (ORNL), Y. Wang (PSC), and others, LSMS received the 1998 Gordon Bell award for best achievement in high performance computing. The Cray XT3 will enable realistic quantum mechanical simulation, e.g. study of the dynamics of magnetic switching processes, of real nanostructures.

LSMS achieves the highest sustained performance to date on the XT3 for a production application. Executing on 2048 bigben processors, LSMS sustains 8.03 Tflop/s (82% of the theoretical peak of 9.83 Tflop/s)[14]. This exceptional performance is a consequence of LSMS effectively mapping its data structures into cache, together with expression of its key algorithms as matrix-matrix multiplications.

LSMS also served as an early (fall 2005) test of the XT3's stability prior to entering production. This 52 hour, 57 minute run on 2000 Cray XT3 processors, terminated only to proceed with a scheduled software upgrade, provided empirical support for a high mean time between failures, even for capability-class simulations.

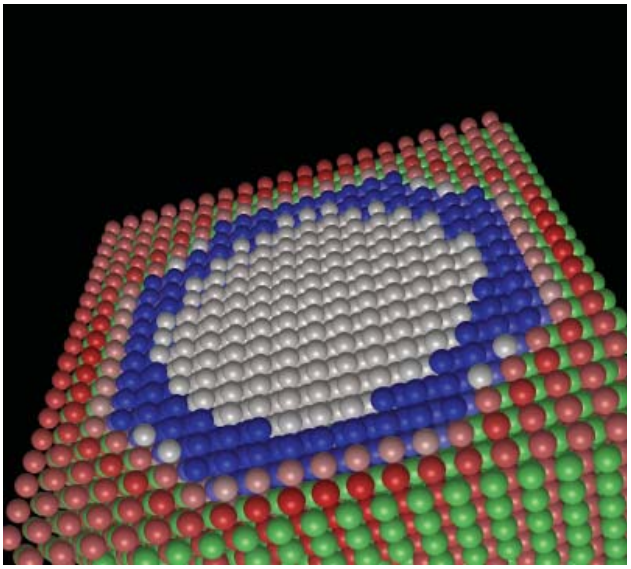


Figure 2. A sliced view of the magnetic Fe nanoparticle (with 4409 atoms on a BCC lattice) together with the surrounding FeAl matrix (with 11591 atoms on a B2 lattice), modeled by LSMS. The calculated charge distribution within the nanoparticle and its surrounding atoms is indicated by the color change from the center to the edge.

4. Applications

The XT3 has proven highly effective for running PSC's diverse and demanding workload, with nearly all production applications running at scale. Table 2 summarizes applications that run at or near scale on BigBen. Applications routinely scale to the full system configuration, and indeed, a larger configuration would benefit almost all. Additional applications are under development by PSC and users.

The few applications listed as not scaling to the full configuration are limited by load imbalance inherent to their algorithms and/or implementations. Even in these cases, scaling is as good or better than seen on other platforms, and absolute performance is exceptionally good. For example, BigBen was the first system to break the 10 ns/day simulation barrier for AMBER's PMEMD module, in the context of Factor IX runs done by Bob Duke at NIEHS[15].

Table 2. A wide range of applications spanning diverse fields of science run at scale on the Cray XT3.

