LA-UR- 11-05888

Approved for public release;
distribution is unlimited.

Title:	Simulation for Predictive Science: The Promises and the Challenges of Exascale Computing
Author(s):	Cheryl L. Wampler, XCP-ASC Andrew B. White, ADTSC
Intended for:	Six Lab Conference on Engineering and Materials at Extreme Conditions October 23-28, 2011

Form 836 (7/06)

Simulation for Predictive Science: The Promises and the Challenges of Exascale Computing

Six Lab Conference on Engineering and Materials at Extreme Conditions

October 23-28, 2011

Cheryl Wampler, Andy White

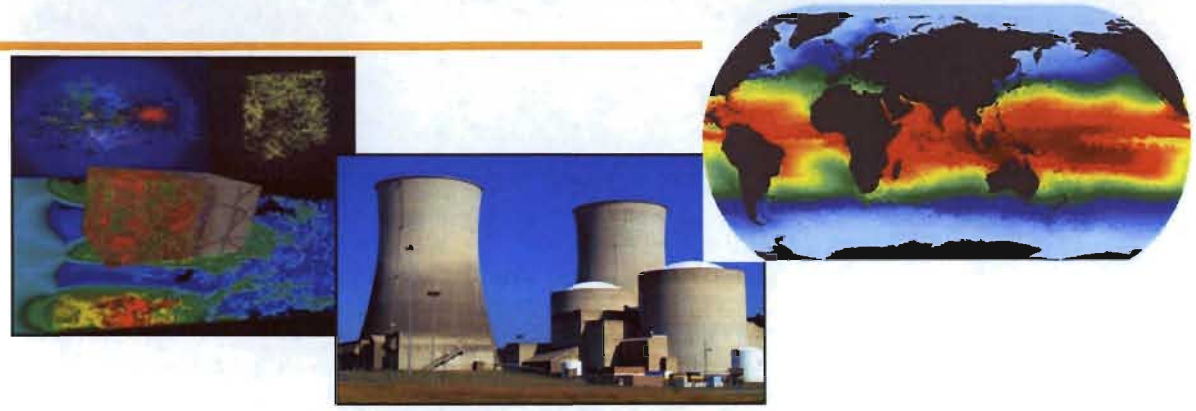
Acknowledgements

- Los Alamos National Laboratory:
  - Andy White -- John Sarrao
  - Bob Webster -- Tim Germann
  - John Hopson -- John Morrison
  - Paul Henning
- Lawrence Livermore National Laboratory:
  - Terri Quinn
- Sandia National Laboratories:
  - Justine Johannes

Simulation for Predictive Science: The Promises and the Challenges of Exascale Computing

- Mission Needs for Exascale Computing
- Exascale Computing – What is it? What's the problem?
- Two Technological Approaches: Roadrunner and Blue Gene
- Strategy for Exascale in the U.S.

Our National Missions Require Exascale Computing

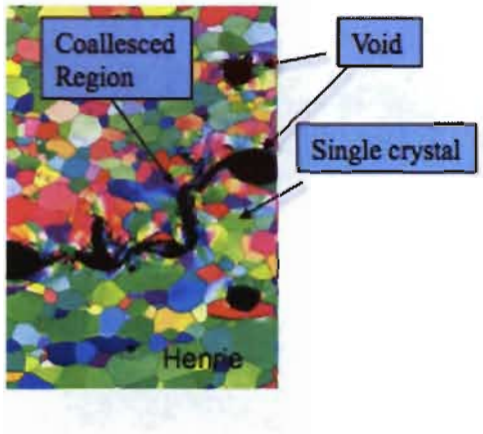


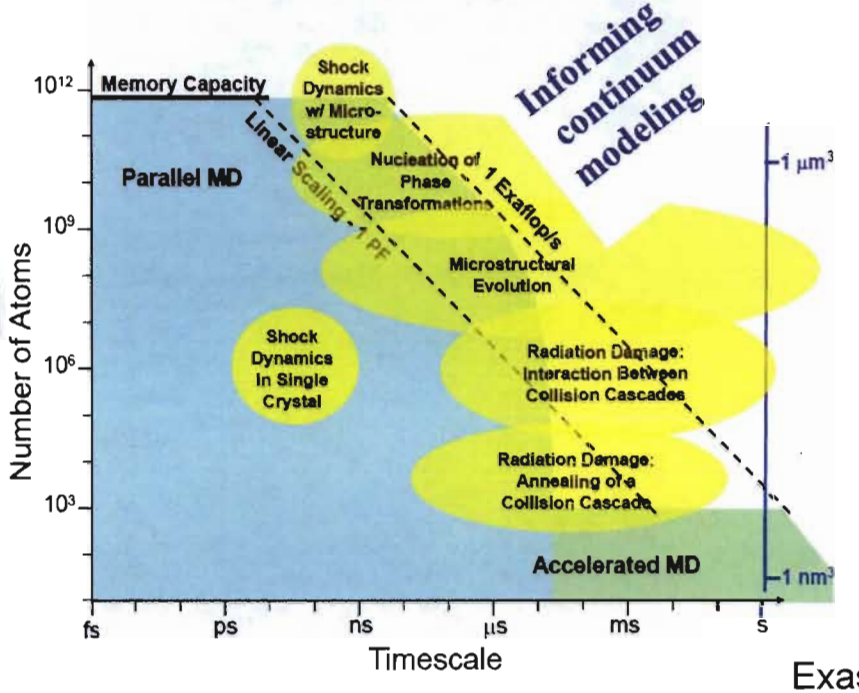
- Enable use of increasingly detailed physics models
- Advances in material modeling and scale bridging
- Enable the development of models that facilitate deduction of engineered systems behavior
- Enable validation of models requiring complex diagnostics

- ***Climate Change:*** Anticipating, mitigating and adapting to the effects of global warming
- ***Energy:*** Reducing reliance on foreign energy sources and reducing the carbon footprint of energy production
- ***National Security:*** Predictive scientific challenges to reducing global security threats
- ***Materials:*** Understanding and design of materials in extreme conditions

The goal is to make reliable scientifically based predictions in regimes that we cannot or choose not to access experimentally, and for which the consequence of error can be very large.

