

High Performance Computing Workshop

**Scale-bridging computational materials
science: heterogeneous algorithms for
heterogeneous platforms**

**Timothy C. Germann
Los Alamos National Laboratory**

Greetings from Los Alamos



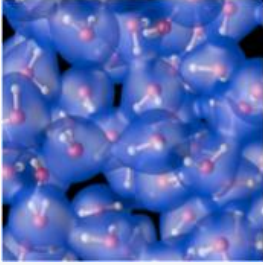
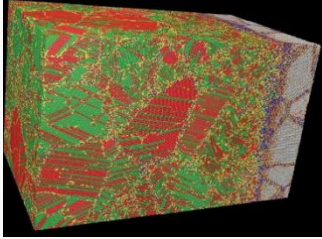
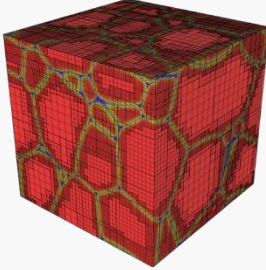
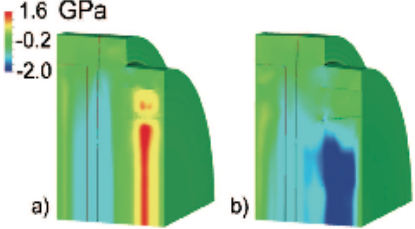
STATE-OF-THE-ART IN COMPUTATIONAL MATERIALS SCIENCE

**Single-scale Gordon Bell-class
applications**



Traditional “sequential” multiscale approach

- Information is passed up a hierarchy of coupled length/time scales via a sequence of subscale models and parameters.

Ab-initio Methods	Molecular Dynamics	Phase-Field Modeling	Continuum Methods
Inter-atomic force model, equation of state	Defect and interface mobility, nucleation	Direct numerical simulation of multi-phase evolution	Multi-phase material response, experimental observables
			
Length/time: nm, ps Code: Qbox/LATTE Motif: Particles and wavefunctions, plane wave DFT with nonlocal norm-conserving, ScaLAPACK, BLACS, and custom parallel 3D FFTs Prog. Model: MPI	Length/time: μm , ns Code: SPaSM/ddcMD Motif: Particles, domain decomposition, explicit time integration, neighbor and linked lists, dynamic load balancing, parity error recovery, and <i>in situ</i> visualization Prog. Model: MPI + Threads	Length/time: $100 \mu\text{m}$, μs Code: AMPE/GL Motif: Regular and adaptive grids, implicit time integration, real-space and spectral methods, complex order parameter (phase, crystal, species) Prog. Model: MPI	Length/time: cm, ms Code: VP-FFT/ALE3d Motif: Regular and irregular grids, implicit time integration, 3D FFTs, polycrystal and single crystal plasticity, Prog. Model: MPI

For a recent example, see: N. Barton *et al*, “A multiscale strength model for extreme loading conditions,” *J. Appl. Phys.* **109**, 073501 (2011)

State-of-the-art electronic structure: Qbox

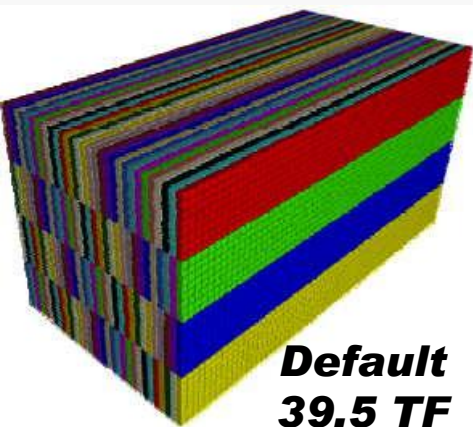
- First-principles molecular dynamics (quantum electrons, classical nuclei) based on the plane-wave, pseudopotential method for electronic structure calculations

- Kinetic and potential terms are sparse in Fourier and real space, respectively, which requires frequent 3D FFTs

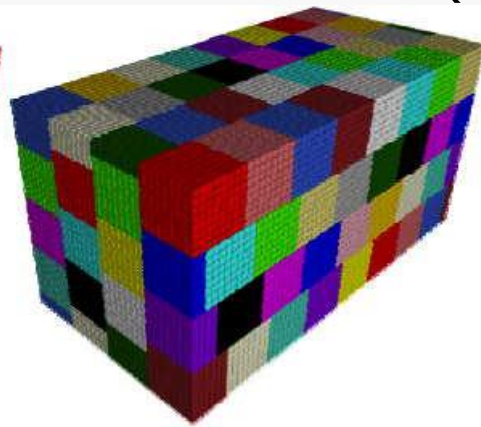
$$\left(-\frac{1}{2}\nabla^2 + V_{eff}(\mathbf{r})\right)\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r})$$

- Maintaining orthogonality constraint requires frequent dense linear algebra
- Optimal node mapping is non-obvious (see below)
- 2006 Gordon Bell Prize winner (207 Tflop/s)

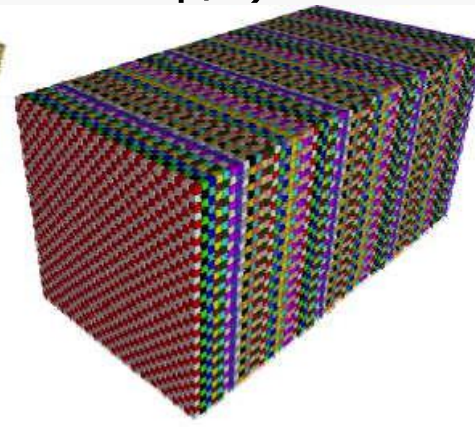
$$\langle\psi_i|\psi_j\rangle = \delta_{ij}$$