application	domain	nodes
HPCC	HPC Challenge benchmarks	5208
NAMD	Molecular dynamics	5000
HOMME	Atmospheric dynamical core	5000
Charm++	Parallel C++ runtime system used by NAMD, OpenAtom, and other applications	5000
WRF	Weather Research & Forecasting	4096
MILC	Quantum Chromodynamics	4096
MPQC	Massively Parallel Quantum Chemistry	2067
LSMS 2.0 LSMS 1.6	Materials science / electronic structure	2048 2048
HYCOM	Ocean modeling	2048
GAMESS	General Atomic and Molecular Electronic Structure System: quantum chemistry	2048[a]
PSCC	Parallel Spectral Channel Code: stratified turbulence	2048
PPM	Piecewise Parabolic Method: computational fluid dynamics and turbulence	2048
Quake	Earthquake modeling	2048
Gasoline	N-body astrophysics	2048
CPMD	Car-Parrinello Molecular Dynamics	2048
DNSmsp	Direct Numerical Simulation	2048
Dynamo	Quantum mechanical/molecular mechanical (QM/MM) modeling of reactive sites in biochemical systems	2048
ABINIT	Electronic structure; plane-wave density functional theory	2048
CHARMM	Molecular dynamics	2048
PARATEC	Parallel Total Energy Code; electronic structure	2048
S3D	Turbulent combustion	2048
MHD	Magnetohydrodynamics	2048
OOCORE	Out-of-Core solver benchmark	2048
Leo	Numerical relativity: solves Einstein equations in vacuum, modeling single black hole spacetimes	1944[a]
Gadget	Smoothed Particle Hydrodynamics	1800[b]
ZEUS-MP	Astrophysics	1728
AMBER	Molecular Dynamics	1024[c]
QChem	Quantum Chemistry	1024[c]

a. The data decomposition of Leo requires $6N^2$ processors. $6 \times 182 = 1944$ is the largest possible Leo run on BigBen.
b. Gadget is fully expected to scale to and beyond 2048 XT3 processors. 1800 processors merely reflects recent production calculations.
c. AMBER and QChem can run on higher node counts; however, for the problems examined, load imbalance limits their effective scalability. This reflects limitations of the algorithms and their implementations, rather than fundamental limitations of the physical processes being modeled and/or the computational platform.

4.1 Interactively Steering Turbulence Simulations: BigBen as a TeraGrid Resource

Paul Woodward and David Porter of the University of Minnesota conceived an ambitious plan: transform simulations of turbulence, long regarded as being computationally intensive, from a batch workflow with timescales on the order of weeks to an interactive session of approximately one hour.

The Cray XT3 enabled this new, real-time execution mode of Woodward and Porter's PPM[16] (Piecewise Parabolic Method) code because of the high-bandwidth, low-latency SeaStar interconnect. Rather than running simulations at progressively higher resolution, Woodward and Porter sought to run simulations that were already at the right resolution faster. Placing extreme demands on interconnect performance, this is actually the most difficult scenario.

Running PPM interactively, however, was only one part of the challenge. To enable real-time, remote visualization and steering, data would also have to flow across the NSF TeraGrid, a high-performance optical network linking sites throughout the Unites States, from PSC to Woodward's visualization lab in Minnesota and other sites where users would be located. Recall that to promote scalability, XT3 compute nodes run the Catamount microkernel, which does not support likely candidates for outbound communications such as sockets and TCP. Direct access to compute nodes can also pose challenges for security and scheduling

To address the need for efficient communications from the XT3 to the TeraGrid (actually, to wide area networks in general), PSC scientists Nathan Stone and R. Reddy devised PDIO[17] (Portals Direct I/O), a library which presents a file-like API through which applications can direct data to a remote system. XT3 compute processors direct data over Portals to PDIO daemons running on service processors, which then redirect the data over the TeraGrid at up to 240 Mb/s. PDIO then assembles incoming data into complete files at the user's site.

In the case of the turbulence simulations, PDIO aggregated distinct data streams from 512 compute nodes into 11 streams for transmission across the TeraGrid. The files