Next generation simulation capabilities are needed to enable discovery science at the micron frontier





Controlled fabrication, high fidelity characterization, novel *in situ* diagnostics, generation of realistic extreme environments, …

Multi-scale approaches to connect fundamental scales to bulk properties, defect generation and evolution, failure…

Exascale computing, multi-scale, multi-physics simulation tools, *ab initio* methods applied to larger, more complex materials, …

Models are integrated into a framework for reliably predicting the behavior of complex physical or engineered systems

Predictive integrated codes

Physics and Engineering Models

+

Computers

+

Experimental Verification and Validations

Simulation is the integrating element of predictive science

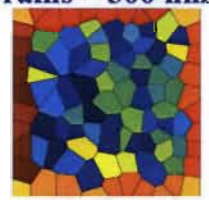
Better Understanding through Integrated Computational Engineering

- To reach a new level of understanding, we must model:
  - Manufacturing processes
  - Assembly
  - Storage and aging effects
  - Transportation history and environment
- To be predictive, the modeling must be multi-physics (e.g., thermal and mechanical) as well as multi-scale (atomistic to continuum)
- Beyond these needs, add multiple events:
  - Drops, crashes, explosive events, fires
- All include materials failure and/or decomposition
- For any one event, we need to include: QMU, Sensitivities, Optimization

Benefits:

- Ability to run complete set of studies for a given accident or event (100's of full-scale analyses)
- **Better understanding of uncertainty in computational results**
- More detailed validation with experimental results
- **Insight into failure initiation at the part level**
- **Ability to compare component designs in full-system accident scenarios**
- Ability to affect manufacturing processes and thereby create more robust system designs

Grains ~ 500 nm



Transporter ~12 m

Complete system modeling from fundamental material science to large-scale fracture and failure is an exascale problem

"Exascale Computing" means many different things…

- Several common interpretations exist:
  - A concept everybody is interested in these days
  - It refers to computing at 1 exaflop (10^{18} flop/s)
- We take it as an icon of an inevitable change in computing technology
  - Power constraints are forcing technology change
  - Our applications, regardless of scale, will have to adapt
  - Processor evolution is already in progress
- An opportunity to rethink the entire solution within a co-design framework
- A bane for the ill-prepared; An opportunity for the inventive

Power Constraints are Driving New System Designs

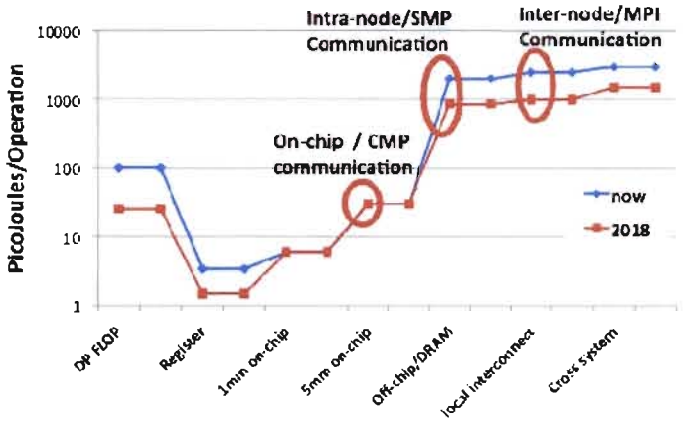
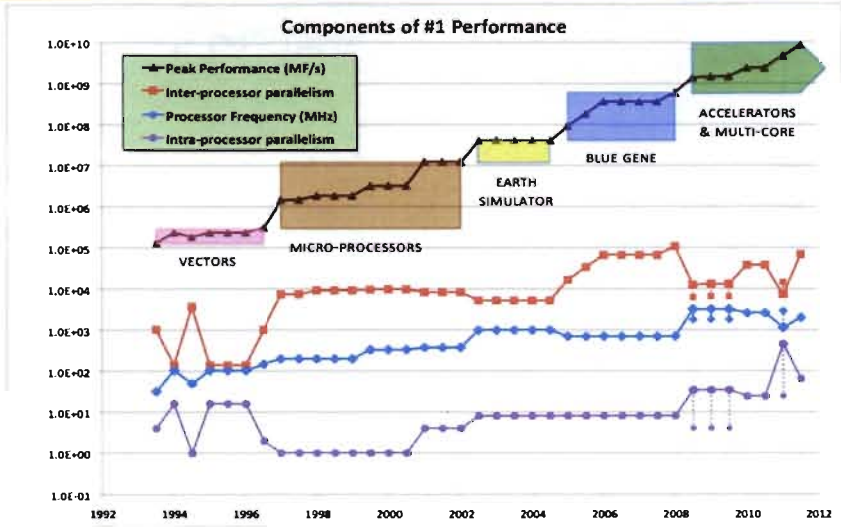


Figure 2: Energy cost of data movement relative to the cost of a flop for current and projected exascale systems



- Demand for "Moore's Law" performance growth continues, but must be realized differently
  - Will no longer be able to count on single-processor performance improvements
- Continued performance gains will only be achieved through increased parallelism
  - Multi-core is here today
- Even parallelism cannot resolve anticipated power constraints – this pushes the drive for radically new architectures - *Future of Computing Performance: Game Over or Next Level?*
- Future programming environments must support the ability to exploit data locality

Integrated performance is the product of* multiple *sub-components, which leads to a whole new system design, not just a new chip

Several significant technology challenges need to be overcome

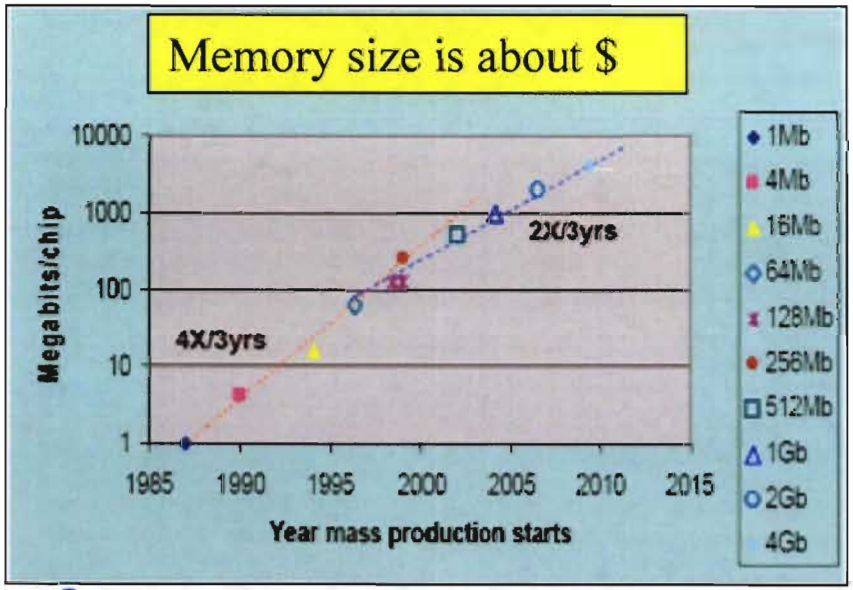
- Power will become the first class constraint on system performance and effectiveness at exa-scale, at peta-scale and at desktop-scale.
- Tomorrow's on-chip multi-processor will have an 100 – 1000x increase in parallelism; architecture is critical to meet power, performance, price, productivity & predictive goals.
- Reliability and resiliency are very difficult at this scale and require new error handling model for applications and better understanding of effects and management of errors.
- Memory is not scaling with performance and memory hierarchies will be higher and deeper.
- Operating and run-time systems will be redesigned to effectively management massive on-chip parallelism, system resiliency and power.

