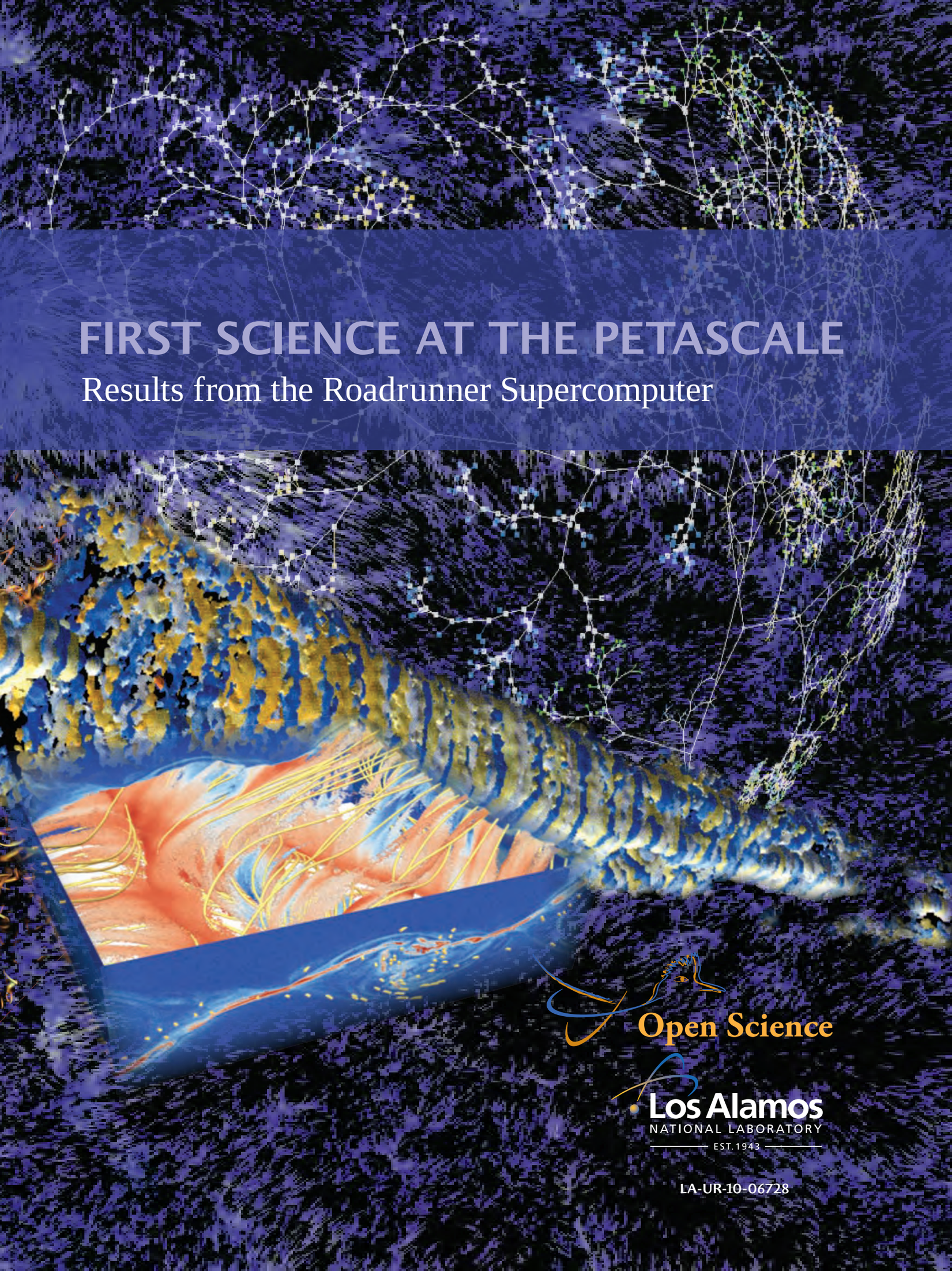


FIRST SCIENCE AT THE PETASCALE

Results from the Roadrunner Supercomputer

Open Science

Los Alamos
NATIONAL LABORATORY
EST. 1943

LA-UR-10-06728

table of contents

From the Roadrunner Project Director 2
Andrew B. White, Deputy Associate Director for Theory, Simulation, and Computation

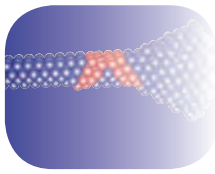
The hybrid architecture of the world's first petaflop/s supercomputer 6
Kenneth R. Koch, Roadrunner Technical Manager

The Roadrunner coding effort 10
Jamaludin Mohd-Yusof, Applications Team Member

Understanding the nonlinear physics of laser-plasma interaction through "at scale" plasma kinetic simulations on Roadrunner 16
Lin Yin[1], Brian J. Albright[1], Kevin Bowers[2], and Ben Bergen[1]

VPIC–Magnetic reconnection 20
William Daughton[1], Brian Albright[1], Ben Bergen[1], Vadim Roytershteyn[1], Lin Yin[1], and Kevin Bowers[2]

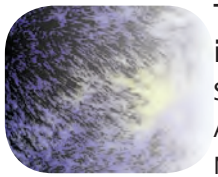
New frontiers in viral phylogenetics 24
Tanmoy Bhattacharya[1], Marcus Daniels[1], and Bette Korber[1]



Simulating the mechanical behavior of metallic nanowires over experimentally accessible timescales on Roadrunner 28
Danny Perez[1], Chun-Wei Pao[3], Sriram Swaminarayan[1], and Arthur Voter[1]

Material dynamics at extreme conditions 32
Timothy C. Germann[1], Sheng-Nian Luo[1], and Sriram Swaminarayan[1]

Direct numerical simulation of reacting compressible turbulence 36
Daniel Livescu[1], Jamaludin Mohd-Yusof[1], and Timothy M. Kelley[1]



The Roadrunner Universe Project: baryon acoustic oscillations in the intergalactic medium 40
Salman Habib[1], Adrian Pope[1], Zrija Luckic[1], David Daniel[1], Patricia Fasel[1], Katrin Heimann[1], Chung-Sing Hsu[1], Lee Ankeny[1], Graham Mark[1], Suman Bhattacharaya[1], James Ahrens[1], Martin White[4], Jordan Carlson[4], and Nehal Desai[5]

Glossary 45

---

1 Los Alamos National Laboratory; 2 D.E. Shaw and Company; 3 Academia Dinica, Taiwan; 4 University of California Berkeley; 5 Aerospace Corporation

From the Roadrunner Project Director

Andrew B. White, Associate Deputy Director, Theory, Simulation, and Computation

Roadrunner is the first petaflop/s supercomputer. Its hybrid design makes it the most power-efficient supercomputer in the world. It is now fully engaged in nuclear weapons science and engineering simulations.

Roadrunner began to take shape in 2002. At that time, scaling a supercomputer to 4,000 nodes was manageable. However, even at that scale, a petascale system would require nodes operating at more than ¼ of a teraflop/s. Achieving this level of performance within a reasonable power envelope required something extraordinary–a modification of the Sony, Toshiba, and IBM-developed Cell chip used in the PlayStation3.

The original design was to be delivered in two phases. The initial phase was delivered in 2006, an AMD Opteron™-based cluster with empty space in the cabinets to accommodate future Cell chips in 2008. However, in late 2006 and early 2007, IBM and LANL redesigned Roadrunner as a standalone integrated 1.4 petaflop/s system. This redesign had two primary advantages: the initial delivery system provided uninterrupted valuable resources to the weapons program, and the connection speeds on the integrated system by were increased by 4x.

In October 2007, the Roadrunner project passed Department of Energy (DOE) review, thus allowing IBM and LANL to proceed to build, deliver, and accept the full system. This review focused on four important DOE applications codes: Implicit Monte Carlo (radiation transport), VPIC (plasma physics), SPaSM (molecular dynamics), and Sweep3D (neutron transport). In June 2008, Roadrunner became the very first system to achieve a sustained petaflop/s.

This report details seven open science projects that were selected to be part of the stabilization and integration of Roadrunner. Selection was based on scientific quality and the ability to develop an accelerated science code within the allotted time, about 6 months. All of the projects described here were transformational, building computational scientific capabilities that did not exist before.

A particle-in-cell code, VPIC, one of the original Roadrunner codes, was the vehicle for predicting energy deposition in fusion experiments, such as at the National Ignition Facility (NIF), at an "unprecedented range of time and space scales." In another project, an innovative computational science technique using SPaSM allowed nanoscale materials simulations "more than a thousand times longer than conventional techniques would have allowed."

Two entirely new computational physics capabilities were developed during this project. The DNS (direct numerical simulation) project performed "the largest reacting compressible turbulence simulations to date," focusing on relevance to astrophysics and including radiation and electron transport. The Roadrunner Universe project developed a hybrid approach to cosmological simulations, using the power of the Cell processor to increase the dynamic range of the simulation by a factor of 10–100.

At a petascale, Roadrunner is providing critical resources for DOE mission and science applications. However, its most lasting contribution will likely be as a harbinger of the future of high performance computing, a future both challenging and replete with opportunity.

First science at the petascale: Results from the Roadrunner supercomputer

The Roadrunner supercomputer at LANL was the world's first supercomputer to exceed the benchmark 1 petaflop/s computing rate in June 2008. Named for the state bird of New Mexico, it was conceived by LANL, and designed and built by IBM. Roadrunner demonstrates the future of high performance computing with its fast, power-efficient hybrid architecture. Over the next decade, supercomputers will be judged as much on their power efficiency as for their speed, and Roadrunner is leading the way.

"From February until September, 2009, Roadrunner was focused on a set of unclassified, open science projects, developing a new suite of hybrid computer codes for modeling and simulation in a wide variety of disciplines."

The secret to Roadrunner's record-breaking performance is the use of compute nodes that consist of two AMD Opteron dual-core processors and four IBM PowerXCell 8i™ processors (used as computational accelerators). The PowerXCell 8i chips used in the machine are a slightly modified version of the Cell processor chip used in the Sony PlayStation 3®. These node-attached Cell accelerators are what make Roadrunner different from typical clusters, and we call it a Cell-accelerated hybrid supercomputer.

The sheer performance and size of Roadrunner has enabled first-of-a-kind computer codes and science simulations of the biggest of the big and the smallest of the small during its first 8 months of operation at LANL. From February to September, 2009, Roadrunner was focused on a set of unclassified, open science projects, developing a new suite of hybrid computer codes for modeling and simulation in a wide variety of disciplines. The machine then switched to a classified mode for its intended mission, and since November 2009, codes and simulations are being used on Roadrunner to help assure the safety, security, and reliability of the US nuclear deterrent. Roadrunner was acquired as part of the DOE National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) Program.

Roadrunner's open science period had two interrelated purposes:
1) check out, stabilize, and optimize the machine and its system software, and 2) develop several hybrid computer simulation codes in order to run full-scale science simulation studies for a set of peer-reviewed projects. Earlier work on a few computer codes had already proved the efficacy of the hybrid design of Roadrunner, thus there was great interest in using the enormous power of the machine to accomplish additional large-scale simulations of significant scientific value.
Ten open science projects were selected from proposals drawn from within LANL for areas of important scientific merit and teams staffed with the appropriate mix of skilled subject matter experts, computational scientists, and programmers. A few projects would build on earlier code demonstration efforts, but for many projects the Roadrunner hybrid architecture would be completely new and would require an entirely new code base. Computational scientists engaged in the earlier demonstration codes worked directly on several of these projects. The range of science represented in these projects was extremely varied, as seen in this report.

During development, open science projects explored and exercised new programming techniques, system software, and tools for the hybrid Cell-accelerated Roadrunner architecture, demonstrating the usability and performance of the unique hybrid architecture across a broad range of disciplines. Once the code development effort was complete, these projects ran scientific studies, often comprising many large simulation runs, to examine important scientific questions or issues. Collectively, these runs provided an early and challenging workload on the system, enabling computer scientists to optimize the system and application software, and to work the kinks out of the operations as well.

Seven successful open science projects are presented in this document.

Laser-Plasma Interaction

This work is focused on understanding the nonlinear aspects of fusion experiments in the NIF at LLNL by studying the physics of onset and saturation of stimulated Raman scattering (SRS) in the fundamental building block of an NIF laser beam, a single laser speckle.

Exploring Magnetic Reconnection

Magnetic reconnection is a basic process that occurs within hot ionized gases known as plasmas. This process often leads to an explosive release of energy that is stored within the magnetic fields, and plays a key role in the earth's magnetosphere, solar flares, magnetic fusion machines, and a variety of astrophysical problems.

The Largest Hiv Evolutionary Tree

Mapping Darwinian phylogenic evolutionary relationships for large numbers of Human Immunodeficiency Virus (HIV) genetic sequences results in an HIV family tree that may lead researchers to new vaccine focus areas.

Modeling Tiny Nanowires at Long Time Scales

Nanowires stretching and breaking under stress is simulated atom-by-atom over a period of time, which is closer than ever to experimental reality, to see how the movement of single atoms can change a material's mechanical or electrical properties.

How Shock Waves Cause Materials to Fail

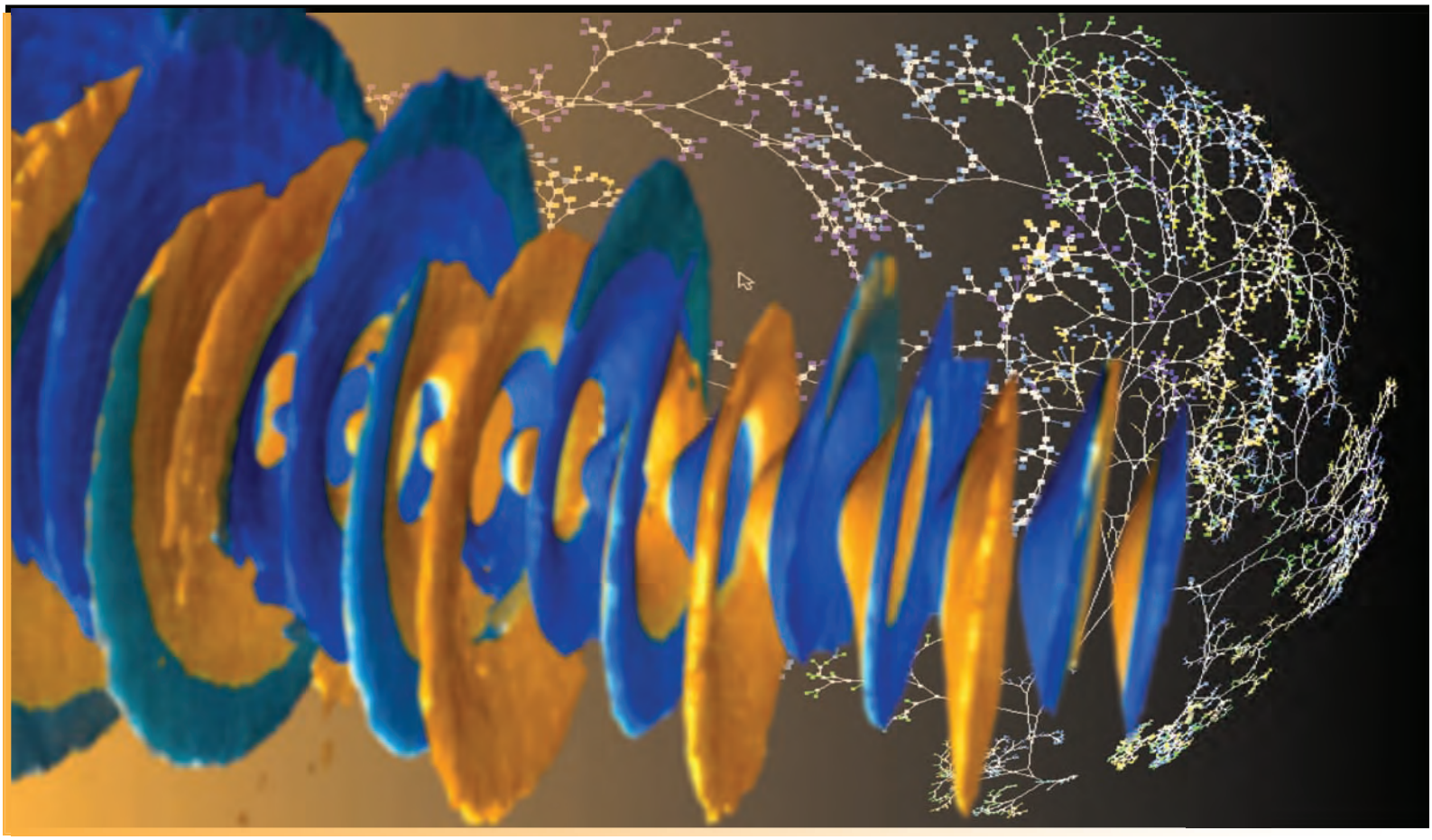
Physicists use the SPaSM (Scalable Parallel Short-range Molecular Dynamics) computer code to conduct multibillion-atom simulations of materials as extreme shock-wave stresses break the materials into pieces. For the first time atomic-scale models reveal the detailed formation and transport of ejecta.

Direct Numerical Simulation of Reacting Turbulence

In the field of fluid dynamics, understanding turbulence, the chaotic behavior of fluids, remains one of the unsolved problems in physics. This study focused on the complex interactions of flame and turbulence, as in the early stages of a type Ia supernova.

Origins of the Unseen Universe

Astrophysicists have created some of the largest high-resolution simulations of the distribution of matter in the Universe. These simulations model an expanding, accelerating Universe to better understand the structure of the Universe, dark energy, and dark matter.



The hybrid architecture of the world's first petaflop/s supercomputer

Kenneth R. Koch, Roadrunner Technical Manager

Roadrunner was conceived as a radical new approach to creating a supercomputer that transcends expected limitations of building larger and larger conventional computers.

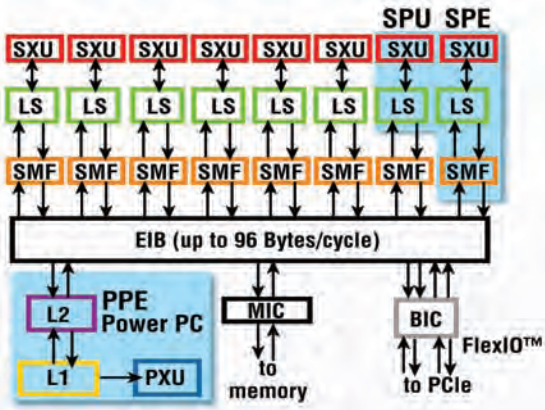
Roadrunner was the first supercomputer to reach the goal of 1 petaflop/s. A "petaflop/s" is 1 quadrillion, or a million-billion, calculations per second (peta signifying the number 1 followed by 15 zeros, a quadrillion, and flop/s is shorthand for floating point operations per second, e.g., a multiply or an add). On May 26, 2008, the Roadrunner supercomputer reached a sustained speed of 1.026 petaflop/s on the industry benchmark known as LINPACK just after it was fully assembled for testing at an IBM site. Then, at the June 2008 International Supercomputing Conference in Dresden, Germany, Roadrunner was named the world's fastest supercomputer by the TOP500 (www.top500.org) organization, which tracks the fastest supercomputers in the world. The machine was delivered and installed at LANL in the fall of 2008. At a slightly augmented size and with a speed of 1.105 petaflop/s, Roadrunner retained its number one ranking as the world's fastest supercomputer according to the TOP500 announcement at the November 2008 Supercomputing Conference, just edging out a strong competitor named Jaguar located at ORNL. A year later, an upgraded version of the Jaguar machine claimed the number one position. During this same time period Roadrunner has also been a leader in energy efficiency for supercomputers and has been near the top of the Green500 (www.green500.org), which ranks computers by their energy efficiency (flops/watt). Only other Cell-based computers have proven more energy efficient than Roadrunner.

Roadrunner is also the world's largest supercomputer employing a hybrid architecture of commodity server nodes with attached acceleration devices. The secret to Roadrunner's record-breaking performance is the use of 3,060 compute nodes that each consists of two AMD Opteron™ dual-core processors plus four IBM PowerXCell 8i™ processors used as computational accelerators. These node-attached Cell accelerators are what make Roadrunner different than typical clusters. We call Roadrunner a Cell-accelerated hybrid supercomputer.

The Roadrunner architecture

The PowerXCell 8i chip used in Roadrunner is a slightly modified version of the Cell processor chip used in the Sony PlayStation 3®. It is very similar to the original Cell chip except that its floating point units can do double precision arithmetic at much greater speed, and its memory subsystem uses conventional DDR2 DRAM (dynamic random access memory). Without these changes, Roadrunner could not have reached its supercomputing goals. The Cell chips are actually hybrid multicore processors themselves, and contain one PowerPC processing element (PPE) and eight special compute cores known as Synergistic Processing Elements, or

SPEs, as shown in Fig. 1. Figure 2 shows how the Cell chips fit into the compute nodes. They have a highly optimized and unique way of processing that achieves great performance in a single chip with great power efficiency. New programming techniques are required, and modified codes and algorithms are needed to effectively run on the Cell processors with good efficiency. Fortunately the Cell comes with a fairly complete programming environment, so most of the effort is in recasting and reprogramming algorithms. Although the Cell design is several years old, it provides performance levels well beyond other chips of that time and just now is being matched by the newest multicore microprocessors.



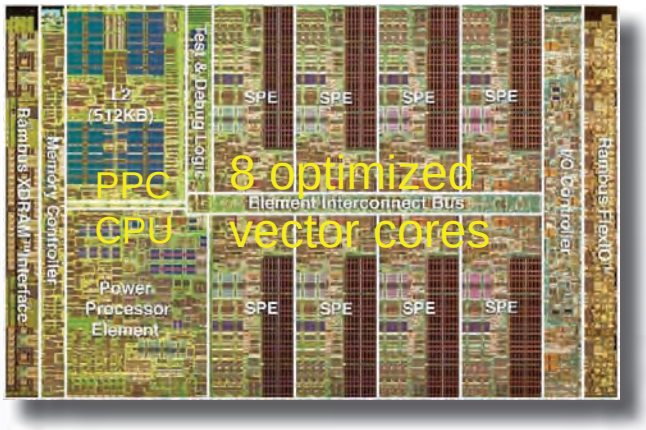
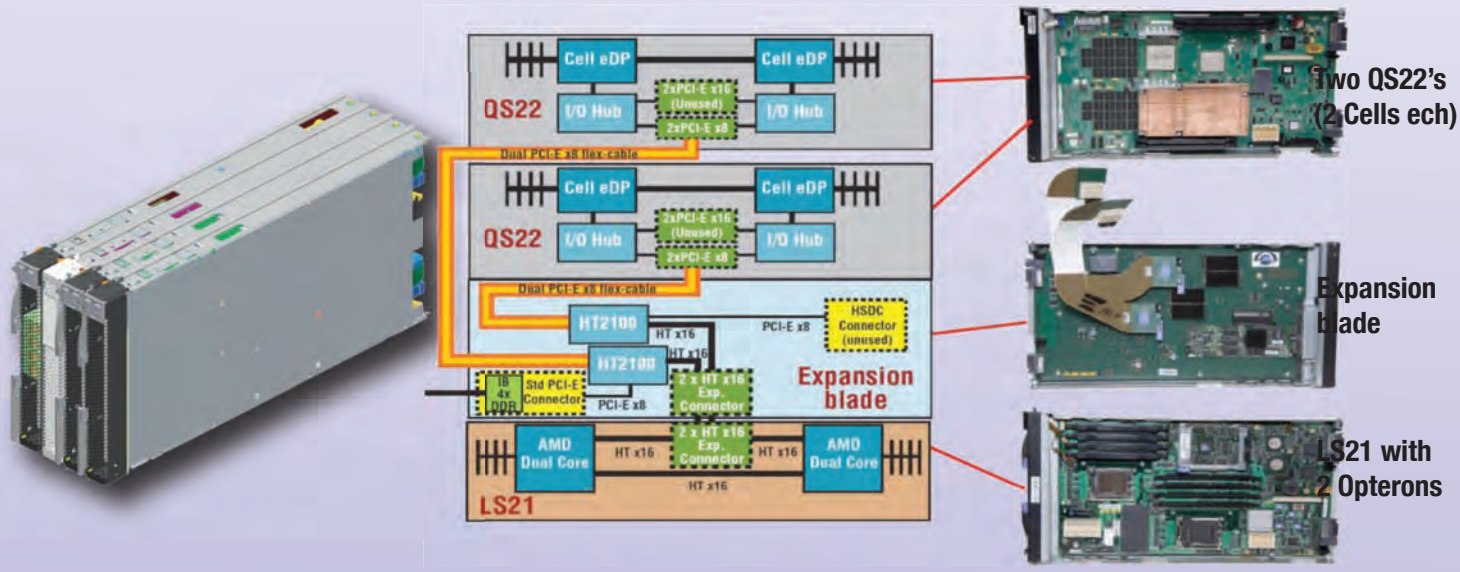


Fig. 1. Eight high-performance Synergistic Processing Elements (SPEs) plus one Power Processing Element (PPE) are in a Cell processor. The SPEs are highly tuned SIMD cores containing a processing unit (SXU), a local 256 KB scratchpad memory (LS), and a DMA memory engine (SMF) to move data into and out of the local store from the large shared off-chip DRAM-based memory. The PowerXCell 8i version of this chip can perform at 102.4 GF/s using all 8 SPEs simultaneously and has an off-chip memory bandwidth of 25.6 GB/s.

