The scaling issue: scientific opportunities

Raymond L. Orbach
3001 Veazey Terrace NW, #525, Washington DC 20008
E-mail: rorbach@earthlink.net

Abstract. A brief history of the Leadership Computing Facility (LCF) initiative is presented, along with the importance of SciDAC to the initiative. The initiative led to the initiation of the Innovative and Novel Computational Impact on Theory and Experiment program (INCITE), open to all researchers in the U.S. and abroad, and based solely on scientific merit through peer review, awarding sizeable allocations (typically millions of processor-hours per project). The development of the nation’s LCFs has enabled available INCITE processor-hours to double roughly every eight months since its inception in 2004. The “top ten” LCF accomplishments in 2009 illustrate the breadth of the scientific program, while the 75 million processor hours allocated to American business since 2006 highlight INCITE contributions to U.S. competitiveness. The extrapolation of INCITE processor hours into the future brings new possibilities for many “classic” scaling problems. Complex systems and atomic displacements to cracks are but two examples. However, even with increasing computational speeds, the development of theory, numerical representations, algorithms, and efficient implementation are required for substantial success, exhibiting the crucial role that SciDAC will play.

1. Introduction and background

The stunning achievement of the Earth Simulator’s official opening on March 11, 2002 changed the approach to high end computation forever. Not only did its peak speed of 40 TeraFlops eclipse any other supercomputer, but its efficiency on real scientific problems (e.g., the Atmospheric General Circulation Model) was in the double digit range (65% in this case). The fastest U.S. computers, based on commodity chips, not only had substantially lower peak speeds, but also exhibited efficiencies only in the single digit range. As a consequence, the Earth Simulator sustained speeds were nearly two orders of magnitude faster than for any U.S. computer.

I was rather shaken by the gulf in capability between the Earth Simulator and anything deployed in the United States. As luck would have it, I was invited to give the Commencement Address at the RAND Graduate School on June 22, 2002 [1]. The title of my talk, inspired by the Earth Simulator, was “Computation for the 21st Century: A New Paradigm.” I wrote:

We are now in an era where computational simulation can inform our approach to science, and I believe the social sciences and humanities. We are now able to contemplate exploration of worlds never before accessible to mankind … we can simulate systems to discover physical laws for which there are no known predictive equations.… This approach to understanding complex systems is to be thought of in the same vein as experiment and analytic theory. In science of the 21st century, simulation and high-end computation are equal partners with theory and experiment … this critical third leg of the triad.
I also wrote:

Bluntly put, we are out of business in some critical areas of computational science…. What does it mean to lose scientific leadership, to be #2?… The great scientific discoveries will take place in Yokohama, Japan, and not here…. I believe our country cannot afford to be second best, to be “good enough.” Our economy, our intellectual environment, literally our national security depends upon our scientific primacy…. The United States needs to face up to our dilemma, a construct of our own doing…. As a product of the Sputnik generation, I can personally attest to the vigor and vitality of the U.S. response. It is now incumbent upon us to repeat this dedication in this new era of computnik, to regain our scientific leadership and primacy.

The ensuing seven years (almost to the date of this conference) can only be described as a sea change. I believe it is fair to say that a revolution in thinking and performance has taken place. The policy of the Office of Science in 2002, in my view, was inward looking. Though we spoke about access, in reality the only way in which a researcher could gain time on our computers was either directly through a DOE contract or grant, or in combination with a researcher supported by DOE.

On top of this insularity, there were so many DOE researchers applying for time on our one supercomputer, the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory, that the number of cycles available for any specific researcher was limited. From its web site, “… what distinguishes NERSC is its success in creating an environment that makes these [scientific computing facilities] effective for scientific research.” That was true then, and continues to be true. But nevertheless, it was a capacity machine, lacking the capability that was exhibited by the Earth Simulator.

All was not bleak, however. A year before the Earth Simulator, the Scientific Discovery through Advanced Computing (SciDAC) program began as a five-year program “to develop the scientific computing software and hardware infrastructure needed to use terascale computers to advance DOE research programs in basic energy sciences, biological and environmental research, fusion energy sciences, and high-energy and nuclear physics.” Serendipitously, I had included in my RAND Graduate School address:

We need to begin with the science, and scientists…. Bring to the table the computer scientists, the applied mathematicians, those who are good at algorithms, together with the chip makers and computer architects…. We shall be able to investigate systems of great complexity, and understand predictive laws for their behavior. We will … learn the rules of collective behavior on a scale previously unknown.