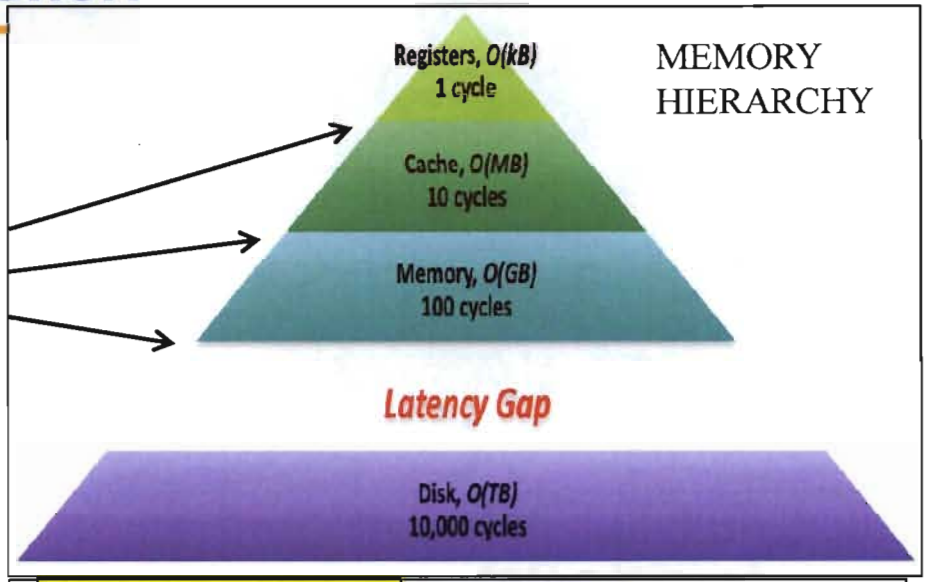
Tomorrow's programming model will be different on tomorrow's chip multi-processors, whether exascale or not. Early investment is critical to provide applications effective access to systems.

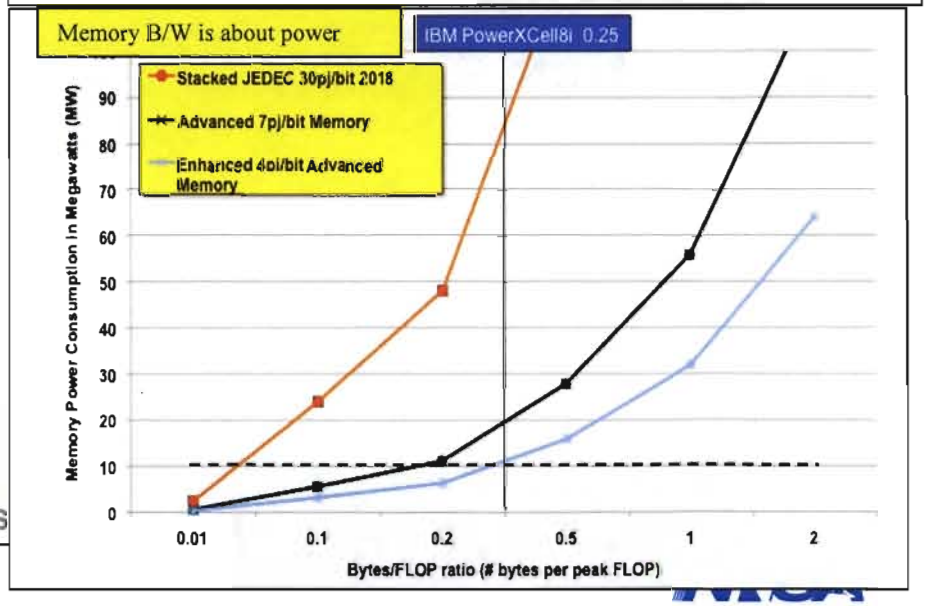
Memory size, bandwidth and hierarchy will be challenges by 2018, if not sooner.

The memory hierarchy will be much richer at the end of the decade than it is now:

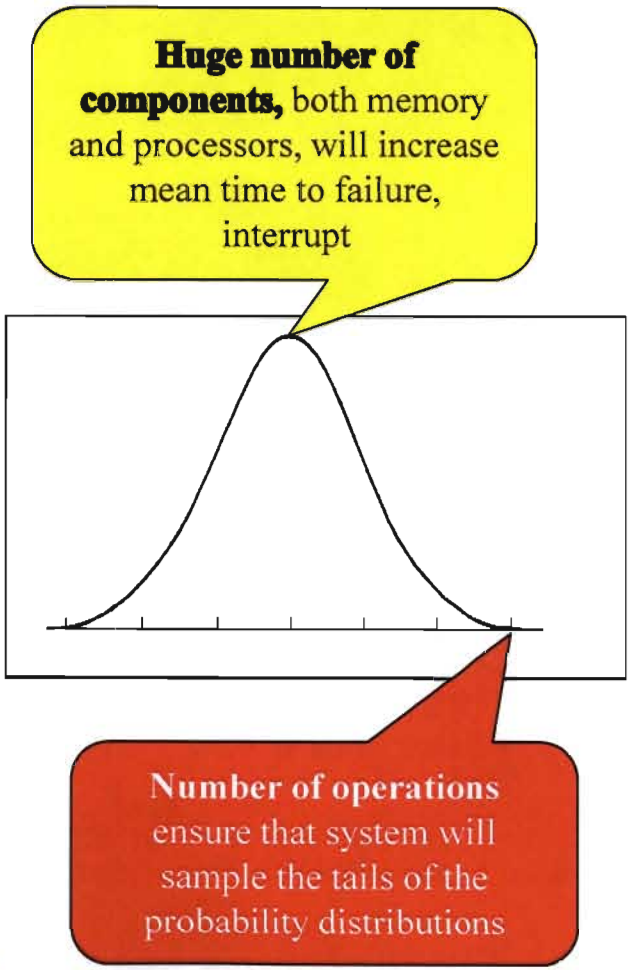
- *Software managed caches or scratch pad memory*
- *Very fast 3D stacked memory*
- *NVRAM for check-pointing and extended memory*