Building Roadrunner from Opterons and PowerXCell 8i processors was an engineering feat that leveraged two existing IBM compute server blade products. One LS21 Opteron blade is connected to two QS22 Cell blades using a Roadrunner-custom expansion blade to form a Cell-accelerated compute node, called a Triblade due to its three active processor blades.

The Triblade achieves a design goal of having one Cell processor for each Opteron core and also provides 16 GB (4 x 4 GB) of shared memory for the four Opteron cores plus 16 GB (4 x 4 GB) for the Cells. IBM and LANL collaborated to create custom drivers, a data transfer protocol, and the Data Communications and Synchronization (DaCS) runtime library to move data between the Cells and Opterons, and to launch, control, and monitor processes running on the Cells from the Opterons. The DaCS application programming interface (API) was designed and implemented to have both direct memory access, or DMA, like interfaces similar to those in the Cell's SPEs, and message passing interfaces similar to MPI; data transfers can be initiated from either side.
The full Roadrunner configuration uses these Cell-accelerated Triblades,

Fig. 2. A Roadrunner compute node comprises one LS21 and two QS22 server blades interconnected by a special expansion blade. The expansion blade connects the four Cell chips on the QS22 blades with four independent PCIe x8 links to the two Opteron chips on the LS21 blade with two independent HyperTransport links. This was the highest performing and most direct connection possible using the available IBM blade-based building block.



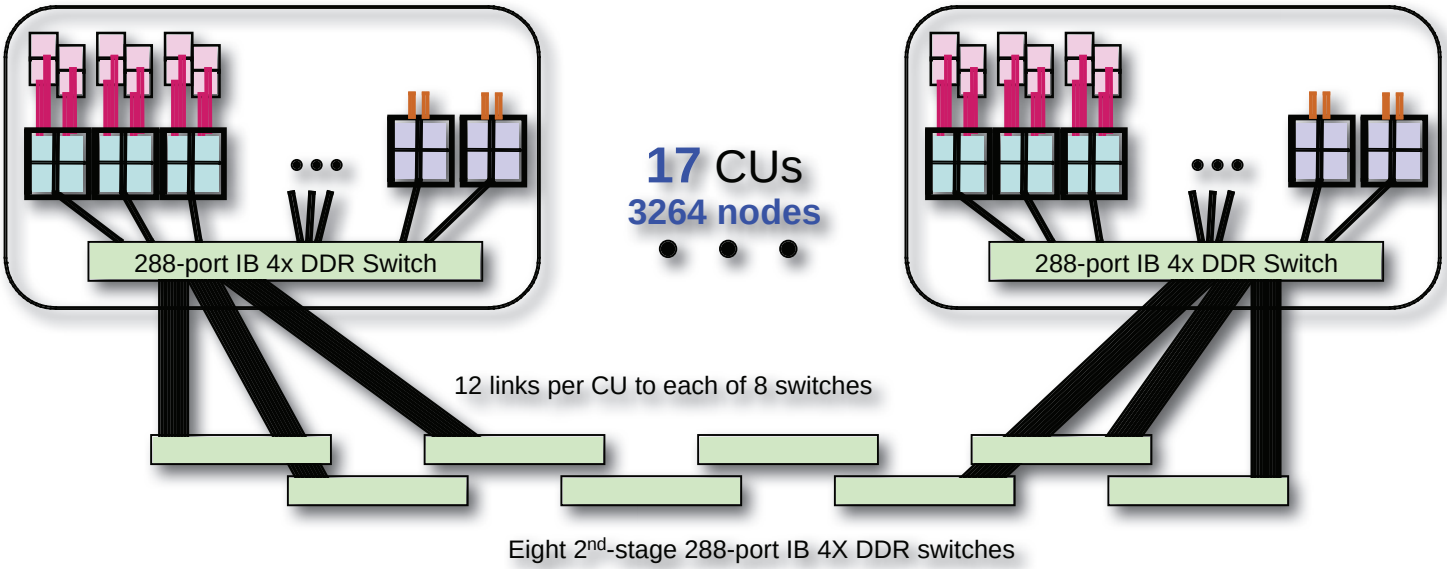


Fig. 3. Roadrunner is composed of 17 Connected Unit groupings, each comprising 180 Triblade compute nodes plus 12 I/O gateway nodes. All nodes in a Connected Unit are connected to a single 288-port InfiniBand 4x DDR switch, and those 17 switches are uniformly interconnected at a second level using eight more of the large switches to form a 2:1 oversubscribed fat tree interconnect. A total of 3,060 Triblade nodes contain the 12,240 PowerXCell 8i chips in the machine and have 1.33 Petaflop/s of double precision performance, while the 6,120 Opteron chips contribute only an additional 44 Teraflops/s.

organized into 17 subclusters of 180 Triblades each with accompanying I/O gateways nodes, to provide off-machine connectivity to a Panasas parallel file system. The overall Roadrunner configuration is illustrated in Fig. 3. The full machine is composed of 3,060 Triblades containing 12,240 PowerXCell 8i chips and 6,120 dual-core Opterons. The overall performance from just the SPEs in the Cell chips is 1.33 petaflop/s, while there is only 44 teraflop/s (~3%) of performance in the Opterons.

Programming Roadrunner

Writing a code to run on Roadrunner requires three interoperating programs, typically created by breaking an existing program into an Opteron piece, a Cell PPE piece, and a Cell SPE piece, and then adding the control logic and API calls to exchange data as needed to make them work together.Each Opteron process is part of a typical MPI-based cluster application and, through the DaCS library, spawns a Cell PPE process, which then goes on to spawn SPE processes.

Since almost all of the performance of Roadrunner is in the Cell SPEs, the most important step is to get the compute-intensive portion(s) of an application running in parallel on the eight SPEs of each Cell chip. This is where the most program restructuring is potentially needed to: 1) present threaded parallelism to eight SPEs in 128-bit SIMD form, and 2) tile or block the algorithm into work chunks that fit and can be used within the small 256 KB scratchpad local stores of each SPE. The SPE DMA engines are called via simple APIs in the SPE code to retrieve work chunks and send them back to main Cell memory after computing on them. Optimal performance is achieved by simultaneously having the next work chunk prefetched and the previous one written back while the SPU computes on a current work chunk.

The Cell PPE process is launched by its Opteron host process via DaCS. It sets up and synchronizes the eight SPEs and works with the Opteron process to exchange data between the Cell and Opteron memories as needed. Typically the PPE code will control the overall advancing of an algorithm's steps or cycles, and in doing so, synchronize the SPEs doing the actual computations. Since the PPE is a stripped-down PowerPC, algorithm and code developers should target as little work as possible for the PPE.

A standard MPI environment is used to launch and run the first part of the code that is distributed on the Opteron cluster, and is the heart of every Roadrunner run. Typical application functions like setup, cycle monitoring, and checkpoint/restart run on the Opterons with only minimal code changes to send data to or retrieve data from the Cells. The Opteron process uses DaCS to start a coupled Cell PPE code. The PPE and the Opteron coordinate and process data exchanges as needed between Cell and Opteron memories.

Algorithms requiring data exchange with neighboring cluster nodes do so only through the Opterons and usually create an MPI-relay form of messaging to do so via the following four steps (see Fig.4): 1) the Cell PPE first sends data to the Opteron process; 2) the Opteron then forwards the data via an MPI send operation to the appropriate neighboring Opteron rank; 3) the Opteron also receives data from other neighboring Opteron ranks using MPI receives; 4) and finally, the Opteron or Cell PPE sends the data up to the Cell.

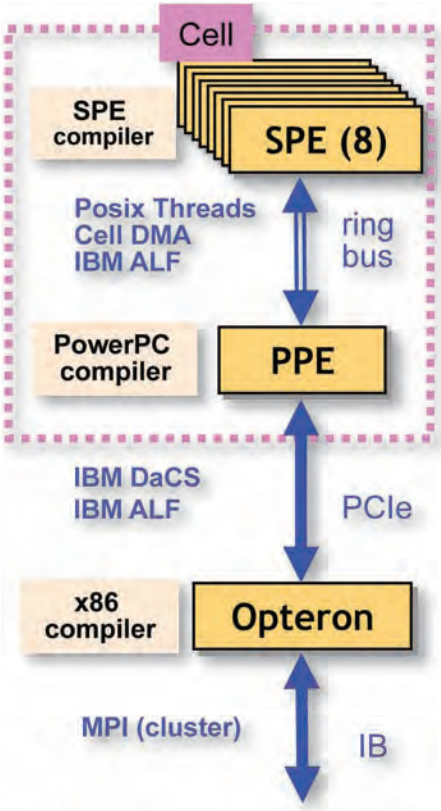
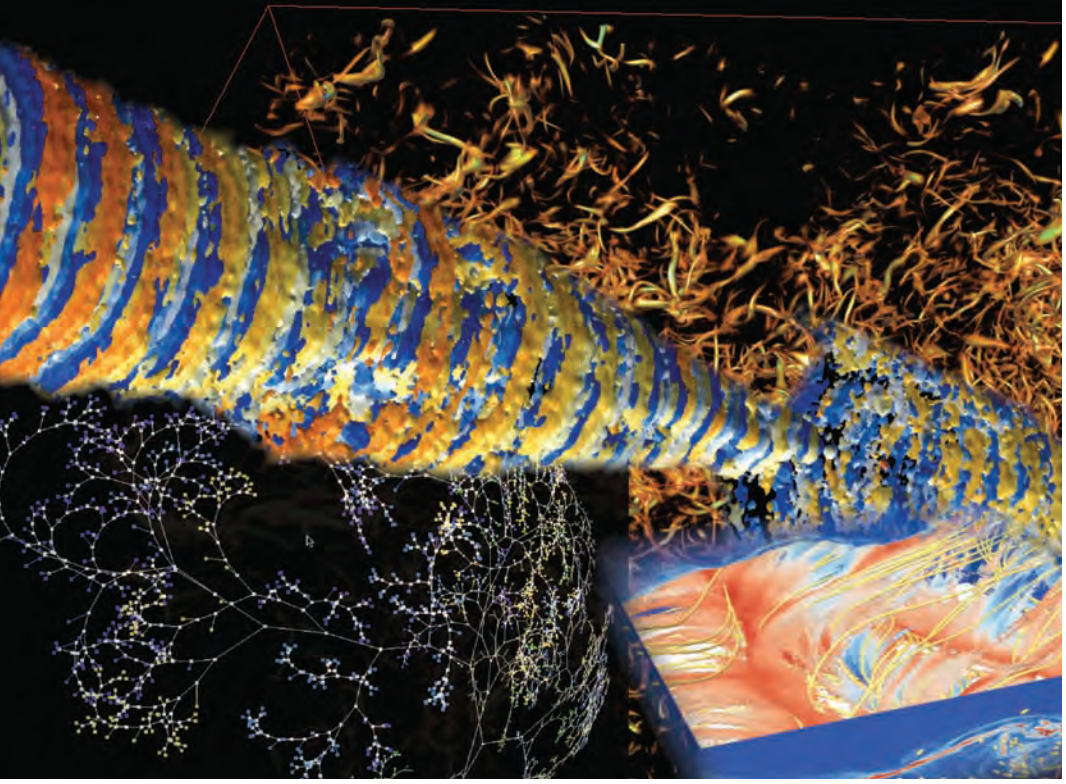


Fig. 4. Codes for Roadrunner are built as three cooperating pieces running on the different cores in a Triblade. The programming model uses MPI across the cluster, DaCS for Opteron-to-Cell communications, and SPE threads with memory DMAs in and out of their local stores.



The Roadrunner Coding Effort

Jamaludin Mohd-Yusof, Applications Team Member

The Roadrunner architecture presented some unique challenges to application programmers, the solutions to which in algorithmic refactoring and data locality are applicable to hybrid and many-core architectures of the future.

The codes selected for the Open Science projects represent a variety of types. Several contain particle-based segments, including VPIC (particle-in-cell), SPaSM (molecular dynamics), IMC (Implicit Monte Carlo), and RRU (Roadrunner Universe Project). Only one selected code, CFDNS, is a primarily grid-based code. Another, the Human Immunodeficiency Virus (HIV) phylogenetics code, requires the evaluation of tree structures using a Markov chain-branching model. The good use of the coarse hybrid aspect of the Roadrunner system is an issue that has yielded different engineering approaches, as well as different algorithms. Some codes are feasible and appropriate for centering on Cell processors (termed "Cell-centric"), thereby using the Opterons simply to handle data movement across the cluster. Other developers worked with very complex legacy codes that had to move in a more incremental fashion to Cell acceleration. Nevertheless, the basic considerations involved in successful porting of the codes are those that should apply to the development of any hybrid application, and, to a large extent, to hybrid architectures in the near future.

From a coding perspective, the Roadrunner hybrid architecture presents some unique challenges. Most obvious is the need to write three separate, but coupled, codes for the Opteron, PPE, and SPE, each with its own programming paradigms and languages (see previous article for an overview of the hardware). Due to the exacting data-alignment requirements on the PPE and SPE, most of the Open Science codes utilize C or C++ (with appropriate processor-specific extensions) for the Cell portions. Some codes have retained Fortran for the Opteron code, whereas others were originally C/C++. In the case of the DNS code, the entire code was rewritten in C, because the modification of data structures and algorithms needed for optimal Cell performance meant that virtually none of the original code would have been retained, regardless of the language choice. The choice of protocols and languages (as well as diagnostic tools) available to programmers constantly evolved throughout the project. Since most of the codes used in the Open Science runs were derived from the earlier evaluation codes, they mostly retained the earlier protocols (e.g., DaCS and libspe) rather than the later ones (e.g., ALF).

Less obvious are the limitations imposed by the SPE, most notably poor branching performance. Although the need to explicitly manage data movement between Cell main memory and local store is a significant departure from "normal" programming models, it is, along with the excellent memory performance of the Cell architecture, responsible for the speedup that is observed even in the absence of significant arithmetic intensity within kernels. By and large, the transition to

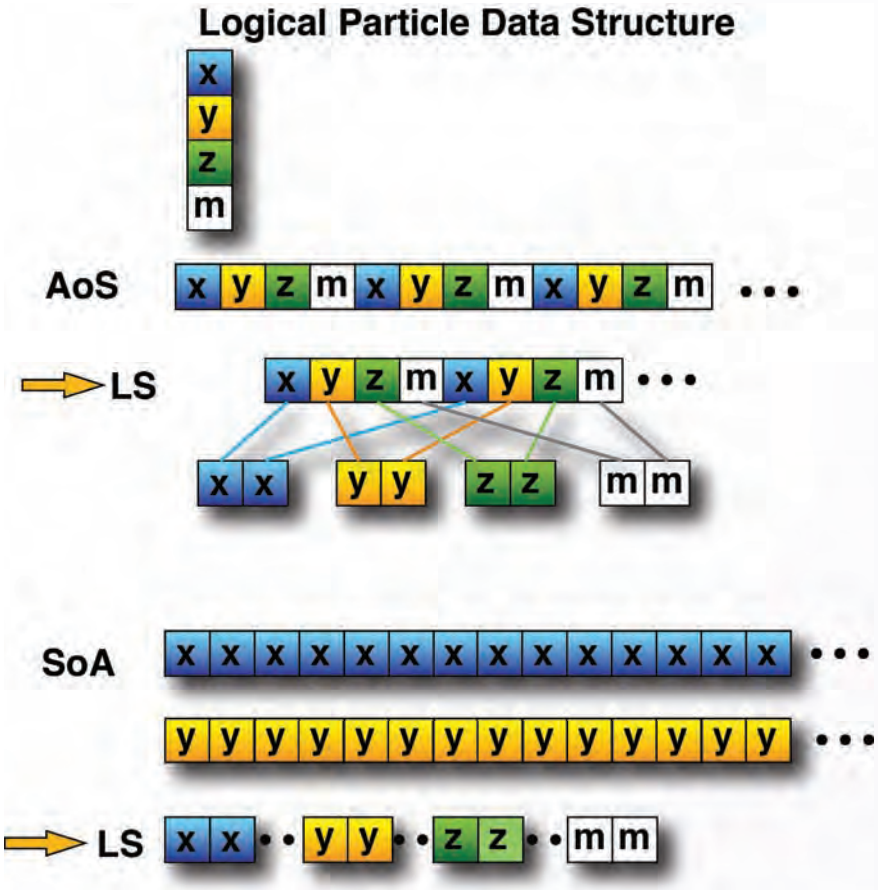


Fig. 1. Comparison of Array of Structures (AoS) and Structure of Arrays (SoA) data layouts. If the data is stored as AoS, then values must be shuffled in local store to allow vector math operations. If the data is stored as SoA, then no reshuffling is required and the computation is significantly faster.

explicit direct memory access (DMA) management has been straightforward. In general, the PPE was shown to be a weak performer, so most codes minimized its contribution, except where the PCIe bus speed made it necessary – that is, there are cases (e.g., RRU) where having the PPE perform computations is faster than moving the data back to the Opteron, where the same operations would have been quicker in isolation.

One consequence of the speedup provided by the Cell is the exposure of the effects of data movement and latency, which are inherent in all the non-computation portions of any code. These include not just the usual inter-node communications overhead, but also the time needed to move data between Opteron and Cell main memory (across the PCIe bus), and the overhead associated with kernel invocation on the Cell. The last means that for a small work block, the Cell kernel cannot be viewed as simply calling a subroutine from the main program, since the kernel invocation overhead will dominate the total execution time. Rather, most of the codes invoke a kernel once on the Cell and then use signals among the three levels for managing the progression of the code. This extra level of data movement is also a factor when inter-node communication is present, because it introduces extra latency. Although some users developed programming abstractions to mask the complexity of these data movements (allowing "direct" messaging between Cells), the topology of the machine is such that the extra lag cannot be avoided.

We present here an overview of the important features shared by the code development teams, with examples drawn from each of the three code types: particle, grid-based, and tree-based Markov chain. We also provide a prescription for porting future applications to similar accelerated architectures, including graphical processing unit (GPU)-enhanced clusters.

Identification of Kernels to be Ported

Typically, the identification of kernels to be ported involves identifying "hotspots" in the code, where significant computational time is spent during a simulation. Because the PCIe bus is a potential bottleneck, the data movement between kernels is also a factor in the overall computational time.

For the particle-based methods, there is typically a compute-intensive ($O(N^2)$) local particle-particle interaction kernel that accounts for a large fraction of the computational work. These particles usually interact with underlying fields in the mean sense only, and thus these kernels are ideal for movement to the Cell broadband engine (CBE). Once the particles have been moved to Cell main memory, there is no need to move the data back to the Opterons – except when they need to move out of the local domain. The underlying mesh fields and the accumulated particle contributions to those fields may need to be communicated to the Opteron at every time step, which places a practical lower bound on the number of local particles that are required to make the Cell acceleration practical.

For the CFDNS code, the large data transfer between kernels (entire local blocks) and the relatively uniform distribution of work across kernels indicates that the entire time step should be moved to the CBE, leaving only I/O and inter-node communication on the Opteron. In this case, the field variables once again need to be moved onto the CBEs at initialization time, and then only "halo" or ghost-cell data need be moved

during the course of the simulation. For the phylogenetics code, initial profiling indicates that the main kernel consumed 90% of the CPU time. The code uses a master/slave protocol, whereby a single processor generates candidate histories and all others evaluate them. This system works at large scales (hundreds of processors) because it is very cheap to generate a candidate history and very expensive to evaluate them. At this stage, the team did not require acceleration of the individual tree evaluation speed (e.g., across processors), but rather acceleration of an aggregated evaluation speed across millions and billions of tree evaluations.

Modification of Data Structures

The CBE is a (short) vector processor that operates on 128-bit words. Experience has shown that the number of DMA requests affects the performance (a few large DMAs execute faster than many small ones for the same total data movement). As such, the underlying data structures must be amenable to large coalesced loads and vectorizing without data reorganization on the fly. While the original CFDNS code used an array-of-structures (AoS) data layout, the Roadrunner DNS (CFDNS) code uses structure-of-arrays (SoA) to accomplish this vectorizing. Some of the particle-based codes (e.g., RRU) also use this type of data transformation. This method allows the SPEs to operate on the data without the need to shuffle data within the local store (see Fig. 1). As a result, DNS and RRU have achieved acceleration of the local kernels by factors on the order of 100x.

In contrast, some particle-based schemes retained the AoS data layout in order to minimize the amount of modification required to the original code. In these cases, they use the excellent shuffle intrinsics of the SPE, although there is still a performance penalty compared to fully vectorized code. Speedups of the local kernel in these cases appear to be on the order of 10x. SPaSM initially adopted a hybrid scheme, whereby the particle data was converted from AoS to SoA by the Opteron, prior to being sent to the Cell, thus allowing more of the original Opteron code to be retained. A later version of the code resulted in a full redesign of the data structures.

Redesign of the Local Kernel

In addition to being a short vector processor, the SPE has other idiosyncrasies that have occasionally mandated modifications to the local algorithms. In particular, the SPE has poor branching performance, which affects the design of the particle-based kernels. Typically, these kernels involve a series of checks to determine if the particle has passed out of the grid cell, and a branch is then executed based on the result. Since a failed conditional check results in an 18-cycle stall, these operations can become extremely inefficient if executed on single particles. Instead, most of the particle codes have adopted strategies where all particles are treated as if they remain within a grid cell (and within the local domain) and then outliers are flagged and treated in a separate pass. Similarly, the SPaSM local kernel does not attempt to enforce reciprocity during the force calculation, since this complicates the data access patterns. This action results in doubling the number of flops, but has no effect on the total data movement, which is a more significant factor in the overall execution time.