Putting all this together, in 2002:

- The Earth Simulator outpaced U.S. computers in sustained speed by almost two orders of magnitude;
- In practice, only DOE supported researchers had access to DOE computers;
- Users received only small portion of the cycles required to address their problem of choice;
- SciDAC had begun a year earlier on a five-year trial basis.

The Office of Science addressed these four issues directly. Though grid computation was popular at the time, we recognized that a large fast mainframe was the only route to rapid sustained speeds. Time has shown that this was the correct choice. The Department of Energy put in place a U.S. competition for a National Leadership Computing Facility. A peer reviewed process led to investment in the first Leadership Computing Facility at Oak Ridge National Laboratory in 2004. A second Leadership Computing Facility was established at Argonne National Laboratory in 2006. These facilities are designated as capability facilities, while the facility at NERSC, along with that at Pacific Northwest National Laboratory, are designated as a capacity facilities.

All Office of Science computational facilities were opened to researchers worldwide since 2003. Allocations are based solely on the quality of the proposal. Industry is encouraged to make use of the DOE leadership class computational facilities.

In order to advance scientific discovery, the number of users at the capability facilities is limited. Allocations are typically in the millions of processor hours. The Leadership Computing Facilities at
Oak Ridge and Argonne National Laboratories will provide over one billion processor hours this year to the INCTE program.

Finally, SciDAC was recognized as an essential element for efficient use of leadership class computing facilities, and its first five-year existence has been extended indefinitely.

What has happened is truly astonishing [2]. “Long term partnerships have been forged, exciting science is being done, and high-performance simulation on leadership-class computers is keeping the United States at the forefront of science and technology.” A panel of computational scientists, applied mathematicians, and computer scientists gathered together in February 2008 at the invitation of the Director of the Office of Advanced Scientific Computer Research, Dr. Michael Strayer, to identify recent breakthroughs in computational science and enabling technologies supported by INCITE, SciDAC and/or the base program. They identified the 10 most significant science accomplishments over the past 18 months, singular achievements that tell an incredible story of computational discovery [3], exhibited in figure 1.

| Rank | Title |
|------|-------|
| 1    | Modeling the Molecular Basis of Parkinson’s Disease (Tsigelny) |
| 2    | Discovery of the Standing Accretion Shock Instability and Pulsar Birth Mechanism in a Core-Collapse Supernova Evolution and Explosion (Blondin) |
| 3    | Prediction and Design of Macromolecular Structures and Functions (Baker) |
| 4    | Understanding How Lifted Flame Stabilized in a Hot Coflow (Yoo) |
| 5    | New Insights from LCF-enable advanced kinetic simulations of global turbulence in fusion systems (Tang) |
| 6    | High Transition Temperature Superconductivity: A High-Temperature Superconducting State and a Pairing Mechanism in 2-D Hubbard Model (Scalapino) |
| 7    | PETSc: Providing the Solvers for DOE High-Performance Simulations (Smith) |
| 8    | Via Lactea II, A Billion Particle Simulation of the Dark Matter Halo of the Milky Way (Madau) |
| 9    | Probing the properties of water through advanced computing (Galli) |
| 10   | First Provably Scalable Maxwell Solver Enables Scalable Electromagnetic Simulations (Kovel) |

**Figure 1.** Top 10 Computational Science Accomplishments. INCITE: 1, 3, 4, 8. SciDAC: 2, 5, 6, 7, 9, 10.

U.S. industry has successfully competed for INCITE resources. Over 75 million processor hours have been allocated to American business since 2006. Some of their accomplishments are exhibited in figure 2.