Resiliency issues will affect hardware, software and perhaps even applications

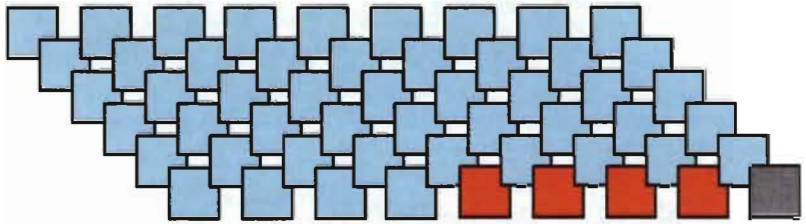


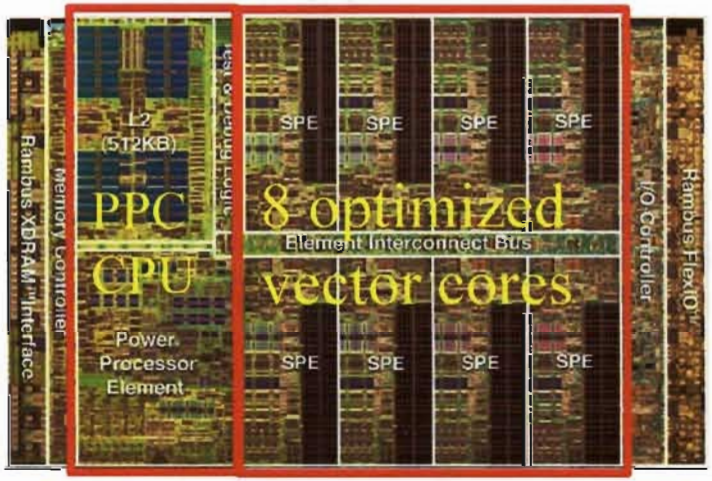
- Running at lower voltages to reduce power consumption increases the probability of errors
- Heterogeneous systems make error detection and recovery even harder, for example, error recovery on GPU system will require managing up to 100 threads
- Increasing system and algorithm complexity makes improper interaction of separate components more likely.
- The rate and effect of undetected (aka silent) errors must be better understood
- In will cost power, performance and $ to add additional HW detection and recovery logic right on the chips to detect silent errors.

New programming model will be required for compute nodes, exascale or not

- Hierarchical approach: intra-node + inter-node
  - Part I: Inter-node model for communicating between nodes
    - MPI scaling to millions of nodes: Importance high; risk low
    - One-sided communication scaling: Importance medium; risk low
  - Part II: Intra-node model for on-chip concurrency
    - Overriding Risk: No single path for node architecture
    - OpenMP, Pthreads: High risk (may not be feasible with node architectures); high payoff (already in some applications)
    - New API, extended PGAS, or CUDA/OpenCL to handle hierarchies of memories and cores: Medium risk (reflects architecture directions); Medium payoff (reprogramming of node code)

Something old: message passing

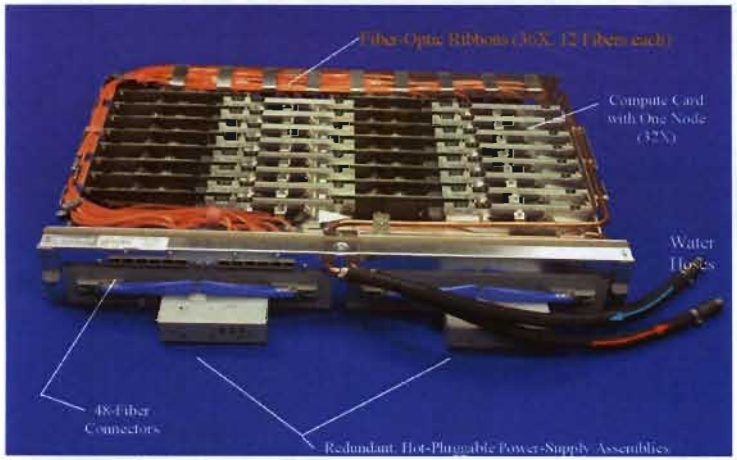




Something new: OpenMP, OpenCL, CAF, UPC, X10, …

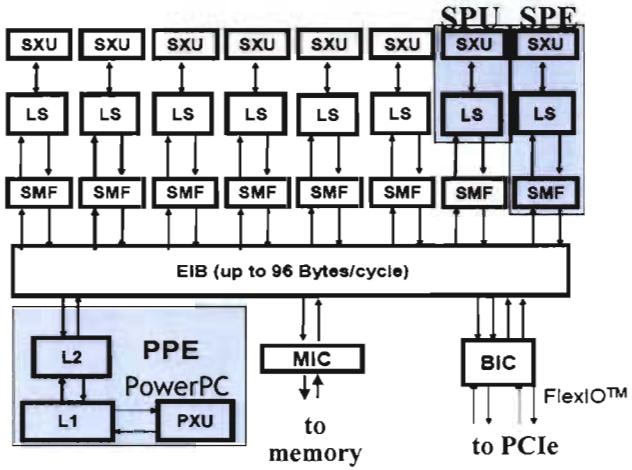
ASC is Pioneering New Architectures Relevant to Exascale Technology Development

BLUE GENE



- Power efficient processor chips allow dense packaging
- High bandwidth / low latency electrical interconnect on-board
- 18+18 (Tx+Rx) 12-channel optical fibers @10Gb/s
  - recombined into 8*48-channel fibers for rack-to-rack (Torus) and 4*12 for Compute-to-IO interconnect
- Compute Node Card assembly is water-cooled (18-25°C – above dew point)
- Redundant power supplies with distributed back-end ~ 2.5 kW