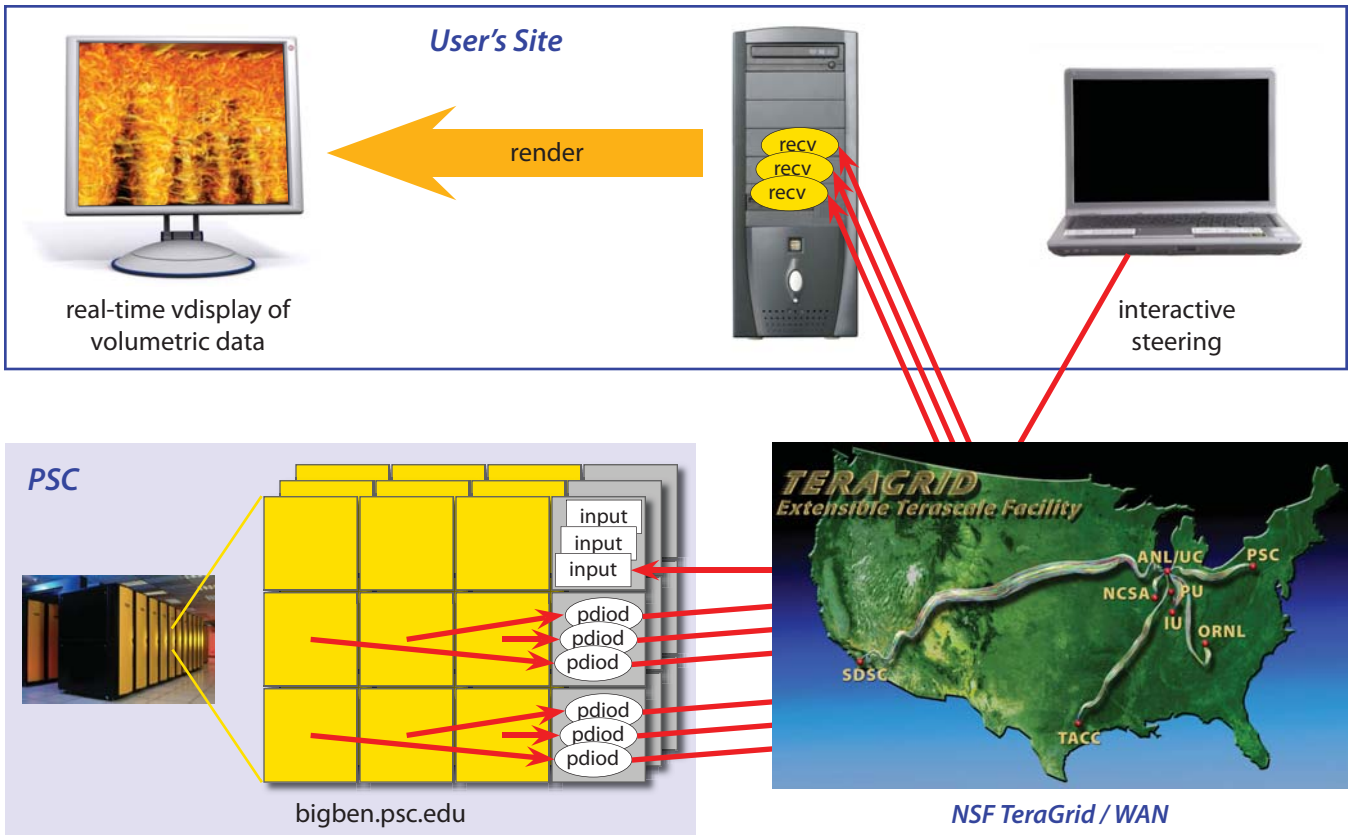
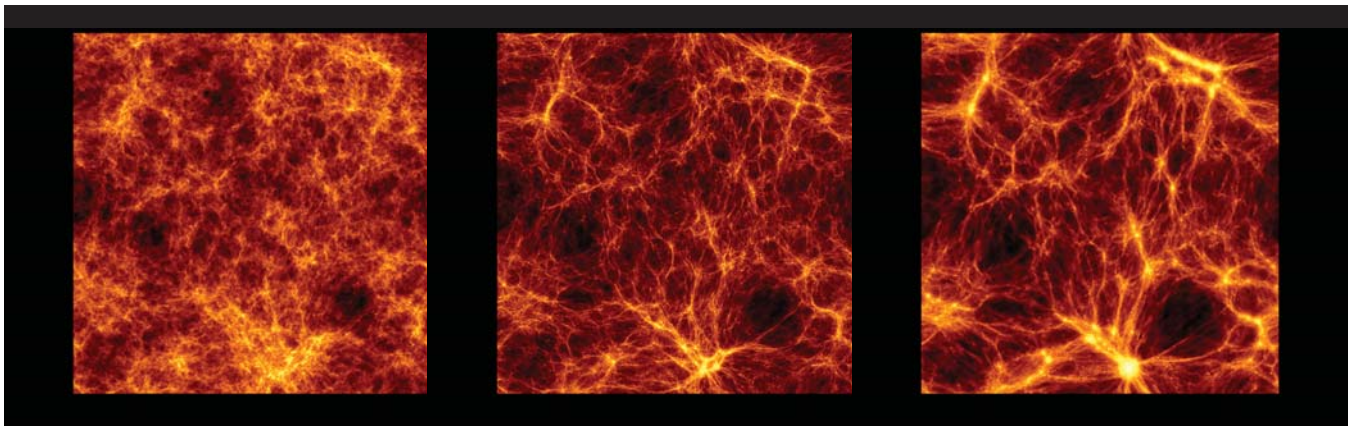