The combination of new architectures and SciDAC programs has allowed INCITE to rapidly increase its allocations, from 4.9 million processor hours in 2004 to 1.3 billion processor hours in 2010. Figure 3 exhibits the actual annual INCITE allocations from 2004 through 2010, and then extrapolates the allocations through 2015.
• General motors (25 million processor hours in 2008 and 2009): thermoelectric materials capable of turning waste heat into electricity
• Proctor & Gamble (11.1 million processor hours in 2007-2009): bubble formation, dynamics, and stability for dissolving of soap and forming of suds
• General Atomics (5.4 million processor hours from 2006-2009): workings of turbulence in a fusion reactor
• Pratt & Whitney (>4.5 million processor hours from 2006-2009): computational fluid dynamics to better understand gas-turbine engines, improve fuel efficiency, control emissions, reduce noise, and boost operability
• DreamWorks Animation (1.8 million processor hours in 2006 and 2007): pushing the limits of real-time ray tracing to improve computer graphics imagery
• Corning Inc. (3.2 million processor hours in 2007-2009): simulation of dense suspensions, calculations of glass transition behavior for silica and silicate glasses
• Boeing (1.8 million processor hours from 2006-2009): next-generation aero-elastic wing simulation capabilities using computational fluid dynamics
• Gene Network Sciences (22,200,000 processor hours in 2009): simulating cardiac rhythm disturbances to identify underlying electrical mechanisms for arrhythmias
• Fluent Inc. (>300,000 processor hours in 2007 and 2008): large-scale engineering simulation early in the process of automotive design

Figure 2. Selection of American business use of INCITE, and accomplishments, since 2006.

Figure 3. Number of projected INCITE processor hours allocated each year from 2004–2015: ♦ actual, ■ extrapolated. INCITE in 2010 uses only Leadership Class Facilities at ORNL and ANL.
The line through the points in figure 3 can be written as,

\[ \log_{10}(\text{processor-hours}) = 0.47 \text{years}, \]

corresponding to a doubling of processor hour allocations every 7.7 months. This doubling rate is actually three times faster than Moore’s law. Hopefully, it will have the same impact [4]: “the driving force of technological and social change in the late 20th and early 21st centuries.”

2. Scientific opportunities

With the current peak speeds (~1.3 PetaFlops at ORNL) and INCITE allocations, and the prospect of ExaFlop peak speeds within the next decade [5], it is important to think of the scientific opportunities that can be addressed. There have been three workshops that addressed this issue [6], and I would like to focus one of the most exciting prospects – scaling. There are two aspects in particular to which I would like to draw attention: scaling in complex systems, and scaling from atomic displacement to cracks. While important advances have been made in both fields, they have been hampered by computational constraints that can be rectified. As I shall discuss, this will involve both increases in computational speeds and in SciDAC advances.

The fundamental problem with advance in scaling computations is computational cost. If the cost is proportional to \( O(N^3) \), then a computer that is 1000X faster can only run a calculation of 10X larger. As we shall see, the two reasonably typical problems cited in this paper require at least 8 orders of magnitude improvement over current scaling ranges. As noted by R.J. Harrison [7], some combination of theory, numerical representation, algorithm development, and efficient implementation are all required to extend current scaling time limits to laboratory measurement times (equivalently, current scaling length scale limits to laboratory measurement length scales). An example of progress along these lines can be found in the paper titled “Extending the time scale in atomistic simulation of materials” by Voter et al. [8].

2.1 Scaling in complex systems

“Spin Glasses” [9] are deceptively simple. The stereotypical example is a dilute magnetic alloy, say 6% of Mn dissolved randomly in a Cu matrix: Cu:Mn 6 at.%. The characteristic individual exchange spin flip time is of the order of \( 10^{-12} \) sec, but the characteristic times for reaching equilibrium as measured in the laboratory are greater than the age of the universe. This is thought to be caused by a highly degenerate set of ground states well separated in phase space so that “communication” between states is strongly hindered.

The localized magnetic moments (here Mn spins) are randomly spatially distributed and are dilute, and there is no net moment in the absence of a magnetic field. This randomness causes the usual two spin correlation function to vanish when ensemble averaged over allowable configurations. Instead, a four spin magnetic correlation function is introduced [10]

\[ G(x, t) = V^{-1} \sum_i \langle \sigma_{i+x, \tau}, \sigma_i, \sigma_i \rangle, \]

where the brackets \( \langle \rangle \) indicate a thermal average, \( V \) is the volume of the sample, \( x \) is the distance from a spin at site-label \( i \) to a spin at site-label \( i + x \), \( t \) is the evolution time for the system, \( \sigma_i \) is the Ising magnetization at the \( i \)th site for realization \( \sigma \), \( \tau_i \) is the Ising magnetization at the \( i \)th site for realization \( \tau \), and the overbar represents the ensemble average over quenched states.