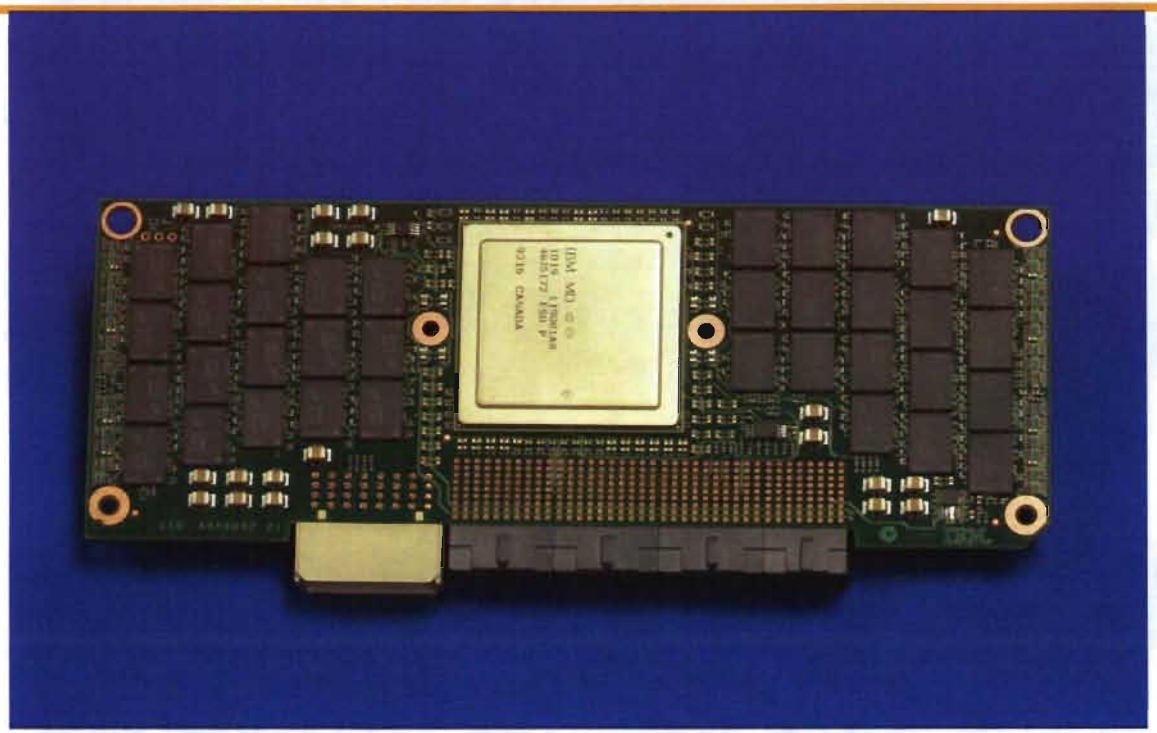
ROADRUNNER



- Cell Broadband Engine (CBE*) developed by Sony-Toshiba-IBM
  - used in Sony PlayStation 3
- **8** Synergistic Processing Elements **(SPEs)**
  - 128-bit **vector cores**
  - 256 kB **local memory** (LS = Local Store)
  - Direct Memory Access **(DMA) engine** (25.6 GB/s each)
  - Chip interconnect **(EIB)**
  - Run SPE-code as POSIX threads (SPMD, MPMD, streaming)
- 1 PowerPC **PPE** runs Linux OS

Design: both technologies focus on delivering optimal flop/s per watt

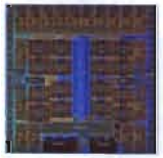
Blue Gene/Q Compute Card



- Basic FRU of a BlueGene/Q system
- Compute Card has 1 BQC chip + 72 SDRAMs (16GB DDR3)
- Two heat sink options: water-cooled → "Compute Node" / air-cooled → "IO Node"
- Connectors carry power supplies, JTAG etc, and 176 HSS signals (4 and 5 Gbps)

Blue Gene/Q packaging hierarchy

1. Chip
16 cores



2. Module
Single Chip

3. Compute Card
One single chip module,
16 GB DDR3 Memory

4. Node Card
32 Compute Cards,
Optical Modules, Link Chips, Torus

5a. Midplane
16 Node Cards

5b. I/O Drawer
8 I/O Cards
8 PCIe Gen2 slots

6. Rack
2 Midplanes
1, 2 or 4 I/O Drawers

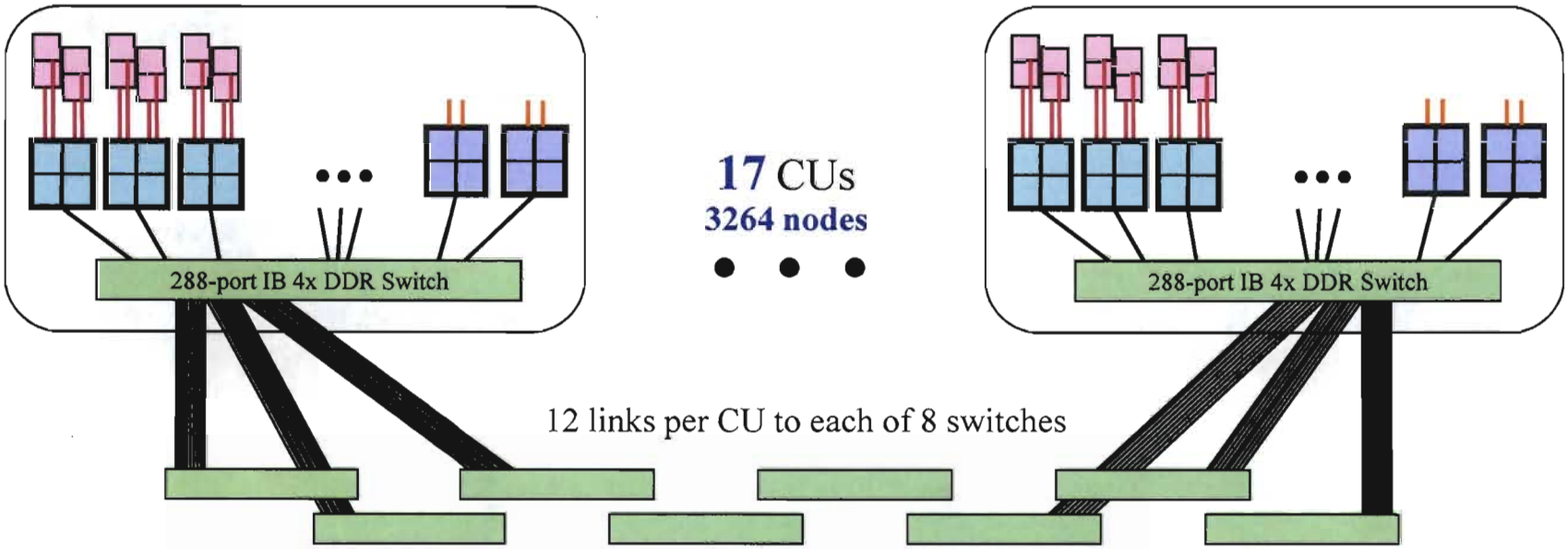
7. System
20PF/s

Roadrunner is a hybrid petascale system of modest node count delivered in 2008

Connected Unit cluster
180 compute nodes w/ Cells
+ 12 I/O nodes

12,240 PowerXCell 8i chips ⇒ 1.33 PF, 49 TB
6,120 dual-core Opterons ⇒ 44 TF, 49 TB