Figure 3. Paul Woodward and David Porter worked with PSC to develop a distributed architecture for interactive exploration of turbulence simulations, made possible by the Cray XT3's remarkable interconnect bandwidth and use of the TeraGrid's high-performance optical backbone. PSC's PDIO library aggregates data through a file-like API from XT3 compute nodes running Catamount through *pdiod* daemons on the XT3's service nodes, which assemble the data into a small number of streams for efficient transmission across the TeraGrid. At the user's site, incoming PDIO data are reassembled into files for interactive volume rendering, and a steering client provides dynamic control as simulations evolve.



Redshift z=14
Very early universe, approximately
200 million years after the big bang

Redshift z=4
About 1.5 billion years old; about 1/10
the present age of the universe

Redshift z=1.6
About 5 billion years old; approximately
what the universe looks like today

Figure 4. Three snapshots of gas density from a very high-resolution, 2×486^3 full hydrodynamical (stars, gas, and black holes as well as dark matter) cosmology simulation run on Cray XT3 by Tiziana Di Matteo (Carnegie Mellon University). The frames are of successively later redshifts. The small points are dwarf galaxies, which can only be resolved through high-resolution simulations.

are then volume-rendered, again in real time, by David Porter's Hierarchical Volume Renderer (HVR) to examine properties such as vorticity, velocity divergence, and density. Utility programs including a steering client allow for interactive control of the simulation, affecting parameters such as Mach number. This functionality was demonstrated at IGrid (San Diego, September 2005), IEEE Vis (Minneapolis, October 2005), and SC|05 (Seattle, November 2005).

Additional work was recently done to parallelize PPM on a more finely grained level, made possible by the XT3's excellent interprocessor bandwidth, to increase the scalability of even moderate-resolution simulations from 512 to 2048 nodes. PDIO is evolving as well: PDIO2, nearing completion, will offer a more convenient interface to PDIO's functionality, intercepting I/O calls directed to a specific device rather than requiring (albeit minor) source code modifications.

Changing the way scientists approach turbulence research through leveraging the XT3, TeraGrid, and local resources will increase scientific productivity by allowing rapid investigation of "what if" scenarios.

4.2 Cosmology

Galaxy formation is one of the most important areas of cosmological research. In collaboration with research groups at Harvard and the Max Planck Institute, Tiziana Di Matteo of Carnegie Mellon University is studying galaxy formation by performing direct cosmological simulations.

Simulations of the formation and evolution of the universe pose an immense challenge. In addition to treating a vast dynamic range of spatial and temporal scales, cosmology simulations must also include the effect of gravitational fields generated by superclusters of galaxies on the formation of new galaxies, which in turn harbor gas that cools to create stars and which is funneled into supermassive black holes.