The calculations are performed using Monte Carlo methods, so the time \( t \) equates to the number of Monte Carlo cycles. The characteristic time for Cu:Mn 6 at. % is \( 2.44 \times 10^{-13} \), corresponding to a single Monte Carlo cycle. The maximum number of Monte Carlo cycles in [10] was \( 10^6 \), equivalent to a
maximum physical time of $2.44 \times 10^{-7}$ sec, while laboratory time scales [11] are typically between $10^2$ sec and $10^4$ sec, or some $10^8$ to $10^{12}$ times longer! The Monte Carlo results are well represented by the functional form,

$$G(x,t) = \frac{A(T)}{x^\alpha} \exp \left\{ - \frac{x}{\xi(T,t)} \right\}^\delta,$$

with $\alpha = 0.5$ and $\delta = 1.48$, and $A(T)$ a temperature dependent constant. The correlation length $\xi(T,t)$ fits to power law dynamics,

$$\xi(T,t) = B(T) t^{\lambda(T)},$$

with $\lambda(T) = 0.16 T$. Experimentally [11], at $B(T) = 0.653$ at $T = 0.72 T_g = 28 K$, where $T_g$ is the “glass temperature” where the transition to the spin-glass state takes place ($T_g = 31.5 K$ for Cu:Mn 6 at.%), and $\lambda(T) = 0.17 T$, in remarkable agreement with the Monte Carlo results, especially because the time scales between theory and experiment differ by at least 8 orders of magnitude.

A competing theory [12], the so-called “droplet model”, uses activated dynamics with

$$\xi(T,t) = C(t^{\lambda(T)})^{1/0.21},$$

that also fits experiment, but with a very small coefficient $C = 10^{-5}$. Figure 4 exhibits the prediction of both theories, along with the experimental results ($N_s \approx \xi^2$).

![Figure 4](image)

**Figure 4.** A plot of the number of spins within the correlation length $\xi$ as a function of time. The solid curve drawn through the points is the prediction for power law dynamics, while the dashed curve is the prediction for activated dynamics.

### 2.2. Scaling from atomic displacements to cracks

Radiation damage is inherently multiscale with interacting phenomena ranging from picoseconds to decades, and nanometers to meters. The difficulty is that first principle simulations are limited to hundreds of atoms for tens of picoseconds. Figure 5 exhibits the requisite length and time scales for simulations to overlap experiment. If one drops the accuracy to the level of empirical interatomic
potentials, millions of atoms for microseconds can be simulated [13]. However, this still falls far short of the length scale and especially the time scale of experiments.

Figure 5. Radiation damage: time and length scales for mechanical property experiments, and for theoretical methods (from [13]).

Perhaps the simplest simulation with relevance to experiment is that of crack propagation in silicon. Instead of trying to simulate the entire system with a common level of accuracy, the atoms are divided into different regions where accurate descriptions are feasible. Of course, the difficulty occurs at the boundaries between these regions. Figure 6 exhibits the geometrical decomposition of a propagating crack tip in a silicon slab into three distinct dynamic regions [14]: a continuum region where finite-element description is accurate (FE), a region describable by atomistic molecular-dynamics (MD), a quantum tight-binding region (TB), and the two “handshaking” regions between: FE-MD and MD-TB. This approach enables a quantum mechanical treatment of the crack tip, where its use is critical, and a mean-field treatment well away from the crack tip where a quantum mechanical treatment is far too costly in computer time, and not necessary.

New methods are being developed that offer promise to increase the number of atoms in an atomistic description of diffusion dynamics [15]. The eventual aim is to describe accurately more complex materials so that simulation can be used to describe materials properties in real systems. For example, understanding irradiation effects on reactor pressure vessels is critical for nuclear reactor design and licensing [16]. A common alloy is Fe-C(0.05-0.2%)-Mn(0.7-1.6%)-Ni(0.2-1.4%)-Si(0.2-0.6%)-Cr(0.05-0.5%)-Cu(0.05-0.4%)-P(0.005-0.025%). Neutron collisions result of hundreds of displacements per atom over the life of the pressure vessel. Reactions take place on the scale of the nuclear to the size and service lifetimes of structural components, spanning factors in excess of $10^{14}$ (length) and $10^{22}$ (time). The problem is really one of [16]
… “multi-physics” involving numerous basic nuclear, atomic and solid-state physics processes linked to complex nano and microstructural evolutions in multi-constituent, multi-phase engineering materials through non-equilibrium thermodynamics and accelerated kinetics, leading to structure-property and property-property relations described by micro and macro mechanics models.