** I/O nodes not counted*



Eight 2nd-stage 288-port IB 4X DDR switches

Future Architectures Pose New Challenges for Computational Algorithms

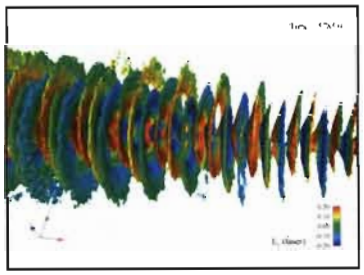
- Research to determine the benefits of various future architectures for particular types of computational algorithms
- Many types of algorithms are of interest to NNSA, e.g.
  - Radiation transport
  - Molecular dynamics
  - Hydrodynamics
  - Plasma physics (e.g. Laser-plasma interactions for ICF)
  - Direct numerical simulation of fluid flow
- The choice of architecture lead to trade-offs in performance, portability, and complexity for each type of algorithm:
  - How much speed increase is possible from the new architecture?
  - What type of changes must be made to the code (or the algorithm!) to enable it to run efficiently on the new architecture? (may favor explicit approaches)
  - Some considerations, such as explicit management of data flow, specialized processors and memory hierarchy, may require extensive alterations to the code
- These trade-offs significantly increase the risk of introducing new architectures to scientific computing
- …but we don't have a choice because the architectures are headed that way

Roadrunner has been a pioneer of new architectures

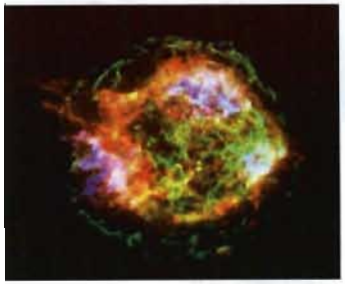
- Roadrunner was a leap into the future
  - *First* computer to reach a petaflop
  - *First* heterogeneous, accelerated supercomputer
  - *First* supercomputer built from non-traditonal commodity processor
- Roadrunner open science provided resources for many important simulations
- Provided important advantages for mission work
  - Increased geometric and physical resolution with reduced turn-around times
  - Implicit Monte Carlo (10X and more speedup) and Deterministic Transport (2-5X speedup) successfully exploited advanced architecture
  - Clarified the requirements for even larger, more highly resolved simulations.
- ORNL Titan machine is next generation
  - Hybrid architecture using GPGPUs
  - 20 PF-class machine in 2012

Shocks in metal

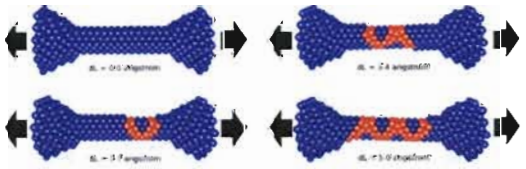
Magnetic reconnection



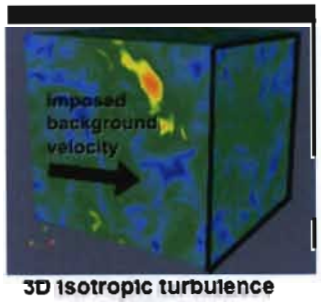
Laser plasma interaction



Core collapse supernova



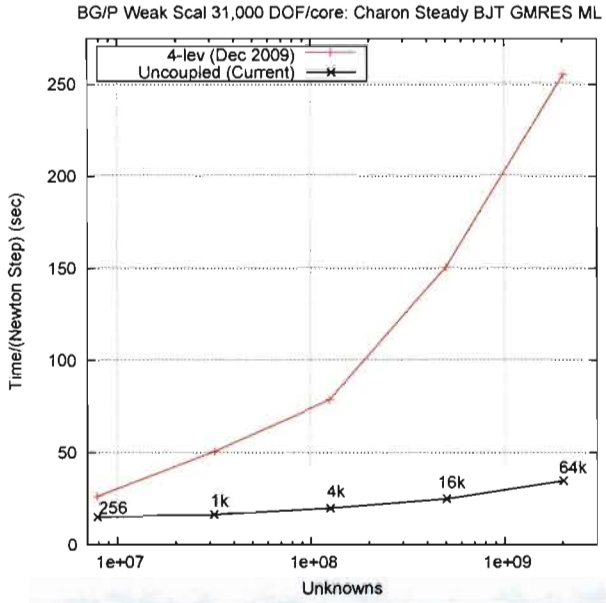
Accelerated MD

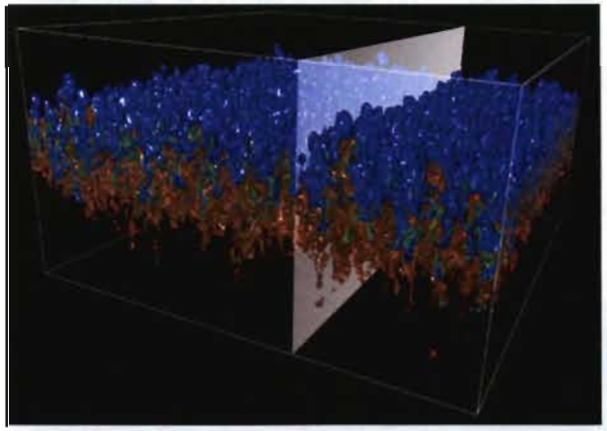


Turbulence with TN burn

Dawn ID system has been used to prepare applications for Sequoia and to conduct weapons science

- Scaling of algorithms (SNL)
  - SNL RAMSES semiconductor physics capability: reduced CPU time by 7× for 2 billion DOF on 64k cores compared with the first scaling study from December 2009
  - Access to very large number of cores critical to improving scaling of some of the key algorithms in Trilinos (e.g., multigrid preconditioner)
  - Many SNL codes depend on Trilinos
- Raleigh Taylor Turbulence (LANL)
  - Generating an extensive database (including the largest fully resolved turbulence simulation ever performed: $4096^2 \times 4608$) to address important open questions related to this instability:
    - New phenomena at large density ratios, effects related to complex acceleration histories and role of initial conditions
    - Data are being used to improve the models under development within ASC





Significant Scientific Advances have Already been Accomplished Using Advanced Architectures

"Three-Dimensional Dynamics of Breakout Afterburner Ion Acceleration Using High-Contrast Short-Pulse Laser and Nanoscale Targets," L. Yin, B. J. Albright, K. J. Bowers, D. Jung, J. C. Fernandez, and B. M. Hegelich, Physical Review Letters, 22 July 2011.