In the IMC computation, for example, the computation of the intersection of a particle trajectory with a cell face has been completely recast to remove the conditional. Instead, the intersection of the particle path with all six faces of the cell is computed and then masking operations are applied to remove the negative values. Finally, the minimum of the remaining values is chosen. This procedure is significantly faster than the original algorithm, despite

requiring six times the computational effort. The speedup comes from the nature of the instructions–the five discarded distance calculations can be pipelined, whereas the (possibly) five missed branches cannot.

In contrast, the grid-based code rarely requires conditional statements except at boundaries. In this case, the implementation of boundary conditions was performed by multiplying the field values by masking stencils that were determined and set by the Opteron at startup. Thus, the CBEs executed some extra flops, but these are cheaper than failed branch checks. The only other conditional statements are those required for the time-stepping scheme. In that case, the logic is performed on the Opteron, which is natural since the time-step restriction is dependent on global values of the error, and these are only available on the Opteron. Additionally, the SPEs perform conditional checks on state flags set by the Opterons, but these are outside large work blocks so that the cost of the 18-cycle stall is amortized over several thousand or million cycles, depending on the local block size.

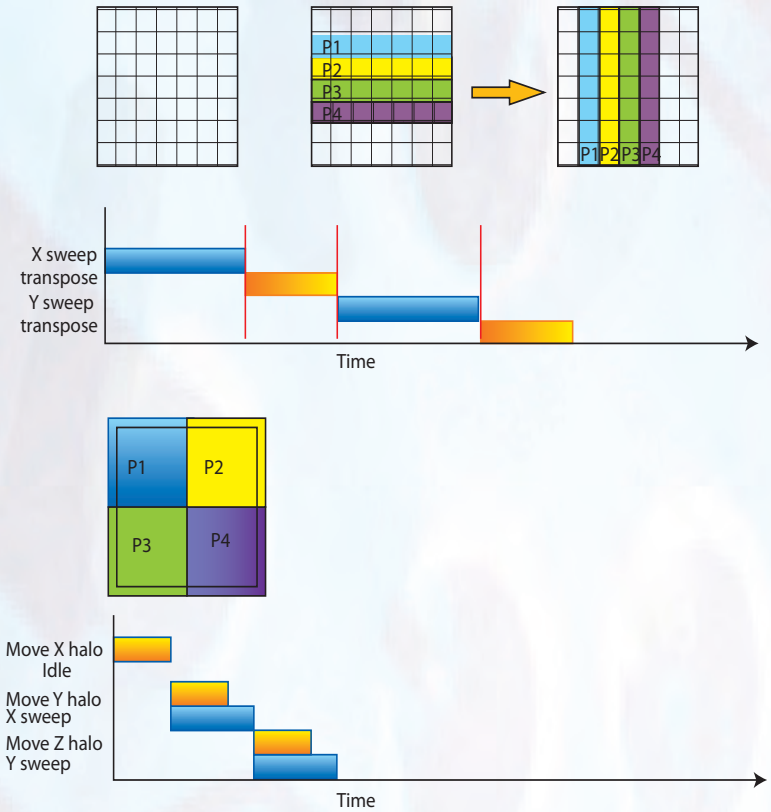


Fig. 2. Data movement patterns for the original and distributed tridiagonal solver. In the original case, the data must be transposed after the x sweep, moving the entire N^3 local data block, and the processors must wait for data. In the distributed solver, only the 'halo' cells ($2xN^2$) need to be moved, and the x sweep can proceed while y halo cells are being moved. Due to reduced data motion, the distributed solver has proved to be faster than the original for large processor counts even on unaccelerated architectures.

Evaluation of Serial Function Performance

After the kernels are ported to the CBE platform, the initial performance is measured in order to assess what other modifications are required.

With DNS, these tests showed that serial performance is improved by 30x compared to the reference Opteron code. This predicated several changes in the overall code execution pattern compared to the original code.

In the case of the particle codes, the kernels ported to the Cell typically showed at least an order-of-magnitude improvement in speed. Those that opted for redesign of the data structures could often obtain a 100x speedup of the kernel in isolation. The overall speedup of the codes in this case was determined largely by the ability to successfully hide the latency of data movement between memory subsystems, as well as the granularity of the work blocks assigned to the Cell.

Modification of Code Execution Pattern

Once the serial speedup has been verified, the overall parallel code pattern is evaluated and modified, if needed. There are two main factors that guide this stage.

First is the need to hide communication overhead by overlapping local computation with inter-node data movement, a problem that is exacerbated by the serial speedup that is exhibited. Second is the desire to reduce the number of synchronization points between the host and the accelerator, since this is a tedious activity and a source of potential error.

In the case of the particle codes, the modification of the communication patterns was usually limited. For the most part, this has involved the change outlined in the previous section, where the treatment of particles that move off-processor (out of the local domain) was done in a separate sweep. Generally, these are aggregated in a separate array and moved at a single intermediate stage, rather than being handled immediately (as they were in the original codes). In the case of SPaSM, the original computation pattern resulted in many small messages being passed between ranks, which proved to be a prohibitive bottleneck. In contrast, the Roadrunner (RR) version of the code passed a set of halo particles between

ranks, after which the entire local domain could be computed without further inter-node communication. In the CFDNS code, inter-node communication is only required during the computation of derivatives, which use a Pade scheme and require a tridiagonal inversion in the direction in which the derivative is taken. In the original code, derivatives of any given field are requested for computation of the update equation as needed. The original code also uses a traditional tridiagonal solver, which requires data transposes to move the entire line of data onto a single processor. In contrast, the RR code accumulates the derivatives into blocks that are directionally split, and uses a novel distributed tridiagonal solver (see Mohd-Yusof et al., 2009). This allows the local x-direction computation to proceed while y-direction data is being moved, and so on (see Fig. 2). In this case, the tradeoff is made between a single large data movement (all-to-all transpose, $O(N^3)$) at the beginning and end of the derivative, and several much smaller (although still relatively large, $O(N^2)$) data transfers throughout the derivative. The accumulation of the derivatives into blocks also serves to reduce the total number of required data transfers.

Benchmarking of Parallel Code

The final measure of success for all these projects is the ability to achieve sufficient speedup to justify the work invested, which is considerable. A secondary benefit of the work done to prepare the codes for Roadrunner is that many of the modifications could also be implemented in the original, unaccelerated codes, with good results. Furthermore, these modifications make the codes more amenable to acceleration on other current and upcoming platforms, many of which have generally similar requirements for data alignment. As an example, the data reorganization that makes a code amenable to SPE vectorizing improves cache performance and will also allow it to take advantage of streaming SIMD extensions vectorizing on conventional architectures, which can result in a 2x speedup. After all the modifications, the parallel RR DNS code at scale is approximately 25x faster than the original. If the modifications to the tridiagonal solver had not been implemented, the data movement required for derivative computation would have limited the performance to only about a 4x improvement over the original.

The particle-based codes showed speedup in the range of 10x, largely due to the retention of significant portions of the code on the Opteron. The phylogenetics code has also achieved a similar speedup.

Strategies for Evaluating Future Codes

Based on the experiences encountered in this exercise, we offer the following prescription for porting other codes to hybrid architectures. Foremost is identification of the kernels to be ported, based on the fraction of execution time in the original code and the amount of data movement required. Once these have been identified, it is safe to assume at least an order of magnitude speedup of the kernel execution time alone. The total serial wall-clock time improvement is then estimated based on the ability of the programmer to hide the data transfer and kernel invocation times by efficiently streaming work to the accelerator. At this stage, the serial speedup may inform the need to redesign the execution pattern of the parallel code in order to avoid poor scaling due to inter-node communication bottlenecks. This can be done either by overlapping the inter-node communication with local compute, or by redesigning the algorithms to require less data movement. Note that for some applications, the acceleration of local work will lead to the execution time being dominated by inter-node data movement, which will lead to poor scaling behavior, albeit with vastly reduced total execution time.

References
Germann, T.C., K. Kadau, and S. Swaminarayan, *Concurrency Comput. Pract. Ex.* **21,** 2143–2159 (2009).
Mohd-Yusof, J., D. Livescu, and T.M. Kelley, *Proceedings, Parallel Computational Fluid Dynamics 2009,* Moffett Field, CA, May 18-22 (2009).
Pope, A., S.Habib, Z. Lukic, D. Daniel, P. Fasel, K. Heitmann, and N. Desai, *Comput. Sci. Eng.* **12,** 17-25 (2010).

Understanding the nonlinear physics of laser-plasma interaction through "at scale" plasma kinetic simulations on Roadrunner

Lin Yin[1], Brian J. Albright[1], Kevin J. Bowers[2], and Ben Bergen[1]

[1] Los Alamos National Laboratory, [2] D.E. Shaw and Company

In 2010, inertial confinement fusion (ICF) experiments commence at the National Ignition Facility (NIF). In these, over a million Joules of laser energy are focused within a gas-filled hohlraum.

Pictured left to right: Ben Bergen, Kevin Bowers, and Lin Yin.

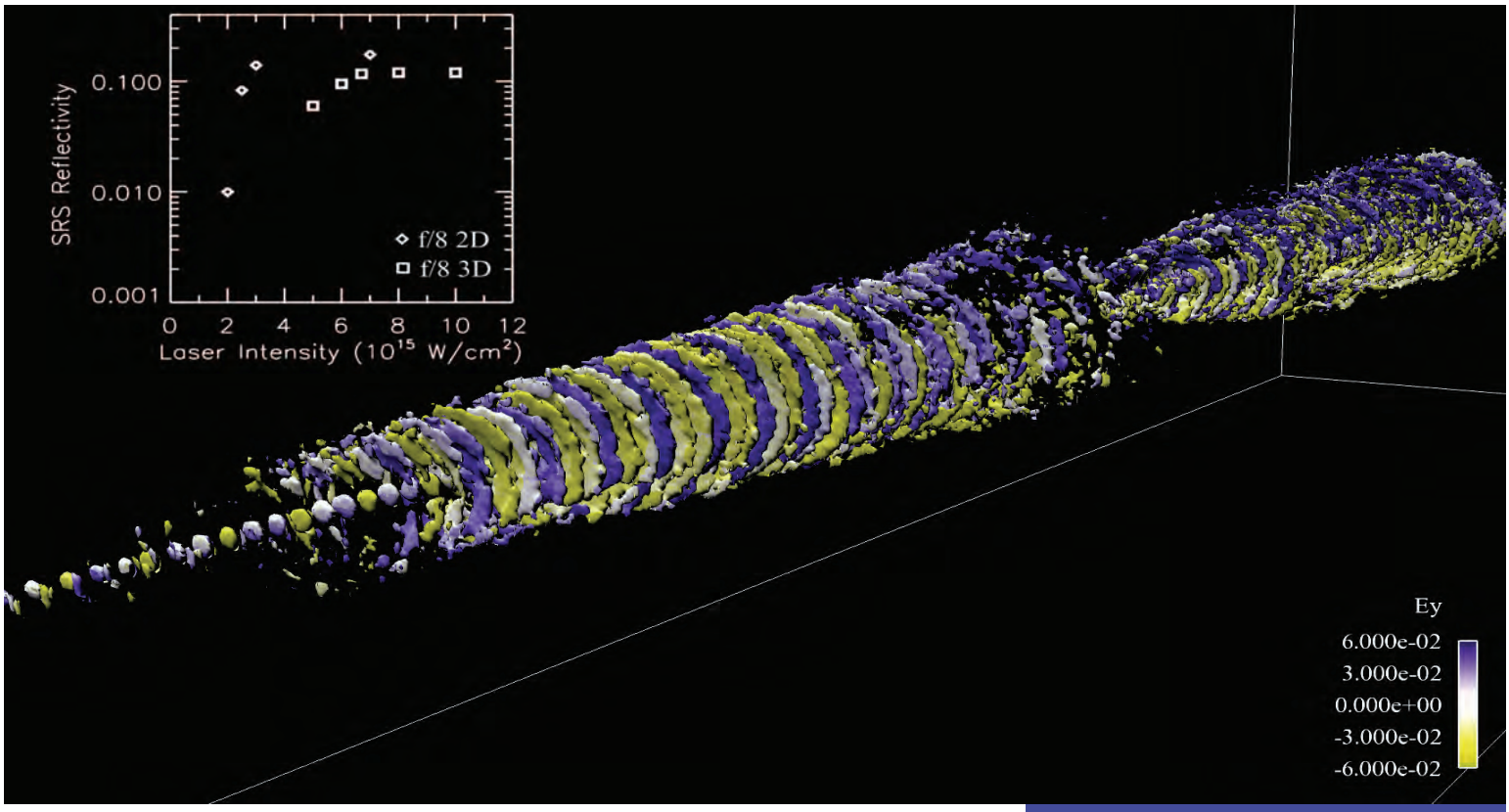
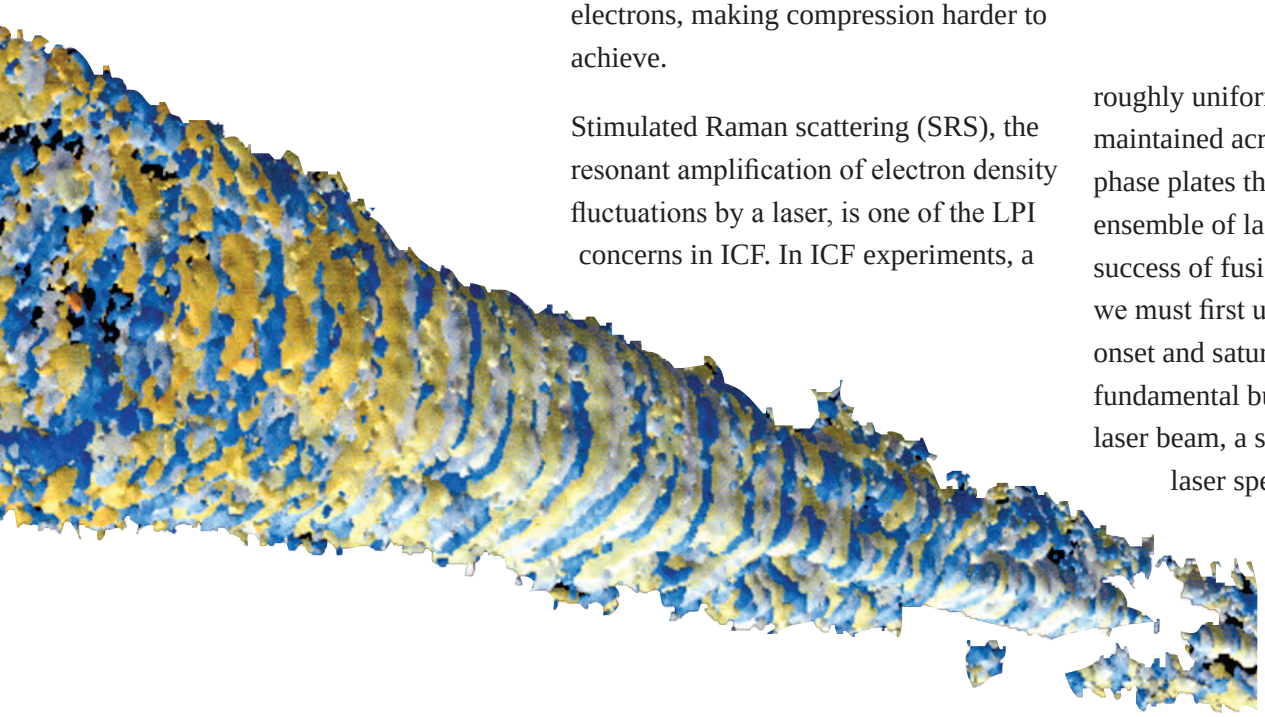


Fig. 1. Single-speckle LPI calculations using 16 CU of Roadrunner (11,520 ranks), nearly the full system; this calculation employed a record 0.4 trillion particles, >2 billion cells, and ran for nearly 58,160 time steps (~10^{19} floating point operations), long enough for two bursts of stimulated Raman scattering to grow from noise to significant amplitude at a laser intensity near the SRS.

The hohlraum walls absorb the laser energy and re-radiate it as X-rays, which absorb in a spherical capsule at the hohlraum center. This causes the capsule to compress, bringing the deuterium-tritium fuel to the high temperatures and pressures required for thermonuclear fusion.

To prevent the hohlraum walls from ablating during the ~10^{-8}-s laser drive, a fill gas of hydrogen or helium is used. As the laser propagates through the fill gas, laser-plasma instabilities (LPI) may arise, which scatter laser light out of the hohlraum, degrade capsule implosion symmetry, and preheat the fuel with hot electrons, making compression harder to achieve.

Stimulated Raman scattering (SRS), the resonant amplification of electron density fluctuations by a laser, is one of the LPI concerns in ICF. In ICF experiments, a roughly uniform laser intensity is maintained across the beam with random phase plates that break the beam into an ensemble of laser speckles. For the success of fusion experiments on the NIF, we must first understand the physics of onset and saturation of SRS in the fundamental building block of a NIF laser beam, a single laser speckle. In a laser speckle, SRS manifests as the amplification of a forward-directed electron plasma wave (EPW) and the backward scattering of



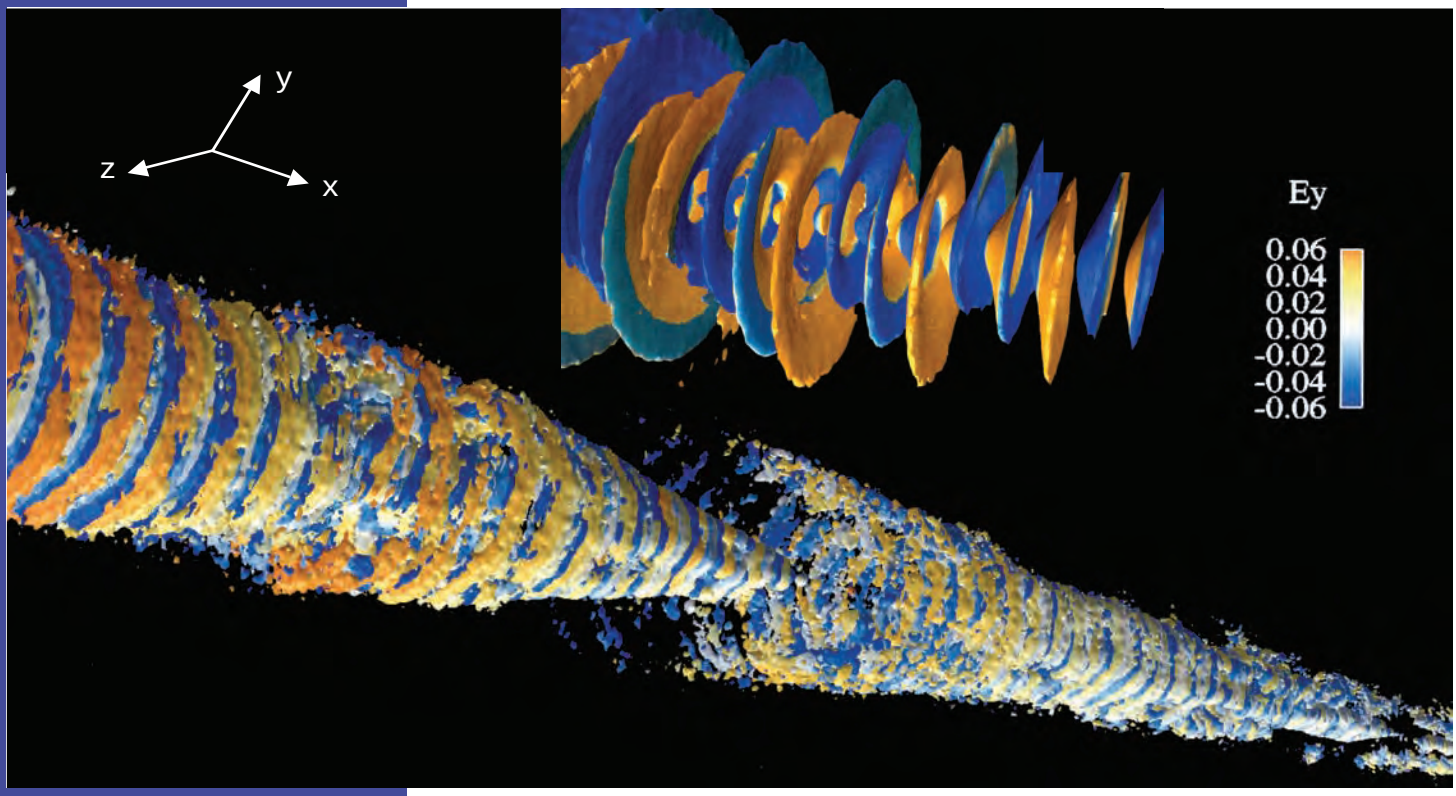


Fig. 2. (Top) A snapshot taken from a 3D VPIC LPI simulation at SRS pulse saturation of a f/4 laser beam, showing bending of iso-surfaces of EPW electric field across the speckle. The iso-surfaces are colored by the laser electric field, graphically showing the source for SRS backscatter has become incoherent. (Bottom) Self-focusing and filamentation in two bursts of SRS, a snapshot taken from a 3D VPIC LPI simulation of a f/8 laser beam. The speckle volume is 16x larger than the f/4 simulation, permitting more transverse self-focusing modes to develop. This leads to chaotic EPW phase variation across the speckle. This further reduces SRS source coherence and increases wave damping, quenching the SRS pulse.

laser light. Unlike the linear growth of SRS, the nonlinear physics was not well understood until recently [1]. Roadrunner has been used to assess the impact of the nonlinear SRS physics on laser penetration and energy deposition in fusion experiments. These fully kinetic plasma simulations employ the VPIC particle-in-cell code [2] and are performed in large plasma volumes in 3D at an unprecedented range of time and space scales.

Until recently it has not been possible to fully simulate the comparatively large 3D plasma volumes of laser speckles. With VPIC on Roadrunner, simulations of the NIF holhraum plasma have been done using 4096 Cell chips at a range of laser intensity values (see the graph in Fig. 1). These simulations [3] showed that SRS reflectivity within a solitary speckle exhibits nonlinear behavior: a sharp onset at a threshold intensity, whereby reflectivity increases abruptly to a level orders of magnitude higher than linear theory predicts over a small range of intensity, with a plateau in reflectivity at higher laser intensity in which SRS nonlinearly saturates. This generic behavior matches that measured in single-speckle experiments at the LANL Trident Laser facility [4] with physics that cannot be captured by linear gain models of SRS growth within the speckle.

As a highlight of the unique simulations afforded by Roadrunner, the largest of these calculations was run on 16 CUs using 11,520 Cell chips and MPI, message passing interface, ranks, nearly the full Roadrunner system, and employed a record 0.4 trillion particles, over 2 billion computational cells, and ran for nearly 58,160 time steps (~10^{19} flops), long enough for two bursts of stimulated Raman scattering to grow from noise to significant amplitude at a laser intensity near the SRS onset.

Figure 1 shows isosurfaces of electrostatic field associated with these bursts; the wave fronts exhibit bending or

"bowing," arising from nonlinear electron trapping, as well as self-focusing, which breaks up the phase fronts. The essential nonlinear physics governing SRS saturation has now been identified. The scattering manifests as a series of pulses, each of which passes through four distinct phases:

1) SRS grows linearly from density fluctuations.

2) Electrons trapped by the EPW reduce the wave frequency and phase velocity by an amount that scales with EPW amplitude.

3) Near the speckle center, where the amplitude is highest, the EPW phase velocity is lower than at the speckle's edge; EPW phase front bending ensues as shown by the top image in Fig. 2.