The GADGET-2 code[18] computes gravitational fields using an optimal combination of hierarchical tree algorithms and Fast Fourier Transforms (tree-PM), and represents fluids by means of smoothed particle hydrodynamics. Both the force computation and the time stepping in GADGET are fully adaptive, with a dynamic range that is, in principle, unlimited. GADGET simulations are memory intensive and, because of the FFT, require very high interconnect bandwidth.

Previously, the largest simulation that had been run was a pure N-body simulation which followed the formation of structure by gravitational instability (only dark matter), lacking any of the hydrodynamics of the gas which is necessary to compare the simulation results to an actual galaxy survey. That simulation required 1 TB of memory and produced 20 TB of data.

The Cray XT3 opened the door to calculations with similar memory usage, but including the full hydrodynamics. These ongoing simulations include all relevant processes for understanding star formation and black hole growth in galaxies, constituting the most detailed physical models for galaxy formation. Figure 4 illustrates snapshots of gas density from the highest-resolution (2×486^3), full-hydrodynamical region of this size ever run. The frames are of successively later redshifts, ranging from only 200 million to approximately 5 billion years after the big bang.

The small points are dwarf galaxies, which can only be resolved through high-resolution simulations. Filaments and galaxies at this resolution show unprecedented structure.

The XT3 speeds both the gravity and the hydrodynamic aspects of GADGET-2. Performance on the Cray XT3 is approximately 4 times that on TCS (see §3.2) due to high interconnect bandwidth and efficient parallel I/O.

4.3 NAMD

NAMD[19], a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems, is critical to the scientific investigations of numerous research groups using PSC's resources. NAMD received the 2002 Gordon Bell award for unprecedented parallel performance on a challenging computational problem, where the award-winning simulation of ATP Synthase (approximately 327,000 atoms) ran on over 2000 processors on PSC's TCS.

Recent performance runs with NAMD have shown good scalability through 5000 Cray XT3 processors for the million-atom STMV protein (167,063 protein atoms + 299,855 waters). Running on 5000 processors, each time step takes only 0.27 seconds, clearly illustrating the importance not only of the XT3's high-bandwidth interconnect but also of its jitter-free operating system.

However, Charm++[20] the parallel C++ library on which NAMD is built, was initially ported to the XT3 using MPI, leading us to believe that additional performance might be realized if direct support for Portals was added to Charm++. A visit to L. V. Kalé's group at the University of Illinois at Champaign-Urbana quickly ascertained that improvements are, in fact, possible. Figure 5 shows two of the screen captures from the Projections[22] performance tool. The left screen capture summarizes all processors' activity. The single lines are the load balancing and the green parts are the actual computation of time steps. If this computation were completely load balanced, then the green areas would be solid. The right screen capture focuses on specific activities: colored sections signify NAMD activity, white signifies idle time, and black signifies Charm++ and MPI overhead, which is substantial. Charm++/projections is able to identify idle time separately from the overhead (which, for MPI, is sometimes hard: idle time looks like time spent in an MPI call). The exploded view shows that there is significant overhead for one processor sending a message to itself, which implies that there is overhead associated with the MPI interface of Charm++. In particular, much time is being lost to MPI_Iprobe(), as has been observed on certain other MPI implementations.

To improve the efficiency of NAMD and other applications which rely on Charm++, the Kalé group, in collaboration with PSC, is now implementing Charm++ directly on Portals (as was done for Elan3 on TCS). This port is also expected to provide optimal handling of the many pending *get* operations that can otherwise overflow the Portals event queue.

5. Conclusion

To ensure that as it entered production the first Cray XT3 would be immediately productive to the national research