"Role of Electron Physics in the Development of Turbulent Magnetic Reconnection in Collisionless Plasmas," W. Daughton, V. Roytershteyn, H. Karimabadi, L. Yin, B. Bergen, and K. J. Bowers, Nature Physics Letters, April 10, 2011

"Monoenergetic Ion Beam Generation by Driving Ion Solitary Waves with Circularly Polarized Laser Light," D. Jung, L. Yin, B. J. Albright, D. C. Gautier, R. Horlein, D. Kiefer, A. Henig, R. Johnson, S. Letzring, S. Palaniyappan, R. Shah, T. Shimada, X. Q. Yan, K. J. Bowers, T. Tajima, J. C. Fernandez, D. Habs, and B. M. Hegelich, Physical Review Letters, 9 Sept. 2011.

C. H. Still, D. E. Hinkel, A. B. Langdon, J. P. Palastro, and E. A. Williams. "Simulating NIF laser-plasma interaction with multiple SRS frequencies." *J. Physics: Conference Series* **244** (2010)

Livescu, D., M.R. Petersen, and J. Mohd-Yusof. Dilatational effects in turbulent combustion with type Ia supernova microphysics. 2011. In preparation for *Combustion and Flame.*

Mohd-Yusof, J., T.T. Kelley and D. Livescu. A distributed tri-diagonal solver suitable for accelerated architectures. 2011. LA-UR-11-00592 submitted to *Parallel Computing.*

Petersen, M.R., and D. Livescu. Forcing for statistically stationary compressible isotropic turbulence. 2010. *Physics of Fluids* **22**, 116101-1—11.

Petersen, M.R., D. Livescu, J. Mohd-Yusof, and S. Dean. High resolution numerical simulations of compressible isotropic turbulence. 2009. *APS DFD09.* (Minneapolis, MN, 22-24 Nov. 2009). Vol. 54, 19 Edition, p. 161.

Mohd-Yusof, J., D. Livescu, and T. T. Kelley. Adapting the CFDNS compressible Navier-Stokes solver to the Roadrunner architecture.2009. *21st International Conference on Parallel Computational Fluid Dynamics (ParCFD09).* (Moffett Field, CA, 18-22 May).

Broad community input helped formulate the U.S. exascale strategy

- Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009
  - Climate Science (11/08),
  - High Energy Physics (12/08),
  - Nuclear Physics (1/09),
  - Fusion Energy (3/09),
  - Nuclear Energy (5/09),
  - Biology (8/09),
  - Material Science and Chemistry (8/09),
  - National Security (10/09)
- Technology workshops
  - Extreme Architecture and Technology Workshop (12/2009)
  - Cross-cutting technologies (2/10)

The E7 has recently released a Request for Information to the industrial community:

Rick Stevens, ANL, and Andy White, LANL, co-chairs

Argonne National Laboratory

Oak Ridge National Laboratory

Lawrence Berkeley National Laboratory

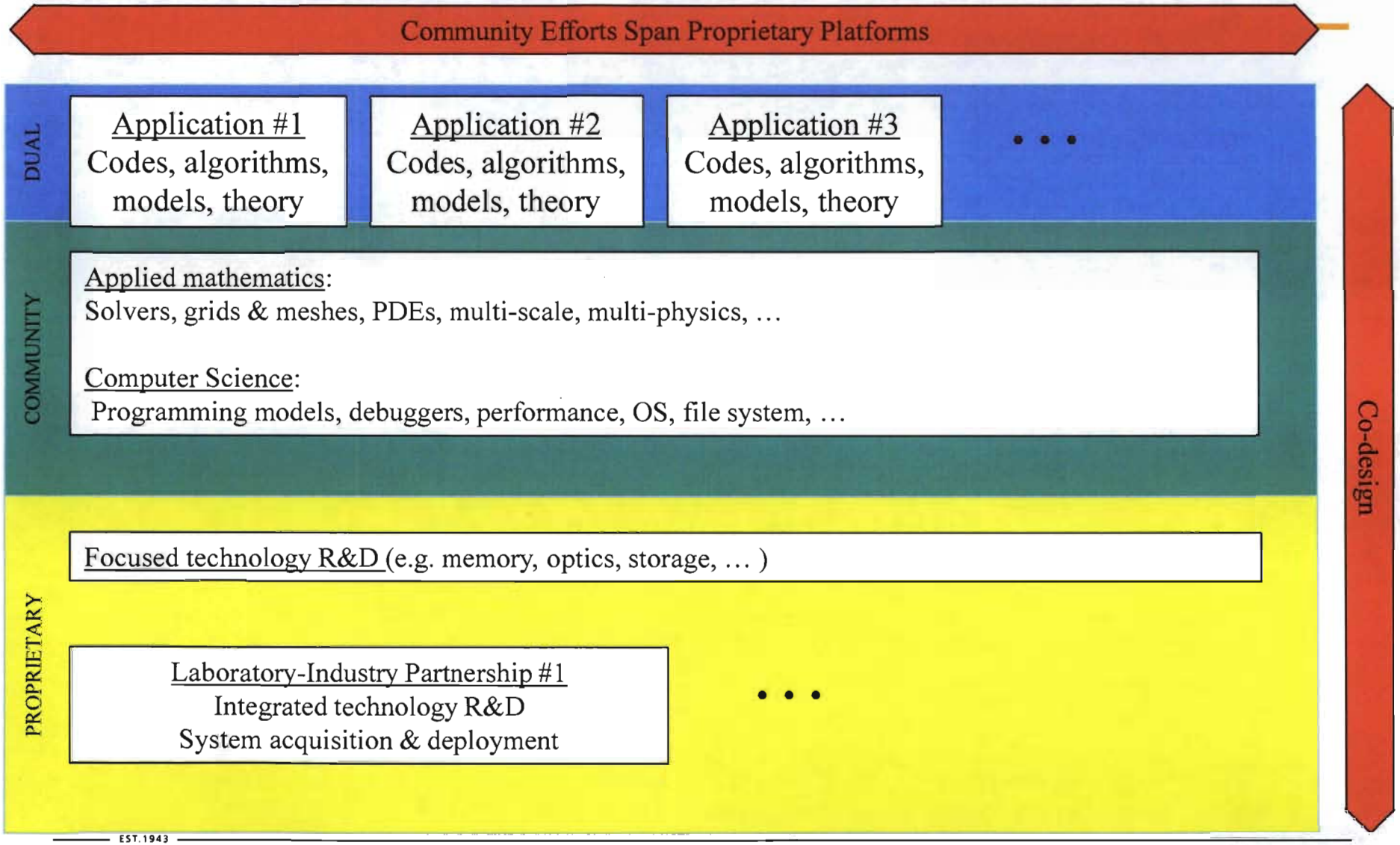
Pacific Northwest National Laboratory

Lawrence Livermore National Laboratory

Los Alamos National Laboratory

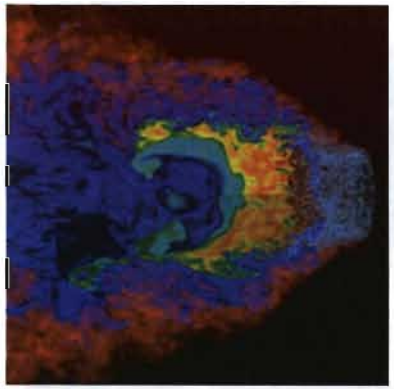
Sandia National Laboratory

Collaboration and co-design are the key ingredients of the exascale approach



Three DOE co-design centers are off and running

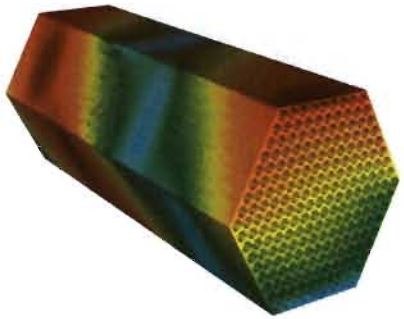
Combustion Exascale Co-Design Center



Director: Jacqueline Chen, SNL
Deputy Director: John Bell, LBNL

With participation from:
SNL, LBNL, ORNL, LANL, University of Texas at Austin, and LLNL

The Center for Exascale Simulation of Advanced Reactors (CESAR)



Principle Investigator: Robert Rosner, ANL/University of Chicago

With participation from:
ANL, LLNL, LANL, ORNL, PNNL, Rice University, Texas A&M University, AREVA, Inc., General Atomics, Inc., IBM, and TerraPower, LLC

Exascale Co-design Center for Materials in Extreme Environments (ExMatEx)

ExMatEx

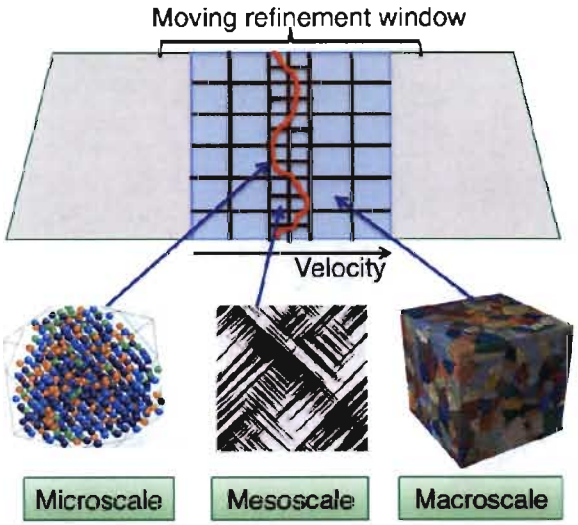
Director: Timothy Germann, LANL
Deputy Director: Jim Belak, LLNL

With participation from:
LANL, LLNL, SNL, ORNL, Stanford University, and California Institute of Technology

Exascale Co-Design Center for Materials in Extreme Environments

ExMatEx

- Establish the interrelationship between algorithms, system software, and hardware required to develop a multiphysics exascale simulation framework for modeling materials subjected to extreme mechanical and radiation environments.
- This effort is focused in four areas:
  - Scale-bridging algorithms
    - UQ-driven adaptive physics refinement
  - Programming models
    - Task-based MPMD approaches to leverage concurrency and heterogeneity at exascale while enabling fault tolerance
  - Proxy applications
    - Communicate the application workload to the hardware architects and system software developers, and used in performance models/simulators/emulators
  - Co-design analysis and optimization
    - Optimization of algorithms and architectures for performance, memory and data movement, power, and resiliency



Exploit hierarchical, heterogeneous architecture to achieve more realistic large-scale simulations with adaptive physics refinement

In Summary

- Exascale computing is needed for:
  - Understanding materials and how they behave under extreme conditions
  - Incorporating that understanding into a broader framework that allows you to reliably predict how a complex, integrated engineered system will behave
  - Both of the above, particularly, in circumstances and under conditions that cannot be directly (for whatever reason) tested through experimentation
- Technology change will effect all scales of computing (and apps)
- Technology change forced on us in order to continue performance improvement gives us the opportunity to rethink the entire solution within a co-design framework

International Snapshot

Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
RIKEN Advanced Institute for Computational Science (AICS) **Japan**	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011 Fujitsu	548352	8162.00	8773.63	9898.56
National Supercomputing Center in Tianjin **China**	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C / 2010 NUDT	186368	2566.00	4701.00	4040.00
DOE/SC/Oak Ridge National Laboratory **United States**	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
National Supercomputing Centre in Shenzhen (NSCS) **China**	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU / 2010 Dawning	120640	1271.00	2984.30	2580.00
GSIC Center, Tokyo Institute of Technology **Japan**	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows / 2010 NEC/HP	73278	1192.00	2287.63	1398.61
DOE/NNSA/LANL/SNL **United States**	Cielo - Cray XE6 8-core 2.4 GHz / 2011 Cray Inc.	142272	1110.00	1365.81	3980.00
NASA/Ames Research Center/NAS **United States**	Pleiades - SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband / 2011 SGI	111104	1088.00	1315.33	4102.00
DOE/SC/LBNL/NERSC **United States**	Hopper - Cray XE6 12-core 2.1 GHz / 2010 Cray Inc.	153408	1054.00	1288.63	2910.00
Commissariat a l'Energie Atomique (CEA) **France**	Tera-100 - Bull bullx super-node S6010/S6030 / 2010 Bull SA	138368	1050.00	1254.55	4590.00
DOE/NNSA/LANL **United States**	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.50
National Institute for Computational Sciences/University of Tennessee **United States**	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz / 2011 Cray Inc.	112800	919.10	1173.00	3090.00
Forschungszentrum Juelich (FZJ) **Germany**	JUGENE - Blue Gene/P Solution / 2009 IBM	294912	825.50	1002.70	2268.00
Moscow State University - Research Computing Center **Russia**	Lomonosov - T-Platforms T-Blade2/1.1, Xeon X5570/X5670 2.93 GHz, Nvidia 2070 GPU, Infiniband QDR / 2011 T-Platforms	33072	674.11	1373.06	