4) The EPW wave amplitude exceeds the electron trapped particle modulation instability (TPMI) threshold [5]. TPMI generates waves off-axis from the laser direction and leads to EPW filamentation, self-focusing and phase front breakup, shown by the bottom image in Fig. 2. Self-focusing increases the transverse loss of trapped electrons and increases EPW damping.

From these basic science simulations, researchers are now able to better understand the essential nature of LPI nonlinear onset and saturation. Current research focuses on determining whether neighboring speckles can interact via exchange of hot electrons or waves to produce higher backscatter than they would individually, the kind of study only possible on Roadrunner, where "at scale" kinetic simulations of laser-plasma interaction in 3D at realistic laser-speckle and multispeckle scales can be prosecuted at unprecedented size, speed, and fidelity.

References

[1] Yin, L., B.J. Albright, K.J. Bowers, W. Daughton, and H.A. Rose, *Phys. Rev. Lett.* **99,** 265004 (2007); *Phys. Plasmas* **15,** 013109 (2008).
[2] Bowers, K.J., B.J. Albright, L. Yin, B. Bergen, and T.J.T. Kwan, *Phys. Plasmas* **15,** 055703 (2008); *Proc. ACM/IEEE Conference on Supercomputing,* Austin, 2008 (Piscataway, NJ, USA: IEEE Press. 1-11 (2008); *Proc. J. Phys. Conf. Ser.* **180,** 012055 (2009).
[3] Yin, L., B.J. Albright, H.A. Rose, K.J. Bowers, B. Bergen, D.S. Montgomery, J.L. Kline, J.C. Fernandez, *Phys. Plasmas* **16,** 113101 (2009).
[4] Montgomery, D.S., J.A. Cobble, J.C. Fernandez, R.J. Focia, R.P. Johnson, N. Renard-LeGalloudec, H.A. Rose, and D.A. Russell, *Phys. Plasmas,* **9,** 2311 (2002).

VPIC—Magnetic reconnection

William Daughton[1], Vadim Roytershteyn[1], Lin Yin[1], Brian Albright[1], Ben Bergen[1], and Kevin Bowers[2]

[1] Los Alamos National Laboratory, [2] D.E. Shaw and Company

Magnetic reconnection is a basic plasma process involving the rapid conversion of magnetic field energy into various forms of plasma kinetic energy, including high-speed flows, thermal heating, and highly energetic particles. This process is usually associated with changes in magnetic topology, giving rise to magnetic islands or more complicated flux rope structures. These types of dynamical changes are conveniently viewed in terms of the *breaking* and *reconnection* of magnetic field lines, thus explaining the origin of the term *magnetic reconnection*. The process is thought to play an important role in a diverse range of applications including solar flares, geomagnetic substorms, magnetic fusion devices, and a wide variety of astrophysical problems. Despite a great deal of effort, scientists are still working to understand a range of basic questions regarding the onset, structure, and dynamical evolution of magnetic reconnection in large-scale systems.

Many of the outstanding scientific challenges are related to the vast separation of spatial and temporal scales inherent to most applications. High-temperature plasmas are very good electrical conductors, which implies that the magnetic flux is constrained to move together with the plasma (commonly known as the *frozen-flux* constraint). For magnetic reconnection to proceed, it is necessary to break the frozen-flux constraint within so-called *diffusion* regions. This can occur either through collisional dissipation, which gives rise to electrical resistivity and viscosity, or through a variety of different plasma kinetic effects. In most applications, the diffusion regions are quite small in comparison with macroscopic scales, but play a critical role in the evolution. Thus researchers are working to understand the

Pictured left to right: Bill Daughton and Vadim Roytershteyn

basic physics of these regions, as well as the coupling to the larger macroscopic systems. Due to the complicated nonlinear physics, simulations have played an important role in scientific progress. Most previous simulation studies have focused on 2D models using a variety of fluid and kinetic descriptions. With increasing computer power, these 2D simulations have progressed to larger scale systems and raised a number of new questions. In particular, recent 2D kinetic simulations have demonstrated that diffusion regions often develop elongated current sheets that are potentially unstable to a variety of different plasma instabilities [1]. Understanding the dynamical evolution of these layers in 3D systems is a formidable challenge that clearly requires petascale computing.

To address these questions, LANL scientists are utilizing the 3D kinetic plasma simulation code VPIC [2], which provides a first-principles description of the physics. The primary science goal is to better understand the role of plasma instabilities on the 3D evolution of reconnection layers in both space and laboratory plasmas. The simulations performed on Roadrunner were of unprecedented scale and complexity, using upwards of 4096 Cell chips and MPI, or message passing interface, ranks, ~200 billion particles, and requiring careful attention to boundary conditions, collisional physics and extensive new diagnostics. Despite these complications, the simulations achieved a factor of ~3x speedup using the Cell processors, and the results are leading to a variety of new insights into the influence of plasma instabilities on reconnection. Here we give a brief overview of these simulations and highlight a few key science results emerging from these efforts.

In space and astrophysical applications, magnetic reconnection typically occurs in highly collisionless parameter regimes and involves dynamical structures on both ion and electron kinetic scales. This is computationally challenging since the kinetic time scales are separated by the ion to electron mass ratio m_i/m_e while the spatial scales are separated by $(m_i/m_e)^{1/2}$.

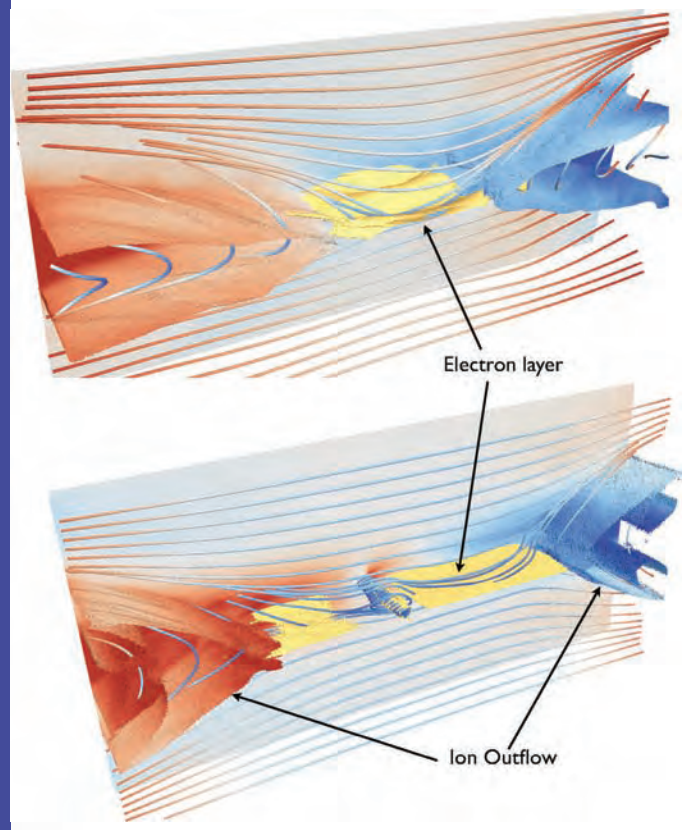
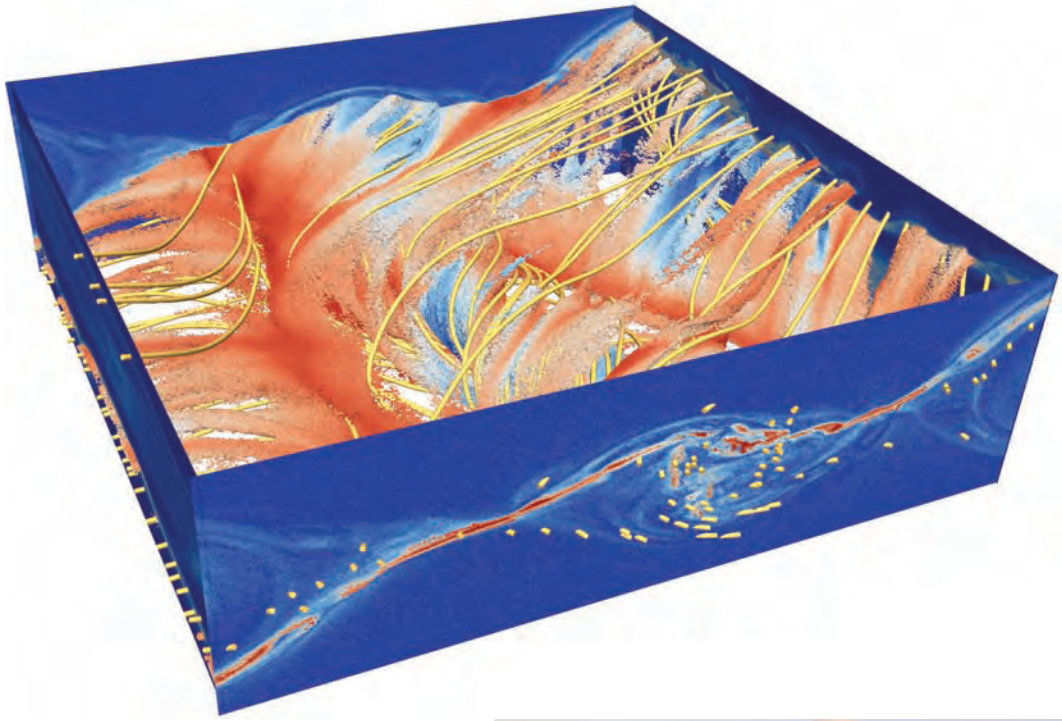


Fig. 1. Open boundary simulations for neutral sheet geometry feature two types of secondary instabilities within the electron layer: an electromagnetic kink wave (top) and flux rope formation (bottom). The boundary conditions permit inflow of new plasma (the top and bottom) and outflow of reconnected plasma (left and right). The outflow jets are visualized with a particle density isosurface colored by ion outflow velocity while the central electron current sheet (yellow) corresponds to an isosurface of electron current density. Sample magnetic field lines are colored by the magnitude of the magnetic field. Simulation was performed with mass ratio m_i/m_e=300, using 4096 ranks, 245 million cells and 147 billion particles.

For example, in the most common case of an electron-proton plasma the electron cyclotron frequency is m_i/m_e=1836 times faster than the ion cyclotron frequency while the electron gyroradius is ~43 times smaller than the ion gyroradius. Furthermore, the global scales of interest in astrophysical plasmas are vastly larger than an ion gyroradius. Due to this immense scale separation, it is not possible to study the global 3D evolution while simultaneously resolving the kinetic scales. To make progress, it is necessary to reduce the scale separation by employing artificial mass ratios m_i/m_e~100-300 and furthermore to focus on the most interesting regions for reconnection, which are considerably smaller. These preferred sites for the onset and initial development of reconnection are thin current sheets, where the intense current density gives rise to a rapid rotation in the magnetic field. As reconnection develops, a flow pattern is set up that brings new plasma and magnetic flux towards a reconnection site and expels the reconnected plasma as a high-speed jet tangent to the initial layer. It is possible to accommodate these reconnection flows and effectively mimic a much larger system by employing a suitable set of open boundary conditions to permit plasma and magnetic flux to cross the boundaries [1]. This approach was employed to study several types of initial conditions on Roadrunner, resulting in a wealth of interesting new results.

Several of these open simulations focused on neutral sheets, where the initial magnetic field reverses sign across an ion scale layer and goes to zero in the center. After the onset and initial evolution, the diffusion region features highly elongated electron scale layers that are unstable to several distinct secondary instabilities. As illustrated in Fig. 1, these modes include an electromagnetic wave that gives rise to kinking of the electron layer and

Fig. 2. Open boundary simulations for guide field geometry feature highly elongated electron current layers that are unstable to flux rope formation over a wide range of angles. These plasma instabilities cause the sheets to break into filaments as illustrated by an isosurface of the current density colored by the plasma density. Some sample magnetic field lines are shown in yellow and cutting planes along the perimeter also show current density. This simulation was performed with a guide field equal to 50% of the reconnecting field, mass ratio m_i/m_e= 64, 2048 ranks, 360 million cells, and 72 billion particles.



a secondary reconnection instability that gives rise to flux rope formation. The unstable kink wave propagates in the direction of the ion drift with a wavelength similar to recent predictions from kinetic theory [3]. The flux rope instability is the 3D analogue of secondary island formation that has been previously reported in 2D [1]. Although these processes are qualitatively similar to recent 3D electron-positron (m_i=m_e) simulations [4], this is the first time they have been observed for the high mass ratio limit m_i/m_e=300 that is now feasible with Roadrunner. While researchers are still exploring the implications of these results, they are potentially of great interest from several perspectives. First, the kink instability is within a range of frequencies where both ions and electrons can exchange momentum through the wave, potentially giving rise to a wave-induced resistivity, a problem of long-standing theoretical interest in reconnection physics. In addition, flux rope formation provides one mechanism to control the length of the electron layer by breaking the diffusion region into two current sheets as illustrated in Fig. 1.

The formation of flux ropes becomes significantly more complicated when a finite guide field is included in the initial conditions. In this limit, the magnetic field rotates across the layer but always remains finite. This type of initial condition is unstable to flux rope formation at oblique angles across the initial ion scale current sheet, leading to new electron scale current layers that are unstable to secondary flux ropes over a wider range of oblique angles. As illustrated in Fig. 2, the current density forms intense filamentary structures leading to complicated magnetic field topologies. LANL scientists are presently working to understand how some of the main features in this complex evolution compare with predictions from kinetic theory and what role the flux rope interactions may play in the acceleration of highly energetic particles.

Magnetic reconnection is also of great interest in laboratory experiments, which offer the ability to study the structure and dynamics in a controlled setting. While there are good reasons to believe that the reconnection physics is similar in laboratory experiments, the plasma parameters and boundary conditions are quite different than in space. Three-dimensional kinetic simulations may serve as a bridge to help extrapolate

ideas, which have been validated in the laboratory, to regimes of direct relevance to space and astrophysical plasmas.

The Magnetic Reconnection eXperiment (MRX) at the Princeton Plasma Physics Laboratory (PPPL) is one leading experiment that has reported detailed measurements of the diffusion region including the kinetic structure of the electron layer [5]. In this experiment, magnetic reconnection is driven by reducing the current within two axi-symmetric flux cores. Weak binary Coulomb collisions are necessary to properly describe these plasmas, but the applicability of fluid models is questionable. The kinetic approach within the VPIC code includes experimental boundary conditions appropriate to MRX [6] and a Monte Carlo treatment of the collisions [7].

On Roadrunner, a series of simulations were performed to examine the influence of collisionality and plasma instabilities on the structure of the diffusion region. An example 3D simulation in Fig. 3 illustrates an electromagnetic instability within the central electron scale current layer. This instability has certain similarities with the kink instability discussed in Fig. 1, and it could potentially play a role in facilitating momentum exchange, or in broadening the electron layer.

Researchers at LANL are working with the scientists at PPPL to perform detailed comparisons with these new simulation results. These validation efforts will include the characteristic thickness and length of the electron layer as well as the observed electromagnetic wave spectra.

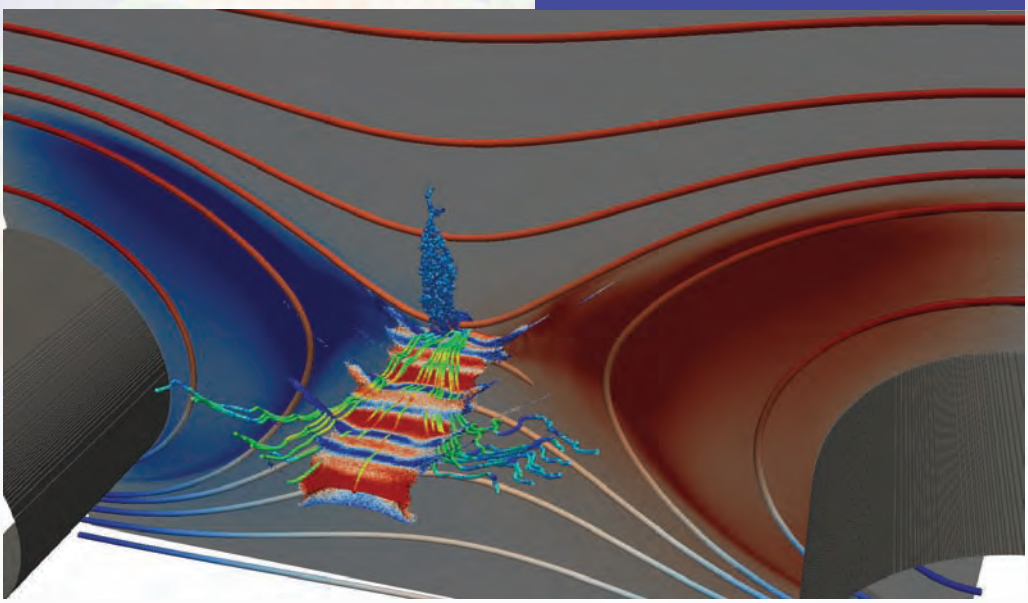
Together with ongoing theoretical work, these efforts are expected to shed new light on the influence of plasma instabilities on magnetic reconnection.

We are grateful for the visualization tools provided by Pat Fasel, Jim Ahrens and John Patchett. The figures in this article were created with ParaView.

References:

[1] Daughton, W., J. Scudder, and H. Karimabadi, *Phys. Plasmas* **13,** 072101, (2006).
[2] Bowers, K.J., B.J. Albright, L. Yin, B. Bergen, and T.J.T. Kwan, *Phys. Plasmas* **15,** 055703 (2008).
[3] Daughton, W., *Phys. Plasmas* **10,** 3103 (2003).
[4] Yin, L., et al., *Phys. Rev. Lett.* **101**, 125001 (2008).
[5] Ren, Y., M. Yamada, H. Ji, S. Dorfman, S.P. Gerhardt, and R. Kulsrud, *Phys. Plasmas* **15**, 082113 (2008).
[6] Dorfman, S., W. Daughton, V. Roytershteyn, J. Ji, Y. Ren, and M. Yamada, *Phys. Plasmas* **15,** 102107 (2008).
[7] Daughton, W., V. Roytershteyn, B.J. Albright, H. Karimabadi, L. Yin, and K.J. Bowers, *Phys. Plasmas* **16**, 072117 (2009).

Fig. 3. Simulations were performed on Roadrunner with geometry and boundary conditions relevant to the MRX experiment [4,6] including the influence of weak Coulomb collisions [7]. The reconnection process is driven by reducing the currents inside the flux cores (grey cylinders) which pulls magnetic flux inward towards the cores. The resulting ion flow velocity is illustrated on the back cutting plane along with characteristic magnetic field lines. The central electron current sheet is illustrated with an isosurface of the current density, colored by vertical component of the current density to show the plasma instability within the electron layer. Sample electron streamlines illustrate the electron flow from the upstream region through electron layer into the downstream region. This simulation was performed with mass ratio m_i/m_e = 300 using 2880 ranks, 720 million cells and 144 billion particles.



New frontiers in viral phylogenetics

Tanmoy Bhattacharya[1], Marcus Daniels[1], and Bette Korber[1]

[1] Los Alamos National Laboratory

Pictured left to right:
Bette Korber, Marcus Daniels, Tanmoy Bhattacharya

Rapidly evolving viruses pose one of the major public health threats today. Among these, the Human Immunodeficiency Virus (HIV) that causes the Acquired Immune Deficiency Syndrome (AIDS) is particularly devastating, infecting 33 million people–with millions of AIDS-related deaths and new infections each year. Vaccines against such highly variable viruses have been unable to cope with the diversity of circulating strains: When a vaccine immunogen is presented to the human body, the elicited immune memory fails to recognize most other strains of the virus. This calls for both a thorough understanding of the adaptability of the virus in its war against the human host and an intelligent design of vaccines that would provide lasting immunity against the virus. Starting with the

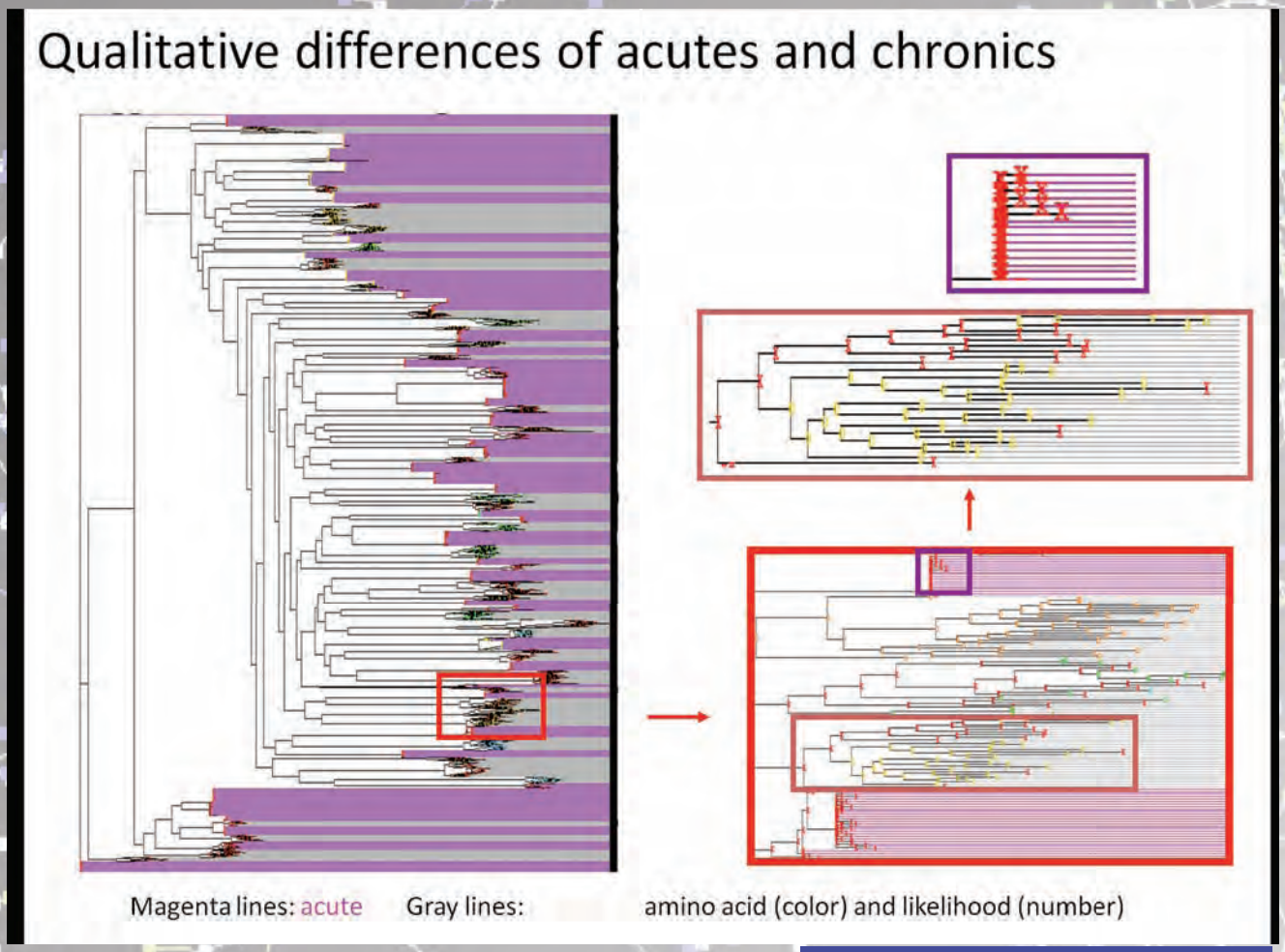


Fig. 1. Founder effects lead to entire clades sharing characteristics. If the sampling of the various clades is nonrandom, this leads to apparent correlations with traits of interest. True causal correlations, however, show up as correlations with changes.

establishment of a central sequence repository for the virus, to establishing that the virus has been circulating in humans since the early part of the 20th century [1], LANL has been at the forefront of such theoretical biology research and has contributed substantially to the field. We also developed ideas of artificial immunogens that better capture the observed diversity of HIV strains than any natural strain can do [2, 3]. While preliminary results on this research are sufficiently promising [4] to move to human trials, it is desirable to advance the field of vaccine design from such data-mining techniques to biological knowledge-based approaches.