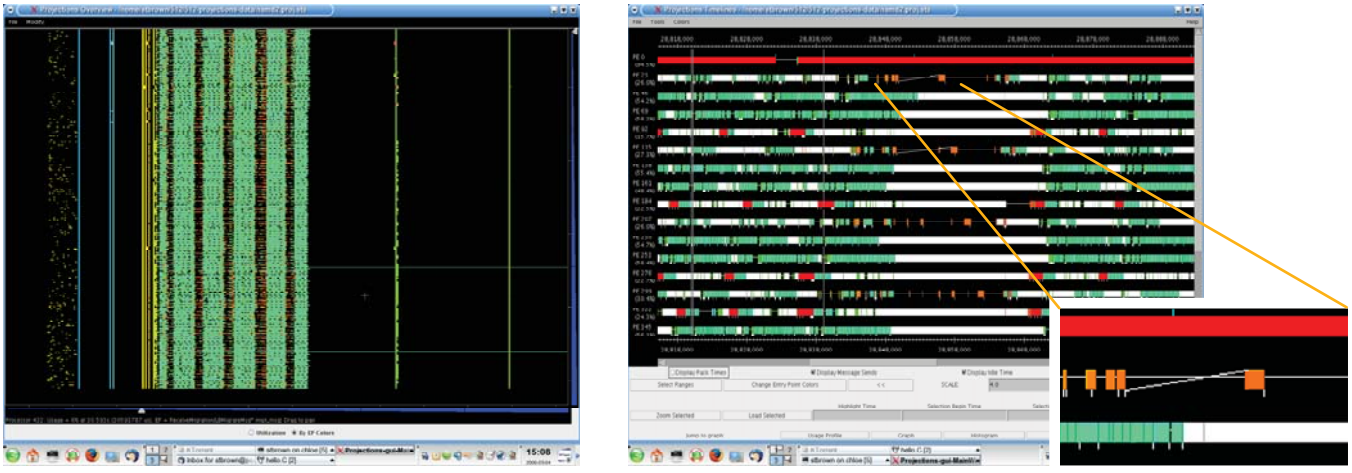


Figure 5. Two views of the Projections tool for analyzing performance of NAMD and other Charm++ applications. The left screen capture summarizes all processors' activity. The single lines are the load balancing and the green sections are the actual computation of time steps. If this computation were completely load balanced, then the green would be solid. The right screen capture focuses on specific activities: colored sections signify NAMD activity, white signifies idle time, and black signifies overhead, which is significant. The exploded view shows that there is significant overhead for one processor sending a message to itself, which implies that there is overhead associated with the MPI interface of Charm++.

community, PSC undertook a thorough and methodical approach to porting and optimizing applications, engaging users, and providing advanced training. This process, while time-intensive, was successful: production applications were running and demonstrated at SC2004, within days of receiving pre-release hardware, and scientific progress began with the friendly user period in late 2004. BigBen entered production on October 1, 2005, precisely on schedule, and today serves a broad range of research in biology, physics, chemistry, neuroscience, materials science, engineering, political science, and social science. Performance on standard benchmarks as well as production applications is excellent. Scientific throughput is high, not only for capability-class simulations, but also for moderate simulations which scale to unprecedented levels. PSC continues to develop applications and supporting software, in close collaboration with users, to advance computational science on the XT3 and across the TeraGrid.

Acknowledgments

The authors would like to thank Sue Kelly (Sandia National Laboratories) for access to ASCI Red and Red Storm, as well as Jeffrey Vetter (Oak Ridge National Laboratory) for access to jaguar. John Levesque, Luiz DeRose, and Sarah Anderson provided in-depth expertise at PSC's Cray XT3 workshops. Luiz has also provided invaluable access to and guidance on the use of Cray's performance tools. We would also like to thank PSC computational scientists who contributed to porting and optimizing the many applications mentioned here, including Jeff Gardner, Roberto Gomez, David O'Neal, and John Urbanic.

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. SCI-0456541.