Figure 6. The geometrical decomposition of a periodic thin (10.9Å) Si slab with a thin crack exposing (100) faces placed under uniaxial tension imposed by initializing a velocity gradient along the pulling direction and constraining the outermost surfaces to move at constant speed [14]. The slab is divided into five different dynamic regions for the simulation: the continuum finite-element region (FE); the atomistic molecular-dynamics region (MD); the quantum tight-binding region (TB); the FE-MD “handshaking” region; and the MD-TB “handshaking region [14]. The image is the simulated silicon slab, with expanded views of the FE-MD (orange notes-blue atoms) interface and the TB (yellow atoms) region surrounded by MD (blue) atoms. Note that the TB region surrounds the crack tip with broken-bond MD atoms trailing behind this region. The acronym, MAAD, implies “macroatomistic ab-initio dynamics.”

3. Summary
The history and future of INCITE, and the concomitant importance of SciDAC, promises to address a most recalcitrant problem in science: scaling. The dynamics of complex systems and crack formation and propagation are but two examples where both must work together, even with the prospect of remarkable advances in computational speeds. Theory, numerical representation, algorithm
development, and efficient implementation are necessary ingredients for scientific discovery in the age of progression from PetaFlop to ExaFlop computational speeds.

References
[1] Orbach R 2002 Computation for the 21st century: a new paradigm RAND Graduate School Commencement Address (available from rorbach@earthlink.net)
[2] SciDAC-2: Advancing SCIENCE through Large-Scale MODELING and SIMULATION (SciDAC Review, Special Issue 11 2009) p 8
[3] SciDAC-2: Advancing SCIENCE through Large-Scale MODELING and SIMULATION (SciDAC Review, Special Issue 11 2009) p 33
[4] Moore’s law – Wikipedia, the free encyclopedia p 1
[5] SciDAC-2: Advancing SCIENCE through Large-Scale MODELING and SIMULATION (SciDAC Review, Special Issue 11 2009) p 5
[6] Exascale for Energy, Ecological Sustainability and Global Security (E³SGS) Town Hall Meeting: 2007 April 17-18 Lawrence Berkeley National Laboratory; 2007 May 17-18 Oak Ridge National Laboratory; 2007 May 31-June 1 Argonne National Laboratory (available from www.sc.doe.gov/ascr/WorkshopsConferences)
[7] Harrison R J 2009 (American Chemical Society presentation March 26)
[8] Voter A F, Montalenti F and Germann T C 2002 Annu. Rev. Mater. Res. 46 321
[9] Binder K and Young A P 1986 Rev. Mod. Phys. 58 801; Mydosh J A 1993 Spin glasses, an Experimental Introduction (London: Taylor & Francis); Mezard M, Parisi G and Virasoro M A 1987 Spin Glass Theory and Beyond (Singapore: World Scientific)
[10] Marinari E, Parisi G, Ruiz-Lorenzo J and Ritort F 1996 Phys. Rev. Lett. 76 843; Kisker J, Santen L, Schreckenberg M and Rieger H 1996 Phys. Rev. B 53 6418
[11] Joh Y G, Orbach R, Wood G G, Hammann J and Vincent E 1999 Phys. Rev. lett. 82 438
[12] Wirth B D, Odette G R, Maroudas D and Lucas G E 1997 J. Nucl. Mater. 244 185; private communication
[13] Bernstein N, Kermode J R and Csányi G 2009 Rep. Prog. Phys. 72 026501
[14] Broughton J Q, Abraham F F, Bernstein N and Kaxiras E 1999 Phys. Rev. B 60 2391
[15] Csányi G, Albaret T, Payne M C and De Vita A 2004 Phys. Rev. Lett. 93 175503
[16] Odette G R and Wirth B D 2005 Handbook of Materials Modeling (Berlin: Springer) pp 999-1037