The adaptive arm of the human immune system consists of three basic branches: the first is the Cytotoxic T Lymphocyte arm, which recognizes distinctive fragments of foreign proteins being manufactured in the body (i.e., viral proteins in infected cells) with very high specificity, the second is the Helper T Lymphocyte arm, which produces cytokines that orchestrate the immune response and have antiviral activity, and the third is the B-cell or antibody arm, which recognizes distinctive shapes on the surface of fully folded proteins.

Vast amounts of data on the interaction between HIV and all three arms of the human immune system are available, but the patterns of correlations are cryptic. Evolutionary systems are marked by long time scales, so that observed patterns in data can be due to correlations imposed by the initial historical emergence of a lineage of viruses, or founder effects, as well as due to biological interactions. In fact, not accounting for these effects leads to vastly erroneous statistical conclusions about the effect of the T-cell induced immunity on the evolution of the virus in populations [5]. But, whereas the sequence, i.e., the state of the virus, indeed depends on its evolutionary history, the changes that it undergoes are almost independent of changes in other lineages. Thus, true causal correlations are manifest also in correlations with these changes (see Fig. 1.), and in our work it was shown to detect effects that were validated experimentally. The separation of the two effects, i.e., a phylogenetic correction, thus needs access to these changes, and requires us

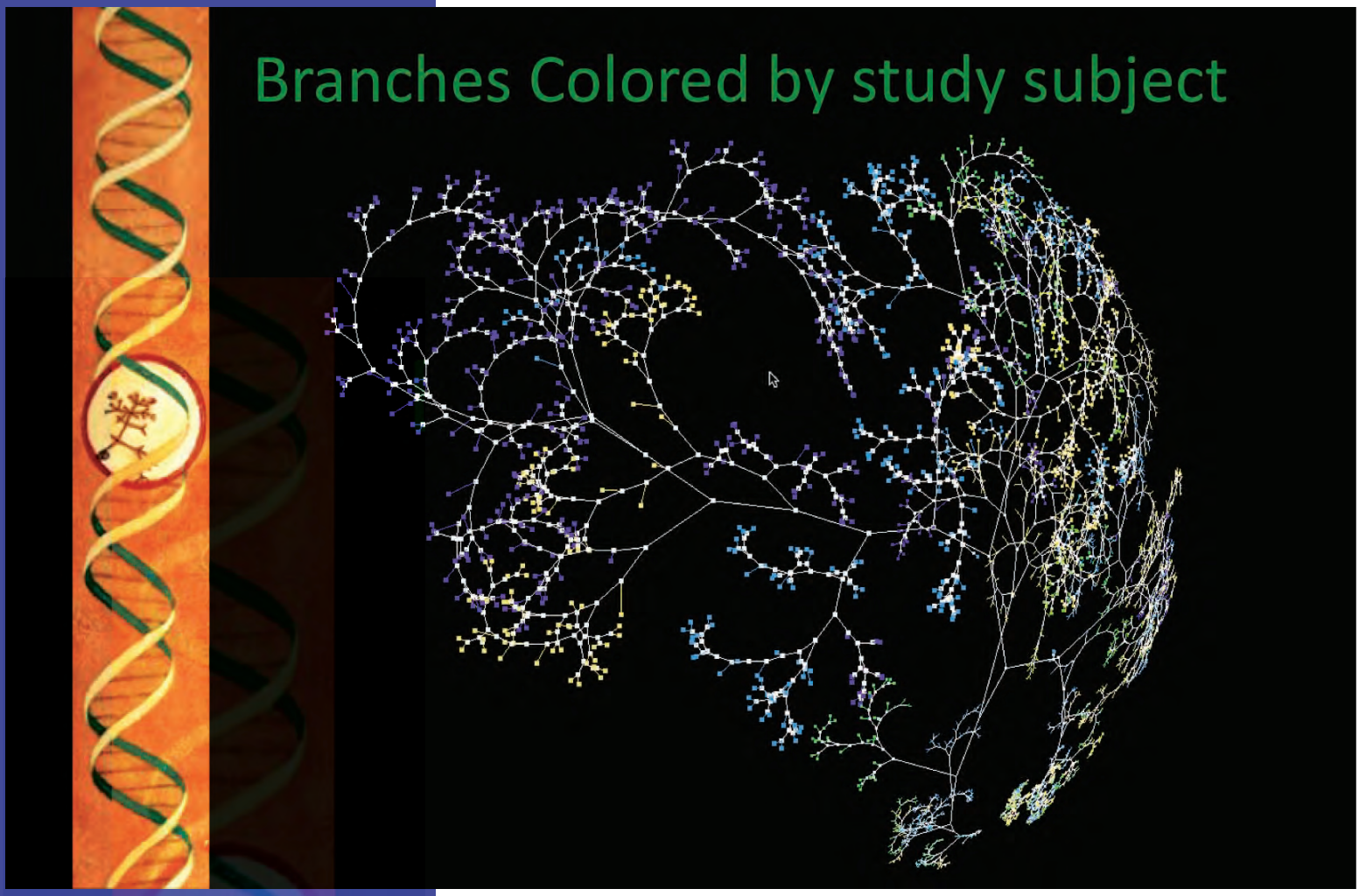


Fig. 2. A phylogeny of about 10,000 HIV sequences coloured by the study subject that was used to implement phylogenetic correction on the observed correlation between genotype and phenotype.

to be able to statistically assess the genealogical relationships between the viruses and reconstruct the ancestral forms of the viruses.

Fortunately, evolution happens by the accumulation of random mutations, most of which are effectively neutral in that they do not affect the fitness of the virus to live and infect its hosts. The covariation of these mutations, then, carries a signal of shared history. This can be used to construct a phylogenetic tree and evolutionary model that leads to random changes, and, simultaneously, the ancestral forms of the virus are also reconstructed statistically. But this reconstruction is technically challenging because the number of possible relationships grows factorially with the number of sequences sampled, and even heuristic searches fail to find reasonable models without extensive computations.

For example, a vaccine needs to prepare the body for fighting an incoming virus that can establish an infection in the healthy body. The virus that exists in a chronic patient, however, results from a long process of virus-host interaction and may be qualitatively different than the virus at the time of infection.

The characterization of these differences is a daunting task: the viral diversity in a chronic patient needs to be represented by at least three to four dozen sequences each, but to control for the phylogenetic effects, we need at least two to three hundred patients infected with various subtypes of the virus. This means that one needs to fit together some 10,000 HIV sequences into a giant family tree of HIV viruses, and find those patterns that distinguish acute and chronic viruses.

We had performed a preliminary study to look at only one subtype of the virus: the B subtype that is predominant in most of the developed world. What we found was that a particular site in the so-called signal peptide region of the precursor of the envelope glycoprotein, which coats the virus, was far more conserved in the early sequences than is expected by chance sampling. It was important to know whether this result is universal to all HIV strains, or specific to the B subtype. We therefore used Roadrunner to construct a phylogenetic tree of about 10,000 sequences from over 400 people (see Fig. 2) and studied the variability of this site. In this much bigger tree, we did not find the site in the signal peptide region to be any less variable in the acute patients, and it is, therefore, likely that the signature we had found was subtype-specific.

We have since started analyzing data that directly measures the immunogenicity of viruses by collecting antibody containing sera and viruses from the same patients. These sera are seen to cluster into groups with markedly different neutralization potencies [6]. We then look at the sequences of viruses from the subset of patients who make potent sera that neutralize the activity of a vast panel of viruses, and compare them with those that make average or poor responses. Even though we did not know a priori whether the difference that we observe is due to host genetics, viral factors, or stochastic events, a preliminary analysis uncovered sites in the viral sequence where changes correlated with the induction of a good immune response. Interestingly, in a model structure, these sites clustered around the part known as the CD4 inducible region of the viral envelope that had long been suspected to be involved in the induction of beneficial antibodies. In the current analysis, however, we used a small number of sera; this analysis is being extended to a much larger panel.

Looking forward, technological advances are making it possible to get hundreds of thousands of sequences from each patient, providing a detailed view of the intrahost diversification in adaptation of the virus. Phylogenetic analysis of such datasets is still a challenge. More importantly, however, in current calculations we are looking only at the best possible phylogeny that we can resolve, and neglecting all the other possible phylogenies. To increase our confidence in the results, we will need to incorporate this phylogenetic *uncertainty*, and such calculations will stretch the limits of conventional computation even for a small number of sequences. The advent of petaflops-scale computing, exemplified by Roadrunner, is coming to the rescue, and in the near future we expect to see a fully detailed phylogenetic analysis of such problems with all sources of uncertainty correctly propagated to the final conclusion. Such computational techniques, complemented with our advance in experimental methods and theoretical understanding, we hope, will usher in a new era that finally stops this deadly epidemic.

References:
[1] Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B.H. Hahn, S. Wolinsky, and T. Battacharya, *Science* **288,** 1789 (2000).
[2] Gaschen, B., J. Tayler, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B.H. Hahn, T. Bhattacharya, and B. Korber, *Science* **296,** 2354 (2002).
[3] Fisher, W., S. Perkins, J. Theiler, T. Bhattacharya, K. Yusim, R. Funkhouser, C. Kuiken, B. Haynes, N.L. Letvin, B.D. Walker, B.H. Hahn, and B. Korber, *Nature Medicine* **13,** 100 (2007).
[4] Santra, S., B.T. Korber, M. Muldoon, D.H. Barouch, G.J. Nabel, F. Gao, B.H. Hahn, B.F. Haynes, and N.L. Letvin, *PNAS USA* **105,** 10489 (2008).
[5] Bhattacharya, T., M. Daniels, D. Heckerman, B. Foley, N. Frahm, C. Kadie, J. Carlson, K. Yusim, B. McMahon, B. Gaschen, S. Mallal, J.I. Mullins, D.C. Nickle, J. Herbeck, C. Rousseau, G.H. Learn, T. Miura, C. Brander, B. Walker, and B. Korber, *Science* **315,** 1583 (2007).
[6] Doria-Rose, N.A., R.M. Klein, M.M. Manion, S. O'Dell, A. Phogat, B. Chakrabarti, C.W. Hallahan, S.A. Migueles, J. Wrammert, R. Ahmed, M. Nason, R.T. Wyatt, J.R. Mascola, and M. Connors, *J. Virol.* **83**, 188 (2009).

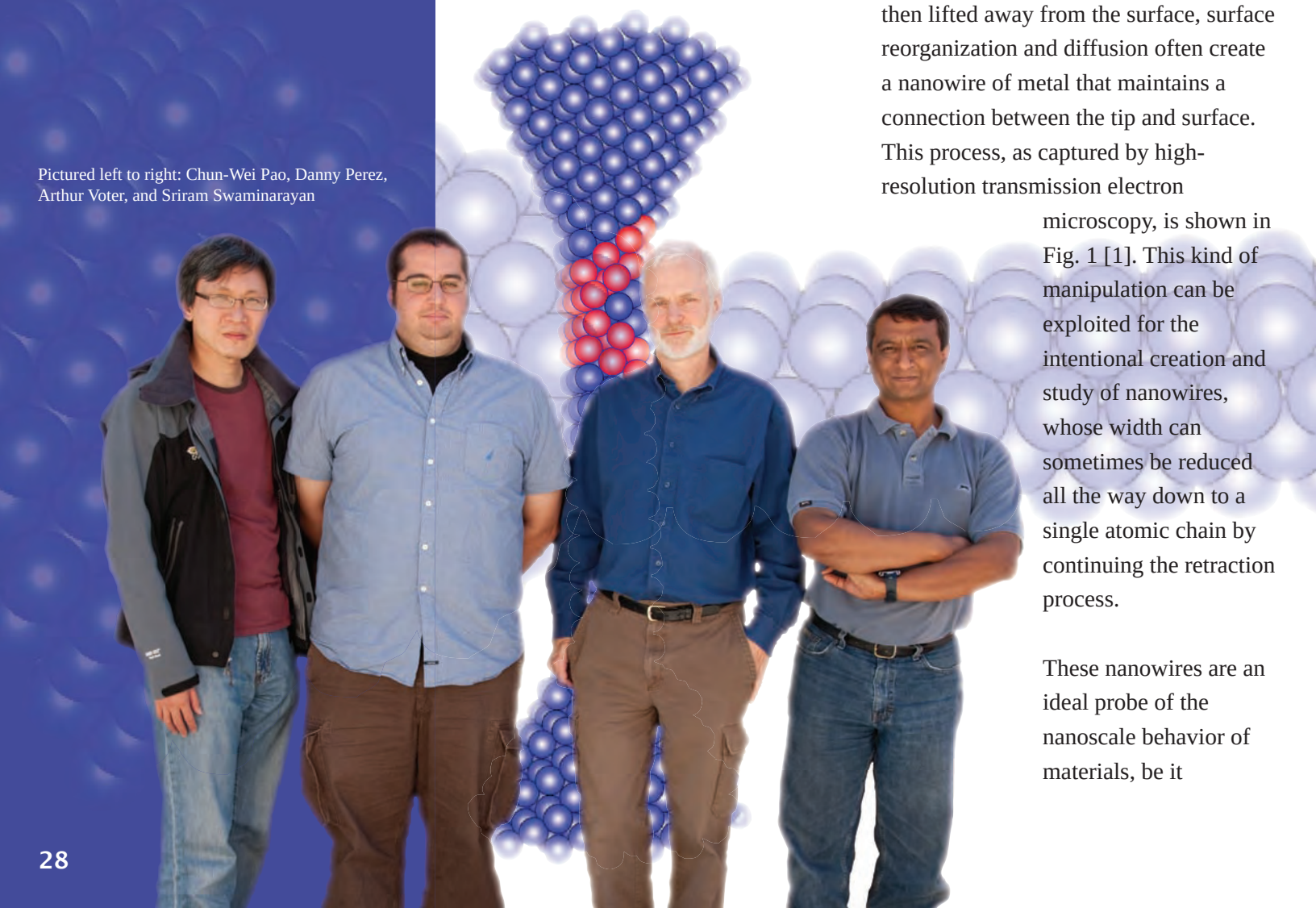
Simulating the mechanical behavior of metallic nanowires over experimentally accessible timescales on Roadrunner

Danny Perez[1], Chun-Wei Pao[2], Sriram Swaminarayan[1], and Arthur F. Voter[1]

[1] Los Alamos National Laboratory, [2] Academia Dinica, Taiwan

For many years now, we have witnessed explosive growth in our ability to control the structure of materials down to the atomic scale. For example, it is now possible to bring the tip of an atomic force microscope (which might be only a few atoms wide at the apex) in contact with a surface. In the case of a metal tip contacting a metal surface, or another tip, a bonded contact forms, and if the tip is then lifted away from the surface, surface reorganization and diffusion often create a nanowire of metal that maintains a connection between the tip and surface. This process, as captured by high-resolution transmission electron microscopy, is shown in Fig. 1 [1]. This kind of manipulation can be exploited for the intentional creation and study of nanowires, whose width can sometimes be reduced all the way down to a single atomic chain by continuing the retraction process.

These nanowires are an ideal probe of the nanoscale behavior of materials, be it



Pictured left to right: Chun-Wei Pao, Danny Perez, Arthur Voter, and Sriram Swaminarayan

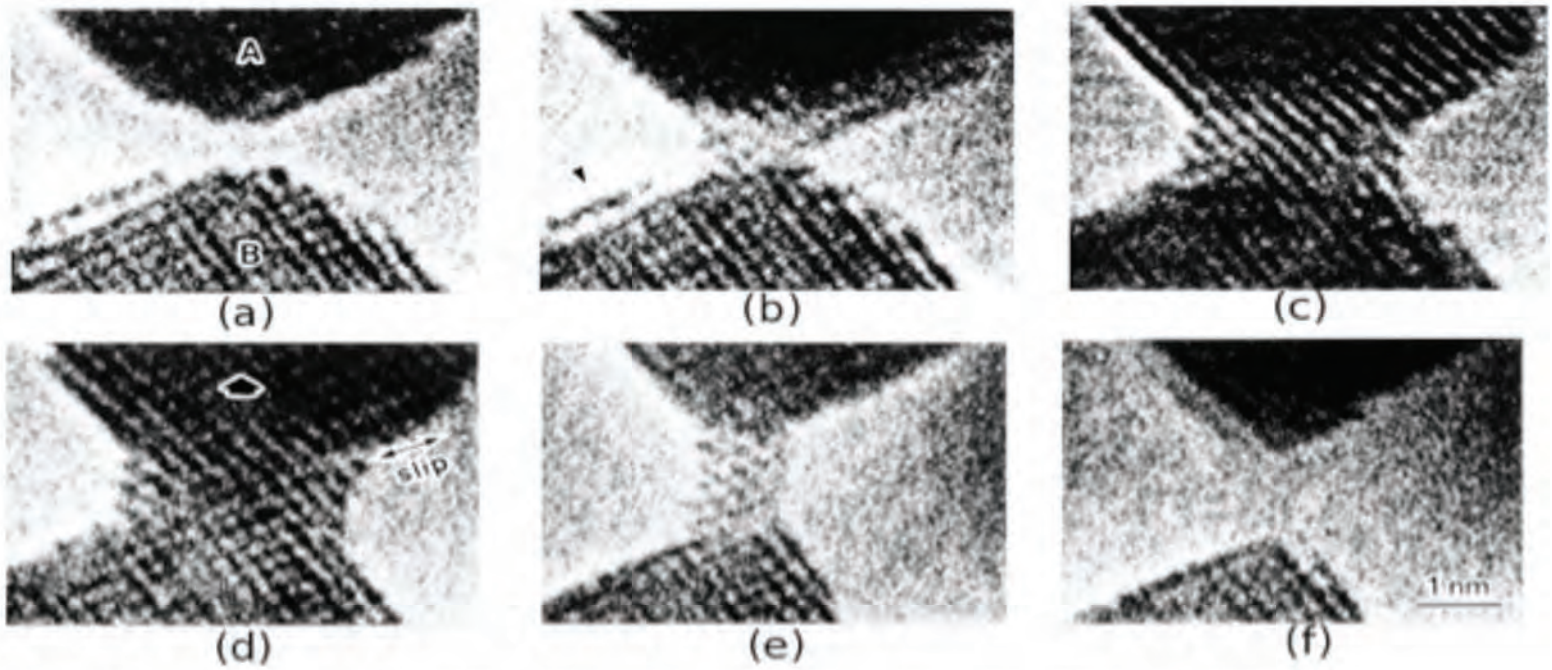


Fig.1. A series of high resolution transmission electron microscope (HRTEM) images showing the contact formation - retraction - and rupture processes of two gold tips. (a)-(c): contact formation process with (a) corresponding to t=0 sec, (b) t=3.7 sec, (c) t=4.3 sec; (d)-(f) retraction and rupture processes: (d)t=0 sec, (e)t=2 sec, (f) t=3 sec.

mechanical [2], or electrical [3], and hence are of interest both for fundamental studies and because of their expected importance in various nanotechnology applications such as electrical conductors and electrical or mechanical switches. However, a deeper understanding of the fundamental nanoscale behavior of materials is required before these applications can become widespread. Indeed, experiments at the nanoscale are hard to control and sometimes lack the resolution necessary to fully understand how the systems behave. For example, metallic nanowires are often seen to completely disappear from one frame to another, leaving one completely in the dark about the basic mechanisms leading to their failure. Further, the act of imaging itself leads to an uncontrolled increase of the temperature of the wire, making it very difficult to perform these experiments under controlled and reproducible conditions.

There is thus a pressing need for atomistic numerical simulations to complement such experiments and help interpret and understand them. The most powerful tool for performing this kind of simulation is molecular dynamics (MD), whereby one integrates the equations of motion of all the atoms in the system, advancing the positions and velocities of the atoms by repeatedly taking small steps forward in time. In this way, one learns about the evolution of the system with full atomistic detail. However, for many systems and processes we would like to study, there is a serious problem with the mismatch in time scale. For example, the nanowire-stretching process discussed above usually takes place over seconds or, at the very fastest, milliseconds. In contrast, conventional MD simulations are limited to a time scale of about one microsecond, even on the fastest parallel computers, i.e., 10^3 to 10^7 times faster than the experimental reality. Because of this extremely large gap, physical arguments suggest that current simulations do not adequately represent reality.

The way to overcome these limitations is to use so-called Accelerated Molecular Dynamics methods to reformulate the problem in a form that is more amenable to computer simulation. For example, the Parallel Replica Dynamics (ParRep) [4] method developed at LANL generates a proper evolution of the system while allowing a parallelization of the problem in the time-domain. This allows one to make optimal use of massively parallel computers to reach timescales that are orders of magnitude longer that what could be done with conventional MD. When implemented on petascale supercomputers like Roadrunner, the ParRep method enables one to study the evolution of nano-scale systems (containing about a thousand atoms or so) over unprecedented timescales. Using 12,000 replicas on 12,000 out of

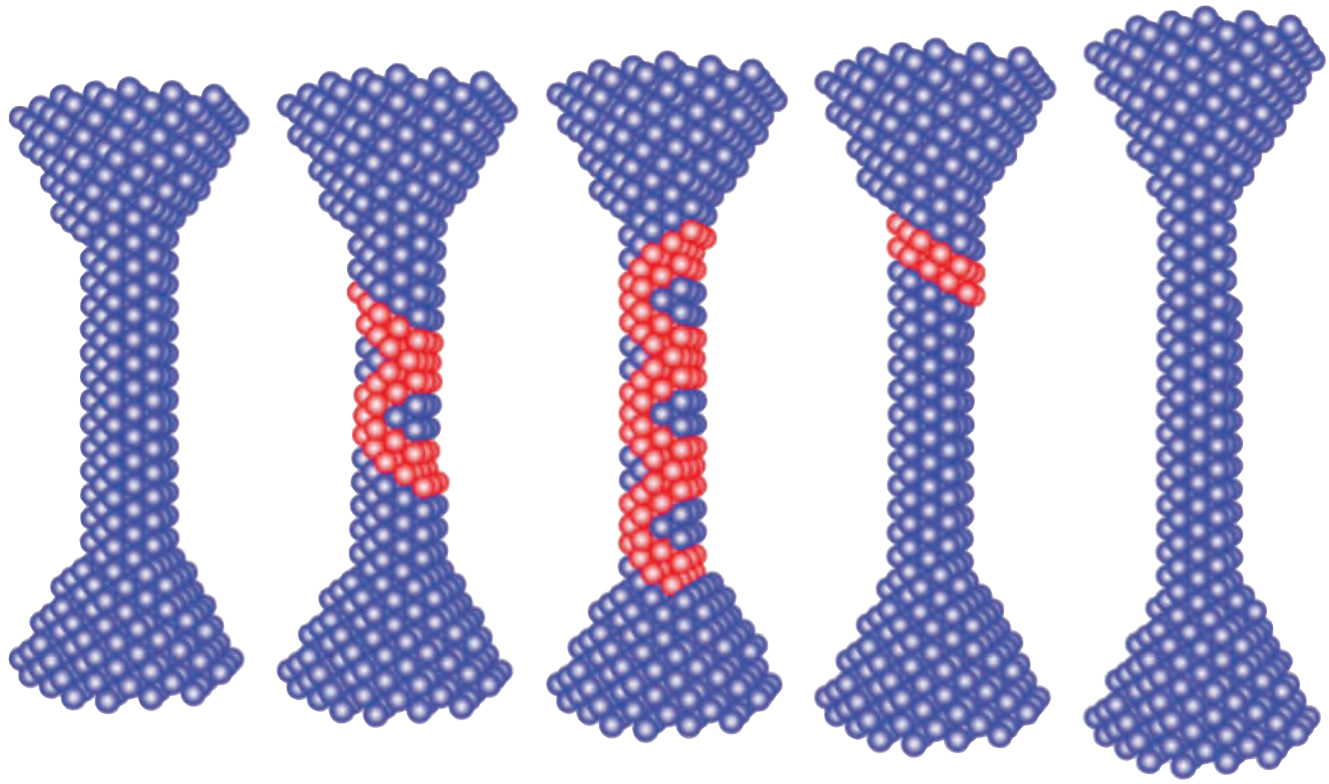


Fig. 2. Early stage of a ParRep simulation of the stretching of a silver nanowire on Roadrunner at a temperature of 300K and a retraction velocity of 10^{-5} m/s. From left to right, t = 0 µs, t = 30 µs, t = 60 µs, t = 90 µs, t = 150 µs. Defective sections of the wire are shown in red.

the 12,240 Roadrunner Cell chips we obtained a simulation rate of about 0.1 ms per wall-clock hour, thereby allowing a direct connection between experiments and fully atomistic simulations.