About the Authors

Nick Nystrom is Director of Strategic Applications at the Pittsburgh Supercomputing Center (PSC), focused on advancing understanding through computational science. Nick has been active in developing and optimizing applications for Cray architectures ranging from the X-MP through the XT3. Deborah Weisser, a Performance Specialist in PSC's Strategic Applications Group, is active in a variety of topics including performance modeling, interconnect optimization, and parallel algorithms. Junwoo Lim is a Senior Scientific Specialist at PSC. His primary interest has been the parallel implementation of computational fluid dynamics (CFD) problems. Yang Wang is a Senior Scientific Specialist at PSC. His research interests are mainly focused on computational materials science, and he is one of the leading figures in the development of the LSMS method. Shawn T. Brown is the Senior Support Specialist in Computational Chemistry at PSC. Shawn has worked in development of quantum chemistry code and is now active in application and development of massively parallel computational chemistry software. Raghu Reddy is a member of the Strategic Applications group at PSC and has been working with researchers in performance analysis, modeling and tuning of various applications. Nathan T. Stone is a member of PSC's Advanced Systems Group, where he has developed high-performance file transfer mechanisms including tcscp and PDIO. Nick, Deborah, Junwoo, Yang, Shawn, Reddy, and Nathan can be reached at PSC, 300 South Craig St., Pittsburgh, PA 15213 USA. Their e-mail addresses are Nick Nystrom: nystrom@psc.edu; Deborah Weisser: dweisser@psc.edu; Junwoo Lim: jlim@psc.edu; Yang Wang: ywg@psc.edu; Shawn Brown: stbrown@psc.edu; R. Reddy: rreddy@psc.edu; Nathan T. Stone: nstone@psc.edu.

Paul Woodward is Astronomy Professor, Minnesota Supercomputer Institute Fellow, and Laboratory for Computational Science and Engineering Director, University of Minnesota. He has focused his research on simulations of compressible flows in astrophysics, studying problems in star formation, supersonic jet propagation, convection in stars, and astrophysical turbulence. Paul can be reached at niversity of Minnesota / LCSE, 499 Walter Library, 117 Pleasant Street SE, Minneapolis, MN 55455, e-mail paul@lcse.umn.edu.

David Porter is Sr. Research Associate at the University of Minnesota's Laboratory for Computational Science and Engineering. David can be reached at University of Minnesota / LCSE, 429 Walter Library, 117 Pleasant Street SE, Minneapolis, MN 55455, e-mail dhp@lcse.umn.edu.

Tiziana Di Matteo is Professor of Physics at Carnegie Mellon University, whose research interests focus on studies of black holes encompassing a wide range of topics in high energy physics and cosmology including theoretical studies of the interplay between black hole growth and galaxy formation and investigations of various aspects of the physics of accretion disks around black holes. Tiziana can be reached at Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, e-mail tiziana@phys.cmu.edu.

Laxmikant (Sanjay) V. Kalé is Professor of Computer Science and Director of the Parallel Programming Laboratory at the University of Illinois at Urbana-Champaign. Sanjay's research interests center on design and implementation of pragmatic tools that facilitate development of efficient, scalable parallel applications, focusing on problems in dynamic load balancing, performance visualization and analysis, parallelization of symbolic and search computations, data parallel programming, and expression of flow of control within objects. Sanjay can be reached at Parallel Programming Laboratory, Computer Science, 405 N. Mathews Ave., Beckman Institute, Urbana, IL 61801,

e-mail kale@cs.uiuc.edu.

Gengbin Zheng is a a Postdoctoral Research Associate at the Center for Simulation of Advanced Rockets. Gengbin's research interests include Charm++ runtime optimization techniques including dynamic load balancing, Charm++ applications including NAMD and rocket simulations, simulations of parallel architectures, and performance analysis. Gengbin can be reached at Parallel Programming Laboratory, Computer Science, 405 N. Mathews Ave., Beckman Institute, Urbana, IL 61801, e-mail gzheng@ks.uiuc.edu.

References

1. www.teragrid.org
2. *Pittsburgh Center Unveils a Bigger, Faster Supercomputer Called "Big Ben"*, http://www.nsf.gov/news/news_summ.jsp?cntn_id=104315 and http://www.psc.edu/publicinfo/news/2005/2005-07-20-xt3.html.
3. *Red Storm to be assembled in New Mexico: Sandia supercomputer to be world's fastest, yet smaller and less expensive than any competitor*, http://www.sandia.gov/news-center/news-releases/2004/comp-soft-math/redstormrising.html.
4. AMD Opteron™ Processor Tech Docs, http://www.amd.com/us-en/Processors/TechnicalResources/0,,30_182_739_9003,00.html.
5. R. Brightwell, K. Pedretti, and K. D. Underwood, *Initial Performance Evaluation of the Cray SeaStar Interconnect*, 13th Symposium on High Performance Interconnects (HOTI'05), Palo Alto, California, 2005, pp. 51-57.
6. M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra, MPI: The Complete Reference, The MIT Press, Cambridge, Massachusetts, 1996. Also available at http://www.netlib.org/utk/papers/mpi-book/mpi-book.html.
7. R. Brightwell, A. B. Maccabe, R. Riesen, and T. Hudson, The Portals 3.3 Message Passing Interface, Revision 1.0, May 16, 2003, available at http://www.cs.sandia.gov/Portals/.
8. S. M. Kelly and R. Brightwell, *Software Architecture of the Light Weight Kernel, Catamount*, Cray User Group 2005 Proceedings, Knoxville, Tennessee, 2005.
9. http://www.lustre.org/
10. P. Nowoczynski, N. Stone, J. Sommerfield, B. Gill, and J. R. Scott, *Slash - The Scalable Lightweight Archival Storage Hierarchy*, 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST'05), Monterey, California, 2005, pp. 245-252.
11. J. Dongarra and P. Luszczek, *Introduction to the HPCChallenge Benchmark Suite*, ICL Technical Report, ICL-UT-05-01, (Also appears as CS Dept. Tech Report UT-CS-05-544), 2005.
12. D. Weisser, N. Nystrom, C. Vizino, S. T. Brown, and J. Urbanic, *Optimizing Job Placement on the Cray XT3*, Cray User Group 2006 Proceedings, Lugano, Switzerland, 2006.
13. Y. Wang, G. M. Stocks, W. A. Shelton, D. M. C. Nicholson, Z. Szotek, W. M. Temmerman, *Order-N Multiple Scattering Approach to Electronic Structure Calculations*, Phys. Rev. Lett., 75, 2867 (1995).
14. Y. Wang, G. M. Stocks, D. M. C. Nicholson, A. Rusanu, and M. Eisenbach, *Quantum Mechanical Simulation of Nanocomposite Magnets on Cray XT3*, Cray User Group 2006 Proceedings, Lugano, Switzerland, 2006.
15. Bob Duke, NIEHS, private communication, September 29, 2005.
16. P. Colella and P. R. Woodward, *The Piecewise-Parabolic Method (PPM) for Gas Dynamical Simulations*, J. Comp. Phys. 54, 174-201 (1984).
17. N. Stone, D. Balog, B. Gill, B. Johanson, J. Marsteller, P. Nowoczynski, R. Reddy, J. R. Scott, J. Sommerfield, and K. Vargo, and C. Vizino, *PDIO: Interface, Functionality and Performance Enhancements*, Cray User Group 2006 Proceedings, Lugano, Switzerland, 2006.
18. V. Springel, *The cosmological simulation code GADGET-2*, MNRAS 364, 1105-1130 (2005).
19. J. C. Phillips, G. Zheng, S. Kumar, and Laxmikant V. Kalé, *NAMD: Biomolecular Simulation on Thousands of Processors*, Proceedings of the 2002 ACM/IEEE Conference on Supercomputing, Baltimore, Maryland, 2002, pp 1-18.
20. L. V. Kalé and S. Krishnan, Charm++: Parallel Programming with Message-Driven Objects, in Parallel Programming using C++ (Eds. Gregory V. Wilson and Paul Lu), pp 175-213, MIT Press, 1996. http://charm.cs.uiuc.edu/
21. L. V. Kalé, G. Zheng, C. W. Lee, and S. Kumar, *Scaling Applications to Massively Parallel Machines Using Projections Performance Analysis Tool*, PPL Paper Number 04-05, http://charm.cs.uiuc.edu/papers/namdPerfFGCS.shtml.