Using ParRep on Roadrunner, we simulated silver nanowire stretching experiments similar to those illustrated in Fig. 1 for different nanowire sizes, temperatures and retraction velocity, reaching more than one millisecond of simulation time in a few instances, more than a thousand times longer than conventional techniques would have allowed. Overall, we were able to study the change of behavior of these wires while varying the strain rate by more than 4 orders of magnitude, a feat that was unthinkable before the advent of Roadrunner.

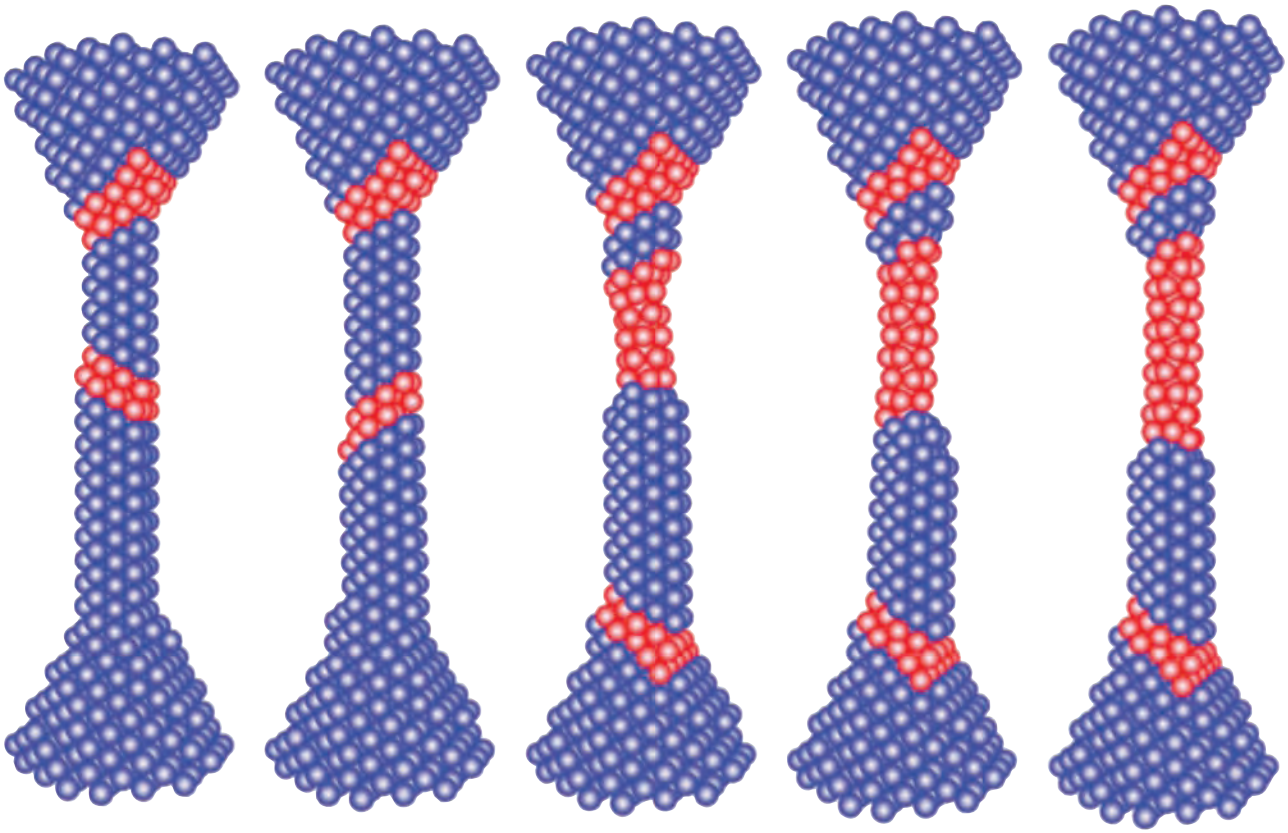
Thanks to these simulations, a picture of the evolution of these systems is emerging. The basic plastic reaction of the system when subjected to strain is to create stacking faults along (111) planes. These stacking faults are highlighted in red in Fig. 2. The formation of a zig-zag network of such stacking faults causes the release of internal stresses while leading to the elongation and narrowing of the wire. Interestingly, all wires, almost independently of temperature or strain rate, initially behave this way. As the wire is stretched further, given enough time, these stacking faults annihilate, leaving behind a defect-free wire that is uniformly thinned down relative to the initial configuration. These simulations illustrate the unique ability of these nanostructures to, under suitable conditions, heal themselves when subjected to severe external constraints. Note that this self-healing behavior emerges only on long timescales that are completely inaccessible to standard MD simulations.

The later stages of the simulation of the evolution of nanowires also revealed some completely unintuitive mechanisms by which plastic deformation occurs at the nanoscale. Shown in Fig. 3, one of these mechanisms is the conversion of bulk-like segments of the wire into low-symmetry helical structures (here a fivefold-symmetric icosahedral structure). These structures appear to be extremely tolerant of mechanical constraints. Indeed, the conversion process between two relatively stable conformations offers a continuous pathway for the stretching to occur, in contrast with competing mechanisms that lead to the accumulation of defects and ultimately to failure. Through this pathway, we have observed elongations in excess of 100% without failure. Once again this process, involving structures that are not allowed by the symmetry of the bulk crystal, demonstrate the unique ability of nanoscale systems to react to their environment in completely unintuitive ways.

The conjunction of innovative algorithms and methods with the unprecedented computational power of Roadrunner enabled us to simulate, for the first time, the mechanical behavior of metallic nanowires, which are widely foreseen as playing a major role in the next generation of nanodevices, on experimentally accessible timescales. With this new capability, it is now possible to directly assist in the interpretation of experiments as well as in the design of novel structures with precisely tailored properties.

References:
[1] Kizuka, T., *Phys. Rev. Lett.* **81**, 4448 (1998).
[2] Liang, W., M. Zhou, and F. Ke, *Nano Letters* **5**, 2039 (2005).
[3] Pascual, J.I., J. Mendez, J. Gomez-Herrero, A.M. Baro, N. Garcia, U. Landman, W.D. Luedtke, E.N. Bogachek, and H.P. Cheng, *Science* **267**, 1793 (1995).
[4] Voter, A.F., *Phys. Rev. B* **57**, 13985 (1998).

Fig. 3. Late stage of a ParRep simulation of the stretching of a silver nanowire on Roadrunner at a temperature of 300K and a retraction velocity of 10^{-5} m/s. From left to right, t =165 µs, t = 180 µs, t = 195 µs, t = 210 µs, t = 225 µs. Defective sections of the wire are shown in red.

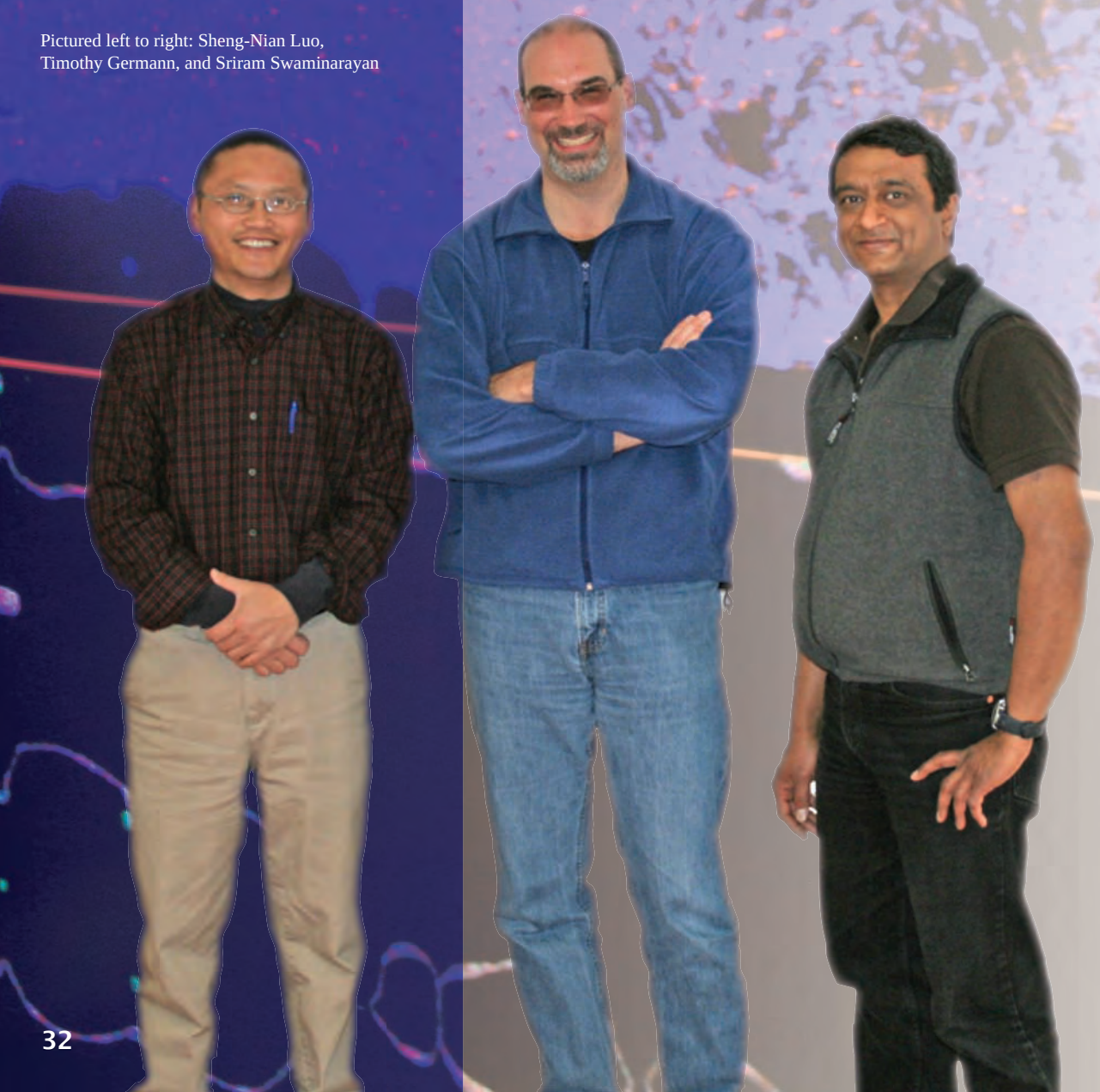


Material dynamics at extreme conditions

Timothy C. Germann[1], Sheng-Nian Luo[1], and Sriram Swaminarayan[1]

[1] Los Alamos National Laboratory

Pictured left to right: Sheng-Nian Luo, Timothy Germann, and Sriram Swaminarayan



Dynamic loading, such as high-speed sliding friction or shock impact, can dramatically affect the microstructure and properties of materials on ultrafast timescales that are difficult–if not impossible–to probe experimentally. Because sound waves travel at a few km/s, or equivalently a few nm/ps, through a typical metal, they traverse the sub-nanometer interatomic lattice spacing in much less than a picosecond. While experimentally challenging, processes occuring at such nm length and ps time scales are ideal for study by non-equilibrium molecular dynamics (MD) simulations.

Over the past decade, large-scale MD simulations have provided significant insight into the

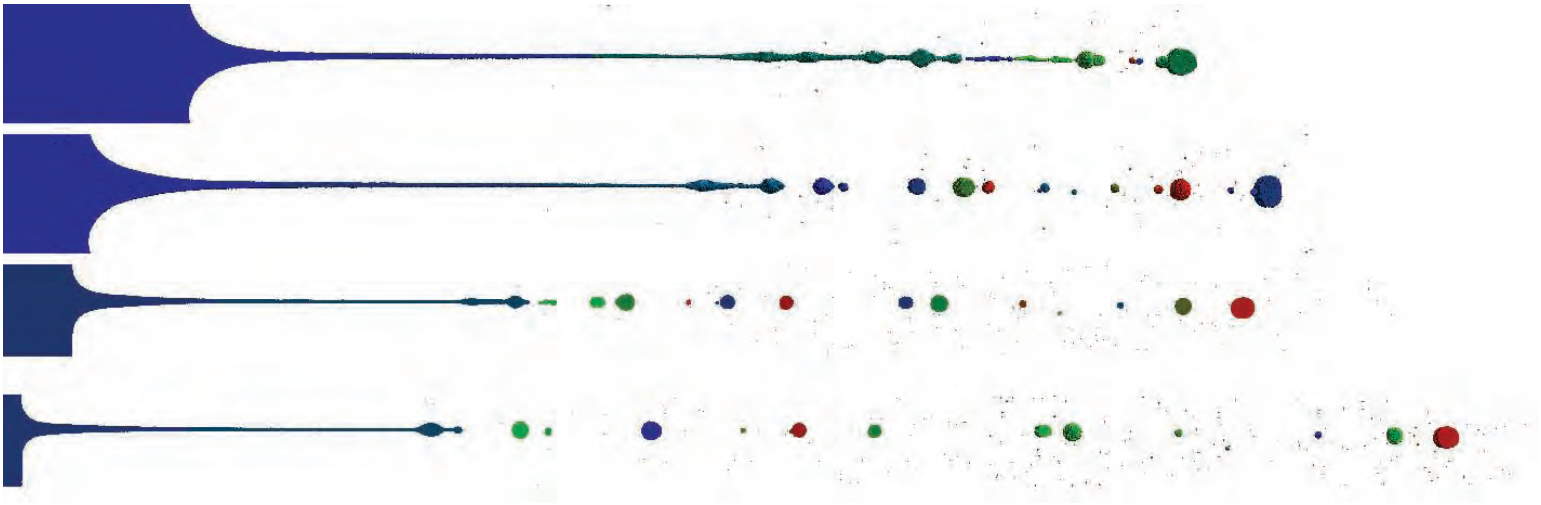


Fig. 1: Ejecta particle formation by fragmentation of an expanding jet produced when a shock wave reaches a copper free surface with a sinusoidal profile.

microscopic pathways and kinetics of shock-induced plasticity and phase transformations in single-crystal metals. Sample sizes of a few million atoms are typically sufficient; representing a cube of material with edge lengths 10-100 nm, this can capture the emergent length scale (average spacing between dislocations or product phase nuclei). Such simulations first suggested that the polymorphic (bcc-hcp) transformation in shocked iron can take place on ps timescales [1]. The predicted orientation relationship, timescale, and product grain size were all subsequently confirmed by ultrafast in situ X-ray diffraction measurements on laser-shocked iron thin foils [2]. With advanced 21st century experimental facilities such as the Linac Coherent Light Source (LCLS) at Stanford, National Ignition Facility (NIF) at LLNL, and the Matter-Radiation Interactions in Extreme Environments (MaRIE) at LANL becoming available in the next decade, we anticipate such simulation-led experiments will provide insight into even more complicated dynamic materials phenomena occurring at nm-µm length and ps-µs timescales.

The SPaSM code was originally developed in the early 1990s for the newly emerging era of massively parallel supercomputers such as the Thinking Machines CM-5, and achieved IEEE Gordon Bell Prize-winning performance by minimizing memory usage and floating-point operations. However, for the modern generation of heterogeneous, multicore architectures such as Roadrunner, it is data *movement* rather than storage or arithmetic operations that is increasingly the bottleneck. Thus, we have redesigned the entire communication infrastructure and data structures of SPaSM to allow for asynchronous interactions between the processors, in particular to accommodate the Cell processor and the unique multilayer hierarchy of Roadrunner's architecture. This effort has paid off, resulting in double-precision benchmark performance of 369 TFlop/s on the full machine [3], and a ~5x speedup for the embedded atom method (EAM) potentials typically used to model simple metals such as copper, silver, and iron.

Using the SPaSM code on Roadrunner, we are investigating the ejection of material that can occur from shocked surfaces [4]. The goal of this work is to develop models that can predict the amount of mass ejected from a shocked interface with a given surface finish and loading history (peak shock pressure, either from a supported square-wave or explosive-driven Taylor wave). We would also like to understand how that mass is distributed, namely its particle size and velocity distributions, as well as evolutions and correlations, if any, between the two. Experimentally, the total mass can be inferred by measuring the resulting momentum transfer onto an Asay foil or piezoelectric probe at some standoff distance, while particle sizes larger than a micron can be imaged using holography or X-ray radiography. However, there is no direct experimental information on the distribution of particle sizes smaller than a micron, nor on the correlation between size and velocity distributions.

On such microscopic scales, MD simulations can complement experiments by providing unique insight into the material dynamics at submicron length and subnanosecond timescales, including key fragmentation and atomization

Fig. 2. Incipient spall failure in a copper bicrystal (shocked left-to-right) by homogeneous void nucleation along the (vertical) plane of maximum tensile stress. Only undercoordinated surface atoms are shown, so that one sees the free surfaces of the impactor and target and the voids which have begun to nucleate, grow, and coalesce. The (horizontal) grain boundaries, one in the center of the sample and the other at the top/bottom periodic boundary, remain intact and are not visible.

mechanisms, but until now this problem had remained computationally intractable. The formation and transport of ejecta involves a complex range of physical processes, including Richtmyer-Meshkov instability (RMI) development in solid materials with dynamic material strength properties, classical and turbulent fragmentation and atomization, and particulate transport in a turbulent gas. We have carried out a systematic study of RMI development from a single sinusoidal surface perturbation in copper to test various RMI theories including material strength effects, which suppress the instability growth. We are using these simulations to study the evolution of the density and velocity distributions of the ejected mass, the modes of particle breakup, and ultimately to develop source theories of ejecta formation based on RMI growth, including material strength effects and transport models that describe the time-evolving particle size and velocity distributions. Earlier MD simulations were able to demonstrate the initial jet formation but could not reach timescales long enough to observe the subsequent necking instabilities leading to jet breakup and droplet formation that have now been revealed (Fig. 1). These fragmentation and atomization processes are also difficult to study experimentally, although various theories have been proposed; atomistic-level simulations such as those presented here are contributing to the development of physics-based models at LANL.

Ejecta is only one form of shock-induced material failure, occurring when a shock wave reflects from a free surface to become an expanding rarefaction (or release) fan. When two such rarefaction fans (one from the impactor free surface, the other from the target) intersect in the interior of the material, they put the material into tension and can lead to spall failure [5]. Ductile spall failure results from the nucleation, growth, and coalescence of voids. Models have been developed that account for each of these aspects, but without direct experimental information–due again to the ultra-fast time and ultra-small length scales.

Using Roadrunner, we have been able to study this process in copper bicrystals, revealing the competition between heterogeneous void nucleation at defects such as grain boundaries and homogeneous nucleation within the bulk single crystal. Large system sizes are required to clearly separate the two processes, and long timescales to explore a wide range of strain rates. Figure 2 shows a copper bicrystal loaded parallel to the horizontal grain boundaries (one in

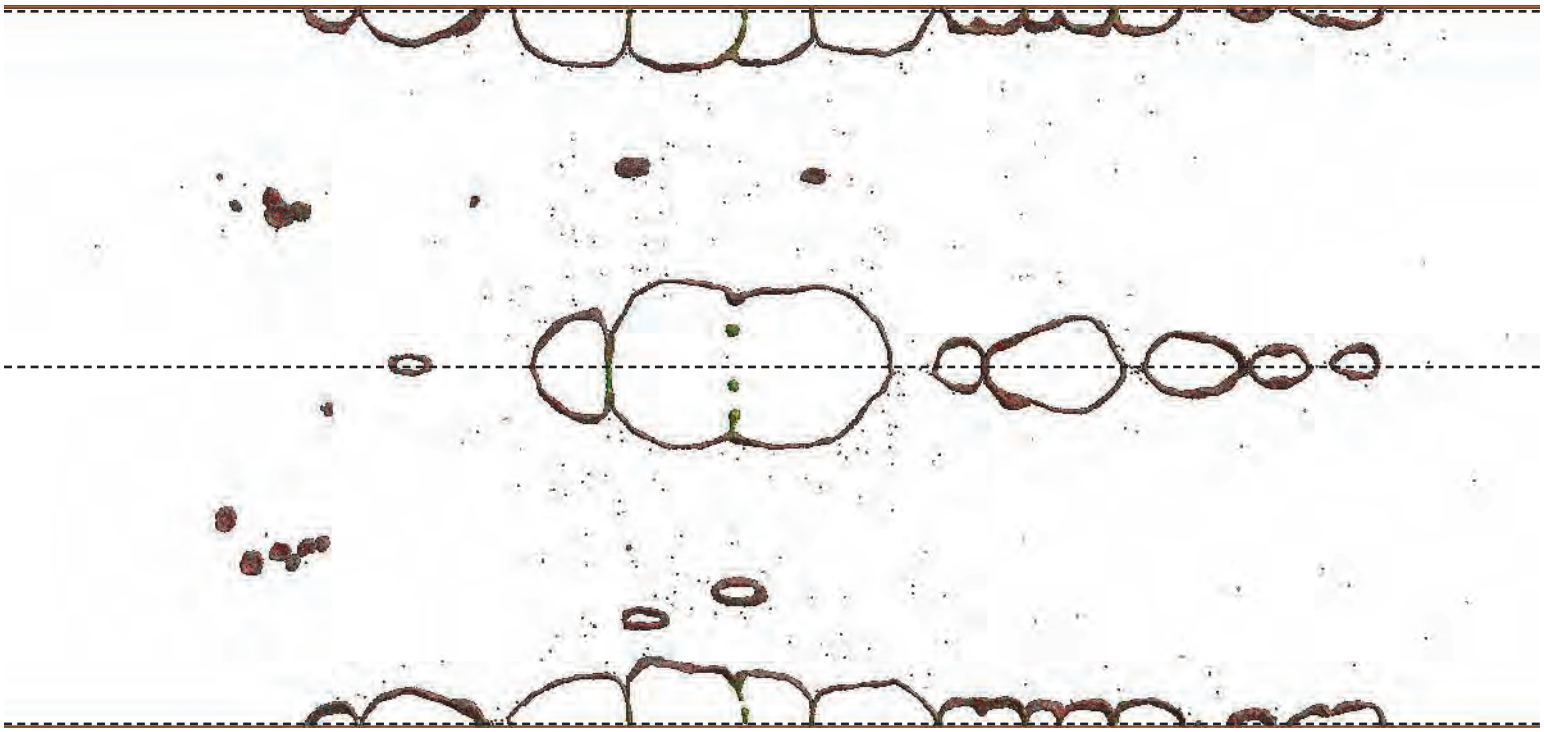


Fig. 3: Incipient spall failure in a copper bicrystal by heterogeneous void nucleation along the (horizontal) grain boundaries.

the center and the other at the top/bottom periodic boundary), with 400 m/s impact velocity.

The sample is 230 nm tall (i.e., each grain is 115 nm), 20 nm thick (into the plane, i.e., images are looking through the entire 20 nm thickness), and 205 nm long (54 million atoms in all). Following shock compression and release, dislocations and voids are produced that leave the sample in an incipient spall state; that is, with a number of voids that remain intact after growing and coalescing, but that have not caused complete fragmentation. In this example, the short sample length leads to a very high strain rate, with voids primarily nucleated homogeneously (within the grains) along the vertical plane of maximum tensile stress. On the other hand, a longer sample length (1 μm, 270 million atoms) leads to a lower strain rate and sufficient time for void nucleation to be localized at the grain boundaries, changing the failure mode from a vertical spall plane to a horizontal grain decohesion (Fig. 3). These results indicate the interplay between grain size and the timescale for nucleation kinetics, with a competition between heterogeneous and homo-geneous nucleation.

The materials science community is extremely excited about the discovery opportunities now presenting themselves with the Roadrunner-class petascale computers. In particular, we are starting to study polycrystalline materials with realistic grain sizes and interface structures at the relevant time and length scales. In addition, our redesign of the SPaSM code for Roadrunner enables us to optimize performance on the new generation of hybrid supercomputers with many-core nodes and/or graphical processing unit (GPU) accelerators.

References:

[1] Kadau, K., T.C. Germann, P.S. Lomdahl, and B.L. Holian, *Science* **296,** 1681 (2002).

[2] Kalantar, D.H., J.F. Belak, G.W. Collins, J.D. Colvin, H.M. Davies, J.H. Eggert, T.C. Germann, J. Hawreliak, B.L. Holian, K. Kadau, P.S. Lomdahl, H.E. Lorenzana, M.A. Meyers, K. Rosolankova, M.S. Schneider, J. Sheppard, J.S. Stölken, and J.S. Wark, *Phys. Rev. Lett.* **95,** 075502 (2005).

[3] Germann, T.C., K. Kadau, and S. Swaminarayan, *Concurrency Comput. Pract. Ex.* **21,** 2143 (2009).

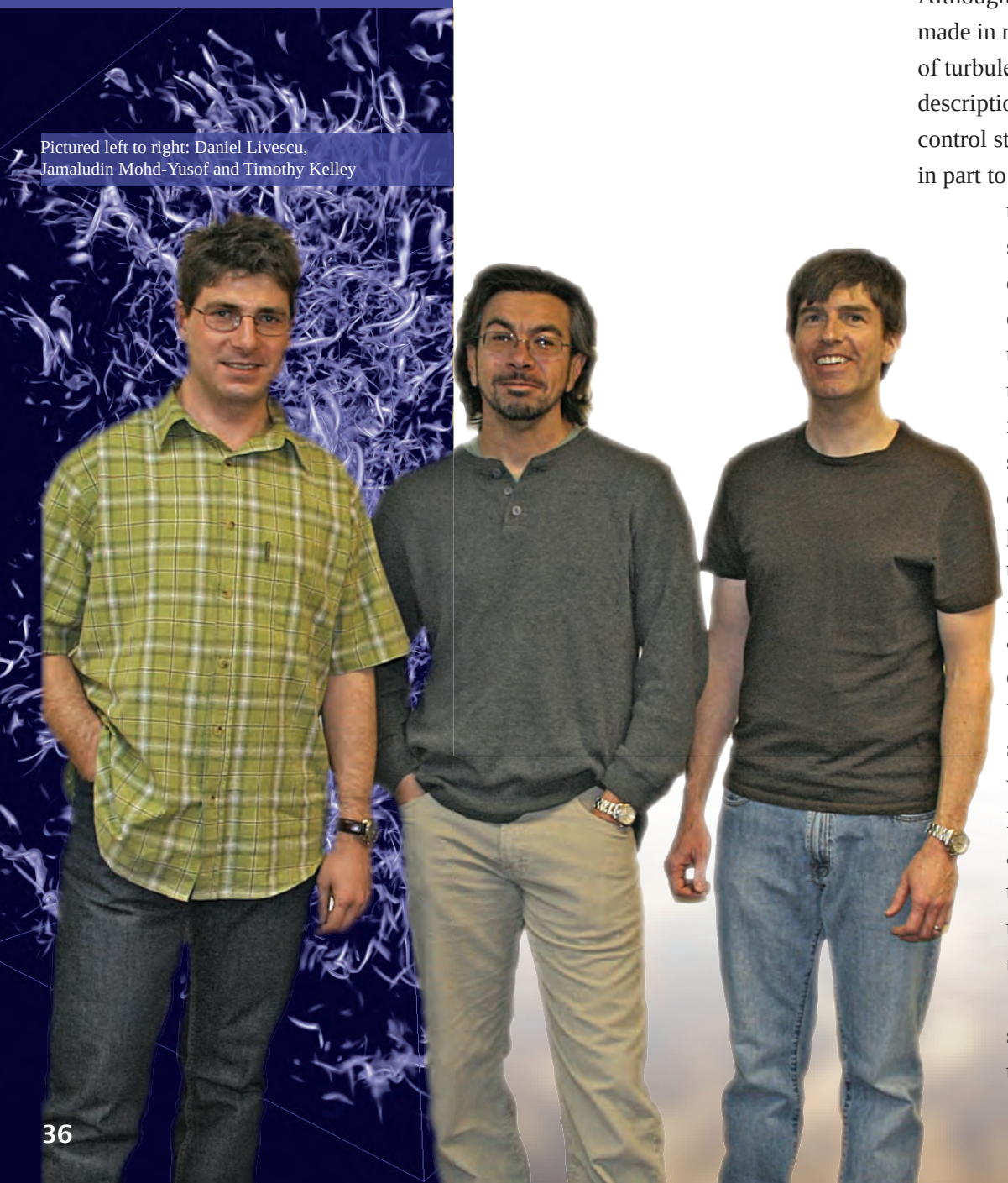
[4] Germann, T.C., G. Dimonte, J.E. Hammerberg, K. Kadau, J. Quenneville, and M.B. Zellner, *DYMAT 2009: 9th International Conference on the Mechanical and Physical Behaviour of Materials under Dynamic Loading*, 1499-1505 (2009).

[5] Luo, S.-N., T.C. Germann, and D.L. Tonks, *J. Appl. Phys.* **106**, 123518 (2009).

Direct numerical simulation of reacting compressible turbulence

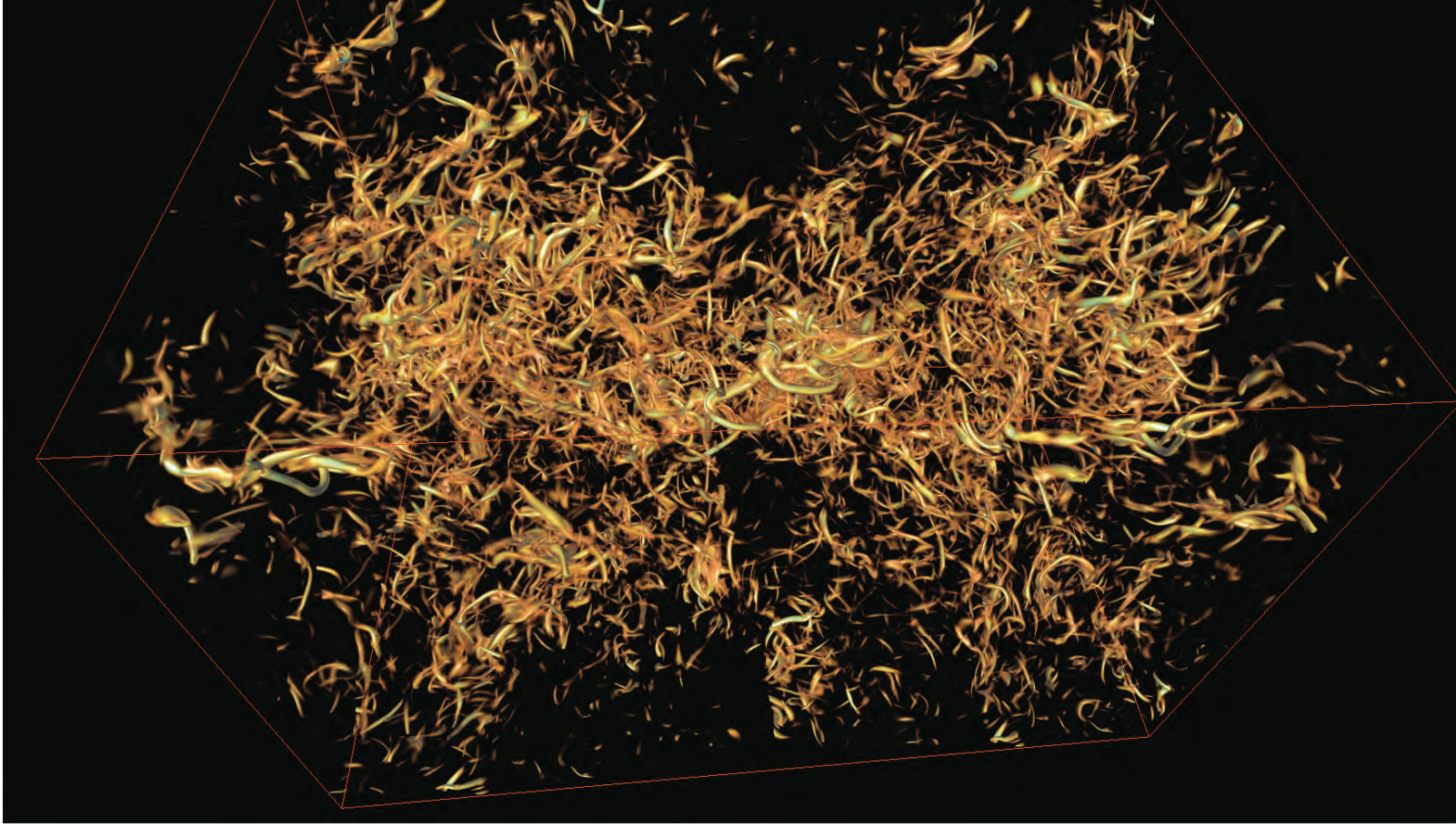
Daniel Livescu[1], Jamaludin Mohd-Yusof[1], and Timothy M. Kelley[1]

[1] Los Alamos National Laboratory



Pictured left to right: Daniel Livescu, Jamaludin Mohd-Yusof and Timothy Kelley

Although important progress has been made in recent years in our understanding of turbulence, complete quantification, description, prediction, simulation, and control still elude us. The problem is due in part to the very large range of spatio-temporal, dynamically relevant scales, but also to the multitude of problems that can be encompassed by the generic term "turbulence." If "ideal" turbulence is in a homogeneous, isotropic Kolmogorov steady state, then "nonideal" turbulence can occur due to many practically relevant effects: time-dependence, anisotropy, inhomogeneity, coupling with active scalars, shock waves, exothermic reactions, etc. Unlike kinetic theory, where significant departure from a weakly perturbed local Maxwellian is exceptional, the analogous state of nonideal turbulence is what is typical, yet the only successful turbulence theory so far is Kolmogorov's 1941 theory. It is then not surprising that the turbulence theory is still centered on the



ideas of Kolmogorov related to the existence of inertial range dynamics (e.g., a -5/3 range in the isotropic turbulence energy spectrum) and small-scale universality. However, more and more evidence points to departures from universal laws in the energy spectrum due to intermittency and a direct connection between the small and the (nonuniversal) large scales, especially in the presence of strong gradients [1] or buoyancy [2-4].

Most of the turbulence research to date has been concentrated on several canonical flows with periodic boundaries or simple jets, wakes, or boundary layers. Numerous modeling strategies have been proposed, and, while there is no "best" strategy, each of the approaches has its own advantages and domains of applicability. Yet, most practical flows are not canonical. In many situations they are driven by acceleration, as in inertial confinement fusion (ICF) or cosmic explosions, and may undergo exothermic reactions. In addition, radiation could have a significant effect, e.g., through heat gain or loss. For these complex flows, the limitations of the current modeling strategies, as well as the turbulence and mixing properties, are unknown.

Turbulence theory and the subsequent model development rely on experimental or high-resolution direct numerical simulation (DNS) data for development and verification and validation. This technique has emerged as a powerful research tool to study the physics of turbulence, for verifying and improving models, and for complementing and even guiding and helping the design of better experiments [5]. The DNS technique seeks "exact" solutions of the governing equations, so that all relevant scales are accurately solved, using high-resolution numerical simulations based on high-order accurate discretization algorithms.

Fig. 1. Enstrophy field in reacting compressible turbulence shows the rich phenomenology of the flame turbulence interaction. The very high resolution simulations highlight the presence of vortex tubes. Across the flame, "fire polishing" damps the smallest turbulence scales; however some of the vortex tubes are enhanced through baroclinic vorticity generation.

DNS relies on nondissipative high-accuracy schemes and is conducted without resort to subgrid modeling or the introduction of "artificial" numerical dissipation or other algorithm-stabilizing schemes. Such computations allow a degree of control in isolating specific physical phenomena that is typically inaccessible in experiments. With the recent advances in supercomputing technology and algorithms, it is now possible to perform simulations of simple flows at ranges of scales comparable or even larger than in typical laboratory experiments. Petascale computing is

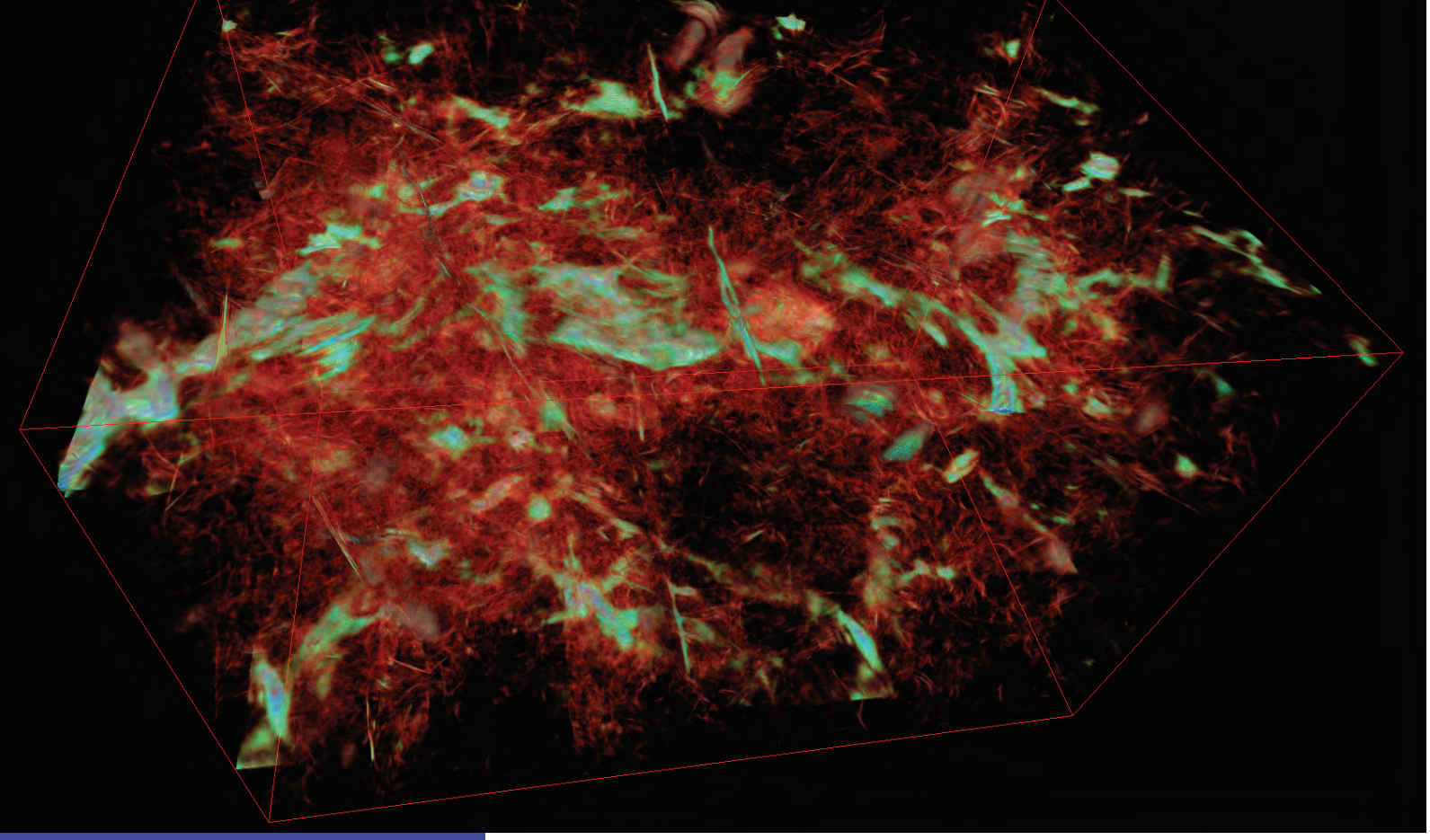


Fig. 2. Dilatational effects, usually neglected in astrophysical calculations, are important for understanding the deflagration-detonation initiation. Our results indicate the presence of shocklets, generated by the turbulence and enhanced by the increased dilatational motions due to the flame. Sufficiently strong shocklets may be the mechanism for producing detonation in type Ia supernovae.

expected to further increase the range of scales of the simulations and allow accurate calculations of more and more complex flows.

This study represents the first successful implementation of a large structured fluid dynamics code (CFDNS) on the Cell processor architecture. The CFDNS code solves the compressible and incompressible Navier-Stokes equations in 3D using high-order compact finite differences or Fourier transforms in the periodic directions. Multiple species are allowed, each with realistic material properties equation of state (EOS), as well as Cartesian, cylindrical, and spherical grid geometries. The serial single processor speedup of the Cell version of the code is approximately 30x faster than the reference Opteron version, which is reasonable when considering the clock speed, parallelism, and vectorization afforded by the Cell. Notably, the excellent performance of the individual memory controllers is responsible for this, since the low arithmetic intensity of the algorithm does not allow the actual compute power of the synergistic processing units (SPE) to be utilized to their fullest. This serial speedup prompted us to perform significant modifications to the parallel code design, which lead to overall speedup in the range of 20x compared with the Opteron-only version [6,7].

The model problem addressed by the study is the flame-turbulence interaction under the complex conditions characterizing the early stages of a type Ia supernova. These conditions are novel and have no direct analogue on earth. This makes them interesting for testing new physics, but it also means that our terrestrial intuition regarding flames can be misleading. For example, it is thought that laboratory flames in the Ka >> 1 regime simply go out because they are unable to maintain their heat in the

presence of so much turbulence. But the flame in a supernovae can never "go out" until the star comes apart and, in terms of local flame variables, that takes a very long time. Although several mechanisms for detonation have been proposed, the debate around deflagration/detonation models is still not settled. Moreover, most turbulent flame simulations so far, under these conditions, are in the low Mach number approximation and no DNS have been performed for such flows.

Using the Roadrunner supercomputer, researchers have performed the largest reacting compressible turbulence simulations to date. The flow conditions considered for the reacting simulations correspond to "well-stirred" single component burning, $^{12}C(^{12}C,\gamma)^{24}Mg$, relevant to Type Ia supernovae. The flame advances into the cold fuel (^{12}C) in a C-O mixture and leaves behind hot product (^{24}Mg). Inflow/outflow boundary conditions are imposed in the flame propagation direction. The physical transport properties are appropriate for the astrophysical situation investigated and are calculated within new modules added to the code. Thus, the thermal transport includes both radiative and electron transport (accounting for degenerate regimes) contributions. The equation of state considers radiative, ion, and electron contributions. To reduce the computational effort, precalculated tables for the transport properties, EOS, and nuclear energy rates are used.

The simulations were performed in three stages:

1) Generate 1D reacting flow profiles as initial conditions.

2) Generate inflow turbulence by performing triply periodic simulations with a background velocity matching the flame speed [8].

3) Simulate flame-turbulence interaction under supernova conditions.

To better study the flame characteristics, the reference frame was chosen such that the flame was stationary in the computational domain. The 3D reacting flow simulations were initialized using the 1D profiles and isotropic turbulence was introduced through the inflow boundary. To understand the effects of various parameters, most notably Da (Damkholer) and Ka (Karlovitz) numbers, as well as the effect of compressibility, several simulations were performed, on up to 2048^3 meshes.

There is a complicated phenomenology associated with turbulent flames under type Ia supernova conditions, from the suppression of the smallest vortex tubes due to the flame "fire polishing," but enhancement of intermediate turbulent scales (Fig. 1), to the rapid acceleration of the flame itself to large velocities, which is one of the important open questions related to the supernova modeling. In addition, the fully compressible simulations allowed considering the dynamics of the dilatational motions, neglected in previous studies. These motions are enhanced by the heat addition due to the flame and can cause shock waves that may lead to detonation (Fig. 2). Current research focuses on the departures of the turbulence properties from the classical Kolmogorov picture and determining accurate turbulent flame speeds.

References:

[1] Livescu, D. and Madnia, C.K., *Phys Fluids* **16,** 2864 (2004).

[2] Livescu, D. and Ristorcelli, J.R., *J. Fluid Mech.* **591,** 43 (2007).

[3] Livescu, D. and Ristorcelli, J.R., *J. Fluid Mech.* **605,** 145 (2008).

[4] Livescu, D., J.R. Ristorcelli, R.A. Gore, S.H. Dean, W.H. Cabot, and A.W. Cook, *J. Turb.* **10,** N13 (2009).

[5] Moin, P., and Mahesh, K., *Ann. Rev. Fluid Mech.* **30,** 539 (1998).

[6] Mohd-Yusof, J., D. Livescu, and T. Kelley, *in proceedings of ParCFD* (2009).

[7] Mohd-Yusof, J., D. Livescu, and T. Kelley, *in proceedings of CSE* (2009).

[8] Petersen, M.R., and D. Livescu, (in press) *Phys. Fluids (*2010).

The Roadrunner Universe Project:

Baryon acoustic oscillations in the intergalactic medium

Salman Habib[1], Adrian Pope[1], Zarija Lukić[1], David Daniel[1], Patricia Fasel[1], Katrin Heitmann[1], Chung-Hsing Hsu[1], Lee Ankeny[1], Graham Mark[1], Suman Bhattacharya[1], James Ahrens[1], Martin White[2], Jordan Carlson[2], and Nehal Desai[3]

[1] Los Alamos National Laboratory, [2] University of California, Berkeley, [3]Aerospace Corporation

Pictured left to right: Salman Habib, Katrin Heitmann, and Adrian Pope

Over the last two decades, critical observational advances in large-volume sky surveys carried out over a wide range of wavelengths, as well as over short time cadences, have revolutionized cosmology. Computational cosmology has emerged as an essential resource for providing detailed predictions for these observations, for providing essential data for assisting in the design of cosmological surveys, and as sophisticated tools for interpreting the final results.

Results from cosmological surveys have cemented a cross-validated cosmological "Standard Model"

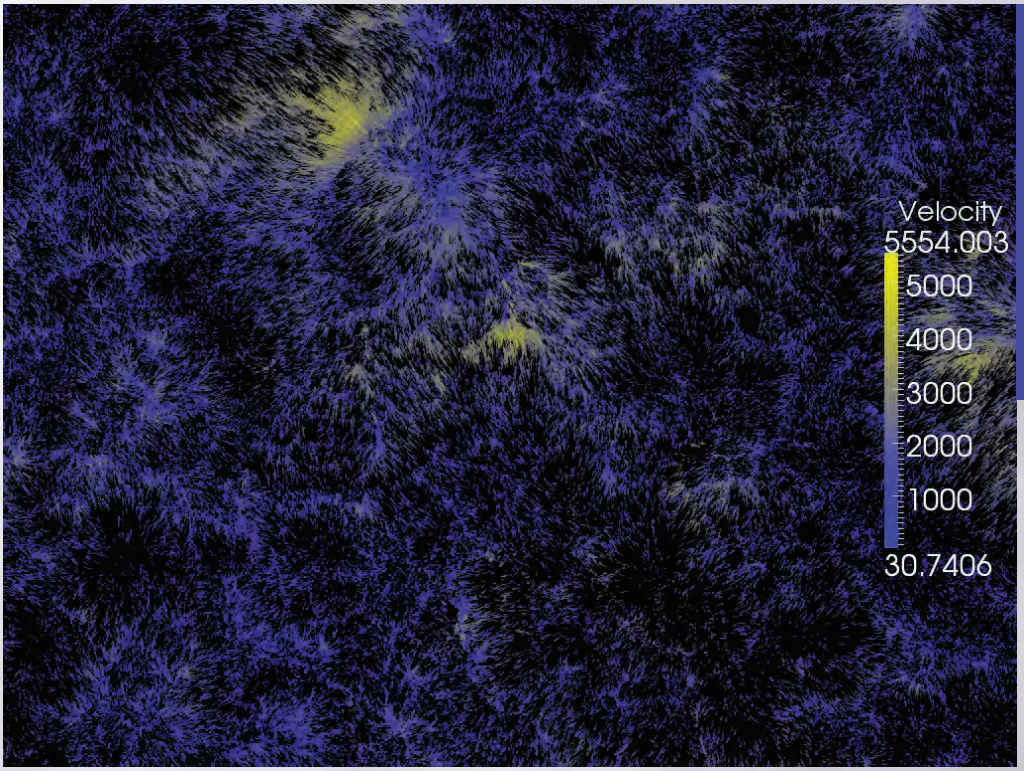


Fig. 1. Dark matter halos from one of the large Roadrunner simulations, with 1/64 of the total (750 Mpc/h)^3 volume displayed. The halos are shown as arrows, colored with respect to their velocity magnitude. This particular simulation was run with 64 billion particles, where each simulation particle has a mass of approximately one billion suns. The snapshot is taken at a redshift of z = 2.5.

presenting a comprehensive picture of the evolutionary history of the Universe and its constituents: 1) its content is dominated by 23% in dark matter which only interacts gravitationally (and a large fraction of which is localized in clumps called halos) and 72% in a smooth "dark energy" component described by a cosmological constant, 2) the initial conditions of the Universe are prescribed by adiabatic Gaussian random initial density fluctuations, and 3) the spatial geometry of the Universe is flat [1]. Although this is a great triumph, it has exposed some of the biggest puzzles in physical science:

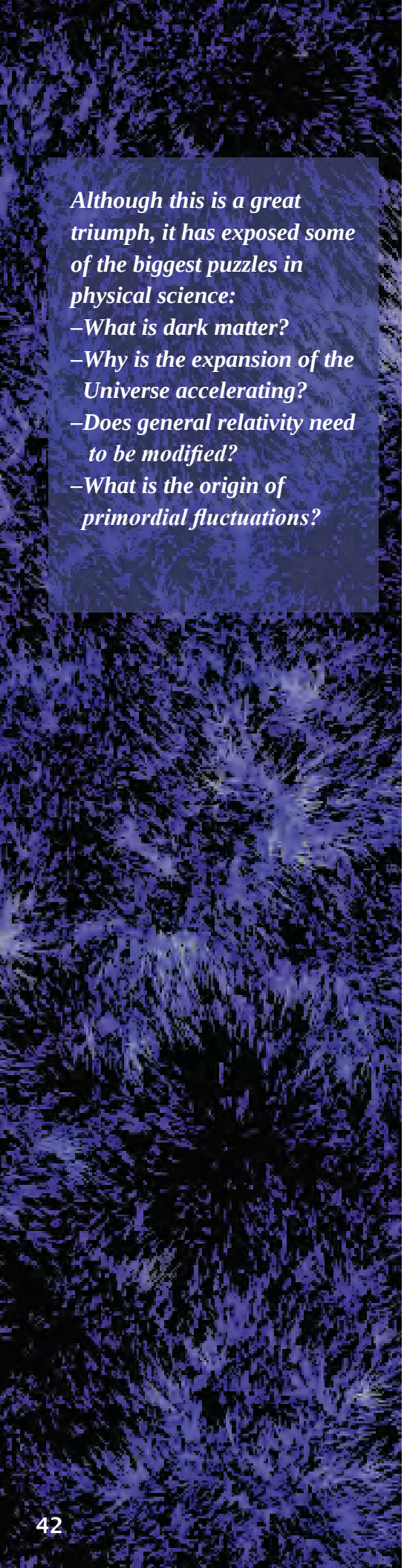
- What is dark matter?
- Why is the expansion of the Universe accelerating?
- Does general relativity need to be modified?
- What is the origin of primordial fluctuations?

To investigate these questions, the observational state of the art is rapidly advancing; surveys now coming on line and within the next decade represent an improvement in capability by roughly two orders of magnitude, translating into a determination of certain cosmological parameters at the 1% level. Remarkable as this is, the effort will only come to fruition if the accuracy of the underlying theory can be controlled to the sub-percent level. This severely demanding task will push the boundaries of computing for the foreseeable future.

Structure formation in the Universe is driven primarily by gravitational instability. Initial density perturbations collapse and merge in a hierarchical fashion to form dark matter halos within a global "cosmic web" structure (Fig.1). On scales smaller than several Mpc (megaparsec; 1 parsec=3.26 light-years), baryonic matter collects in halos, eventually forming stars and galaxies. The collisionless evolution of matter subject only to gravity is described by the Vlasov-Poisson equation in an expanding Universe, which can be solved in detail only by N-body techniques. Next-generation surveys demand simulations with multigigaparsec (Gpc) box-sizes and particle counts in the 10^{11-12} range, all with ~kpc force resolution (a force dynamic range of 10^6). An overall 2–3 orders of magnitude improvement in throughput over the current state of the art turns out to be the minimal requirement.

To meet the challenge of next-generation simulations, the Roadrunner Universe project code MC^3 (Mesh-based Cosmology Code on the Cell) splits the interparticle force problem into two parts, a medium resolution solver based on Fast Fourier Transforms (FFT) augmented by a direct particle-particle short-range solver. The biggest FFT provides up to four orders of magnitude of dynamic range, the remaining factor of 10-100 coming from the short-range force evaluations carried out on the Cell processors. The MC^3 algorithms match to the machine architecture, minimizing data transfer through the narrow communication pipe between the Opterons and the Cell. The global philosophy is to explicitly sacrifice memory and in-place computation to minimize communication and simplify communication patterns. Our approach has two key aspects:

1) Reduction of particle communication across the Cell layer using particle overloading, a mirrored particle cache.



Although this is a great triumph, it has exposed some of the biggest puzzles in physical science:
–What is dark matter?
–Why is the expansion of the Universe accelerating?
–Does general relativity need to be modified?
–What is the origin of primordial fluctuations?

2) Application of digital filtering and differencing in the spectral domain, allowing simplified computations at the Cell layer [2].

The Open Science Roadrunner Universe project (RRU) runs on Roadrunner consisted of nine ultra-large simulations to study the imprint of oscillations in the baryon-photon plasma in the early universe, the baryon acoustic oscillations (BAO). Due to BAO, a distinct, but subtle, signature is imprinted on the large-scale distribution of matter and has been seen in the spatial statistics of the distribution of galaxies [3], confirming one of the most important predictions of modern cosmology. BAO has now become one of the premier methods for determining cosmological distances and hence the expansion history of the Universe.

Traditional galaxy-based BAO surveys require a heavy investment in telescope time, especially as one goes to higher redshifts. Fortunately, there exist tracers of the mass distribution other than galaxies. Neutral hydrogen in the intergalactic medium (IGM) furnishes one such example. At redshifts of $z \cong 2\text{-}3$, the gas making up the IGM is thought to be in photoionization equilibrium, which results in a tight density-temperature relation, with the neutral hydrogen density proportional to a power of the baryon density [4]. Since pressure forces are sub-dominant, the neutral hydrogen density closely traces the total matter density on large scales. The neutral hydrogen density can be probed by obtaining spectra of distant, bright compact sources – the quasars – and studying – the celebrated "Lyman-α forest" of absorption lines which map the neutral hydrogen along the line-of-sight to the quasar. The structure in quasar absorption thus traces, in a calculable way, slight fluctuations in the matter density of the universe back along the line-of-sight to the quasar, with most of the "Lyman-α forest" arising from over-densities of a few times the mean density.

The upcoming Baryon Oscillation Spectroscopic Survey (BOSS) [5] will provide an unprecedented number of quasar spectra for Lyman-α studies, motivating a major simulation effort at understanding the BAO imprint in the IGM. The set of Roadrunner simulations [6] are the first to simultaneously resolve structure down to the Jeans scale of the gas (~100 kpc) as well as properly capture the acoustic scale (~100 Mpc). Using the results of the simulations (the density and velocity fields), mock quasar spectra were constructed by running lines of sight from quasar "sources" to an "observer." These spectra have properties close to those observed at $z \cong 2\text{-}3$. Because these mock spectra will be very useful in testing observational data pipelines, calibrating analysis tools, and in planning future projects, they have been made publicly available.

Given the quasar spectra derived from the simulations, one can compute the flux-flux correlation function, which is related to the underlying nonlinear, redshift-space, mass correlation function. The characteristic BAO "bump" signal in the flux correlation function as measured from our simulations is shown in Fig 2. Although our total simulation volume is large (nine 750 h^{-1} Mpc boxes,) covering effectively 1000 sq. deg. of sky, it is still only 10% of the area planned for BOSS. Thus the fractional errors in the flux correlation function as achieved by BOSS should be a factor of three better than in our simulations.

Once detailed simulations are available, several important effects can be studied, including key sources of systematic errors and bias in the observations. We carried out preliminary investigations of an evolving mean flux, fluctuations in the photoionization rate, and HeII reionization, which generate "extra" power on the acoustic scale and reduce the contrast of the acoustic peak. Gravitational instability produces a well-defined pattern of higher-order correlations, which is not obeyed by nongravitational contributions such as the above, allowing (in principle) a diagnostic of nongravitational physics in the forest. As an example, we demonstrated that the three-point cross-correlation function in models with HeII reionization has a different scale dependence than the three-point function in gravity-only simulations, regardless of the equation-of-state assumed in the latter.

As one part of the Roadrunner Universe project, future BAO simulations will be run with much larger boxes for an improved treatment of planned surveys. Additionally, the quasars will not be randomly placed in the simulation box but will follow the appropriate statistical occupation distribution for being hosted by dark matter halos.

Acknowledgement: We thank the LANL Roadrunner support team for its outstanding effort in helping us to complete the simulation runs for this project.

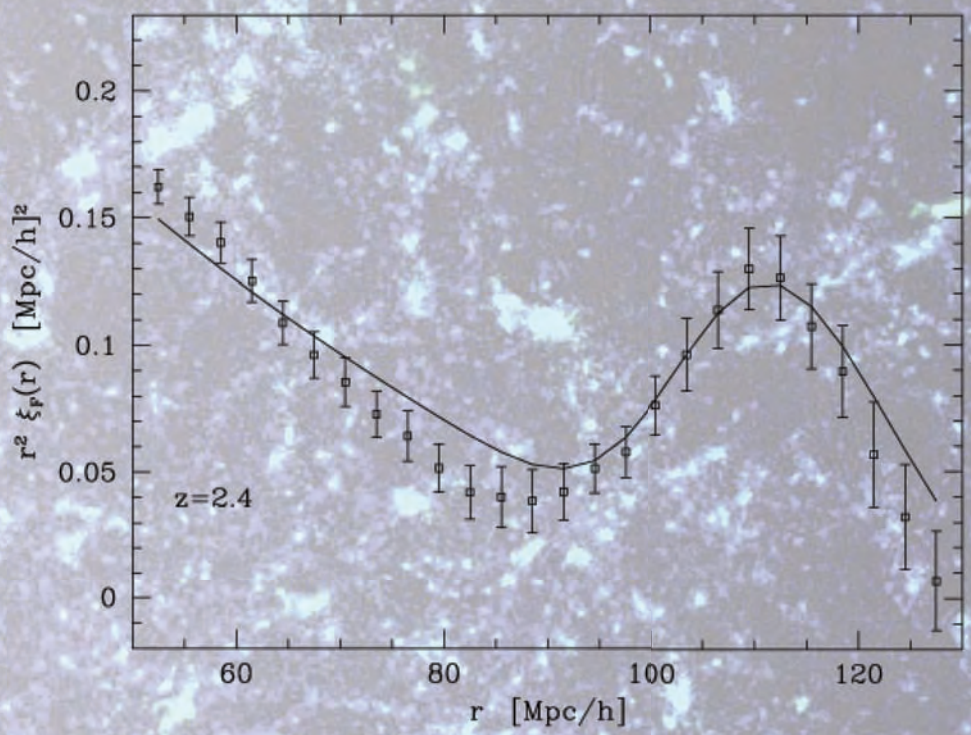


Fig. 2. The redshift space flux correlation function, ξ_F, as a function of the comoving distance, r, measured in Mpc/h. The BAO feature is the "bump" centered around 110 h^{-1} Mpc. The error bars are derived from a bootstrap analysis. The solid line is a Gaussian-smoothed linear theory result multiplied by a scale-independent bias to match the simulations.

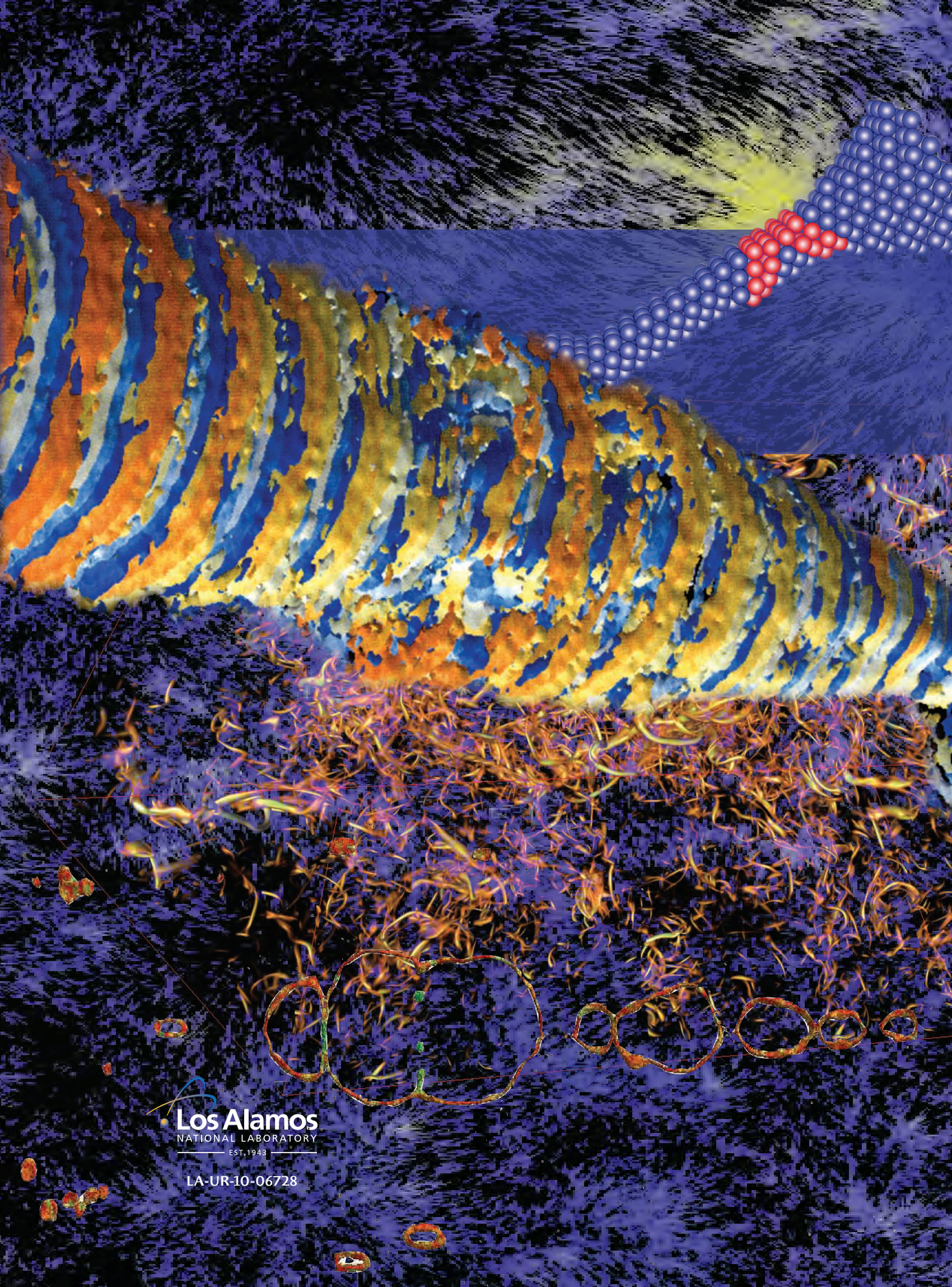
References

[1] Komatsu, E., J. Dunkley, M.R. Nolta, C.L. Bennett, B. Gold, G. Hinshaw, N. Jarosik, D. Larson, M. Limon, L. Page, D.N. Spergel, M. Halpern, R.S. Hill, A. Kogut, S.S. Meyer, G.S. Tucker, J.L. Weiland, E. Wollack, and E.L. Wright [WMAP Team], *Astrophys. J. Supp.* **180,** 330 (2009).
[2] Pope, A, S. Habib, Z. Lukić, D. Daniel, P. Fasel, N. Desai, and K. Heitmann, *Comp. Sci. Eng.* **12**, 17 (2010).
[3] Eisenstein, D.J., I. Zehavi, D.W. Hogg, R. Scoccimarro M.R. Blanton, R.C. Nichol, R. Scranton, H.-J. Seo, M. Tegmark, Z. Zheng, S.F. Anderson, J. Annis, N. Bahcall, J. Brinkmann, S. Burles, F.J. Castander, A. Connolly, I Csabai, M. Doi, M. Fukugita, J.A. Frieman, K. Glazebrook, J.E. Gunn, J.S. Hendry, G. Hennessy, Z. Ivezić, S. Kent, G.R. Knapp, H. Lin, Y.-S. Loh, R.H. Lupton, B. Margon, T.A. McKay, A. Meiksin, J.A. Munn, A. Pope, M.W. Richmond, D. Schlegel, D.P. Schneider, K. Shimasaku, C. Stoughton, M.A. Strauss, M. SubbaRao, A.S. Szalay, I. Szapudi, D.L. Tucker, B. Yanny, and D.G. York, *Astrophys. J.* **633,** 560 (2005).
[4] Meiksin, A.A., *Rev. Mod. Phys.* **81,** 1405 (2009).
[5] Schlegel, D., M. White, and D. Eisenstein, arXiv:0902.4680 [astroph.CO].
[6] White, M., A. Pope, J. Carlson, K. Heitmann, S. Habib, P. Fasel, D. Daniel, and Z. Lukić, *Astrophys. J.* (submitted); arXiv:0911.5341 [astro-ph.CO].

Glossary

Term	Definition
1D, 2D, 3D	One-, Two-, or Three-Dimensional
AIDS	Acquired Immune Deficiency Syndrome
ALF	Accelerated Library Framework
AoS	Array of Structures
API	Application Programming Interface
ASC	Advanced Simulation and Computing program (NNSA)
BAO	Baryon Acoustic Oscillation
BOSS	Baryon Oscillation Spectroscopic Survey
CBE	Cell Broadband Engine
Cell	Cell Broadband Engine
CFDNS	DNS code used by the CCS Fluid Dynamics Team at LANL
CPU	Central Processing Unit
CU	Connected Unit
Da	Damkholer number, used to relate a chemical rection timescale to other phenomena occuring in a system
DaCS	Data Communication and Synchronization
DMA	Direct Memory Access
DNS	Direct Numerical Simulation
DOE	Department of Energy
DRAM	Dynamic Random Access Memory
EAM	Embedded Atom Method
EOS	Equation of State
EPW	Electron Plasma Wave
FFT	Fast Fourier Transform
Flop	Floating Point Operation
GB	Gigabytes
Gpc	Gigaparsec
GPU	Graphical Processing Unit
HIV	Human Immunodeficiency Virus
HRTEM	High Resolution Transmission Electron Microscope
I/O	Input/Output
ICF	Inertial Confinement Fusion
IEEE	Institute of Electrical and Electronics Engineers
IGM	Intergalactic Medium
IMC	Implicit Monte Carlo
Ka	Karlovitz number, used in turbulent combustion and corresponds to the ratio of the chemical time scale and smallest turbulent time scale

KB	Kilobyte
kpc	Kiloparsec
LANL	Los Alamos National Laboratory
LCLS	Linac Coherent Light Source
libspe	SPE library
LINPACK	Benchmark linear algebra code
LLNL	Lawrence Livermore National Laboratory
LPI	Laser-Plasma Instabilities
LS	Local Store
μs	Microsecond
MaRIE	Matter-Radiation Interactions in Extremes
MC^3	Mesh-based Cosmology Code on the Cell
MD	Molecular Dynamics
Mpc	Megaparsec
MPI	Message Passing Interface
MRX	Magnetic Reconnection eXperiment
nm	Nanometer
NIF	National Ignition Facility
NNSA	National Nuclear Security Administration
ORNL	Oak Ridge National Laboratory
ParRep	Parallel Replica Dynamics
PCIe	Peripheral Component Interconnect Express bus
petaflop/s	10^{15} floating point operations per second
PPE	PowerPC Processing Element
PPPL	Princeton Plasma Physics Laboratory
ps	Picosecond
RMI	Richtmyer-Meshkov Instability
RR	Roadrunner
RRU	Roadrunner Universe Project
SIMD	Single Instruction, Multiple Data
SMF	DMA Memory Engine
SoA	Structure of Arrays
SPaSM	Scalable Parallel Short-range Molecular Dynamics code
SPE	Synergistic Processing Element
SPU	Synergistic Processing Unit
SRS	Simulated Raman Scattering
STI	Sony, Toshiba, and IBM
Sweep3D	A neutron transport kernel
teraflop/s	10^{12} floating point operations per second
TPMI	Trapped Particle Modulation Instability
VPIC	Particle-in-cell (plasma physics) code



Los Alamos
NATIONAL LABORATORY
EST. 1943

LA-UR-10-06728