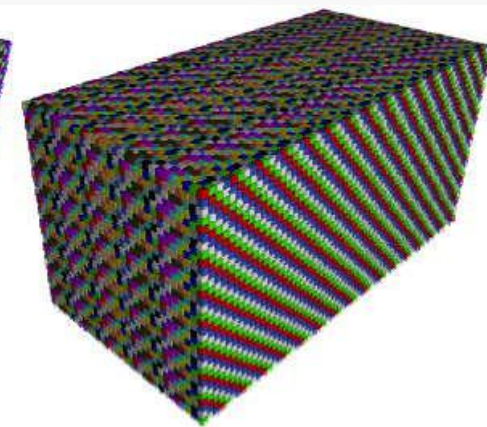
Default
39.5 TF



Compact
38.2 TF



Bipartite
64.0 TF



Quadpartite
64.7 TF

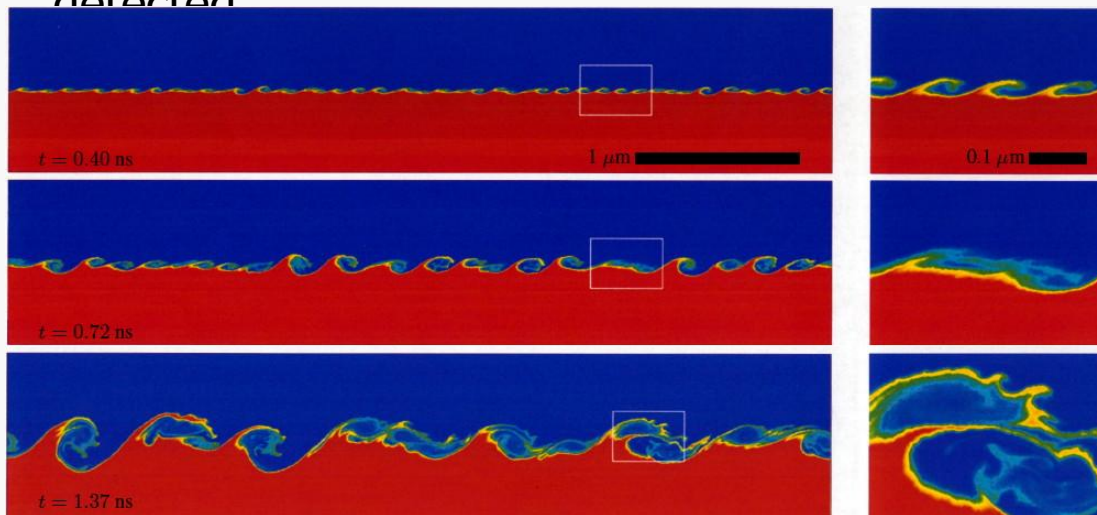


State-of-the-art molecular dynamics: ddcMD

- domain decomposition Molecular Dynamics
- Particle-based domain decomposition with dynamic load balancing
- MGPT potential: computationally expensive, avoid redundant communication and computation

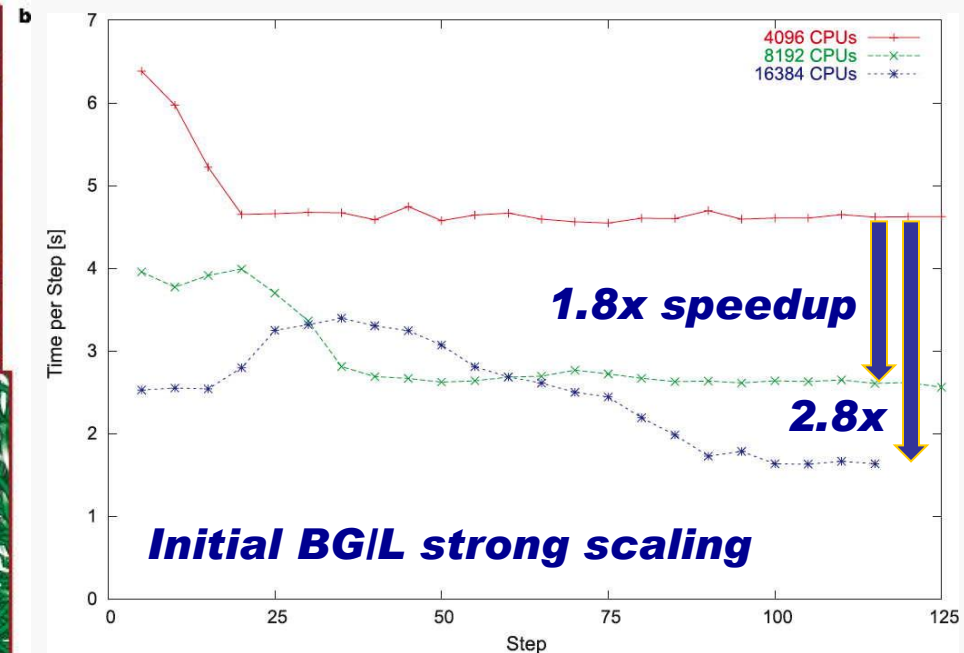
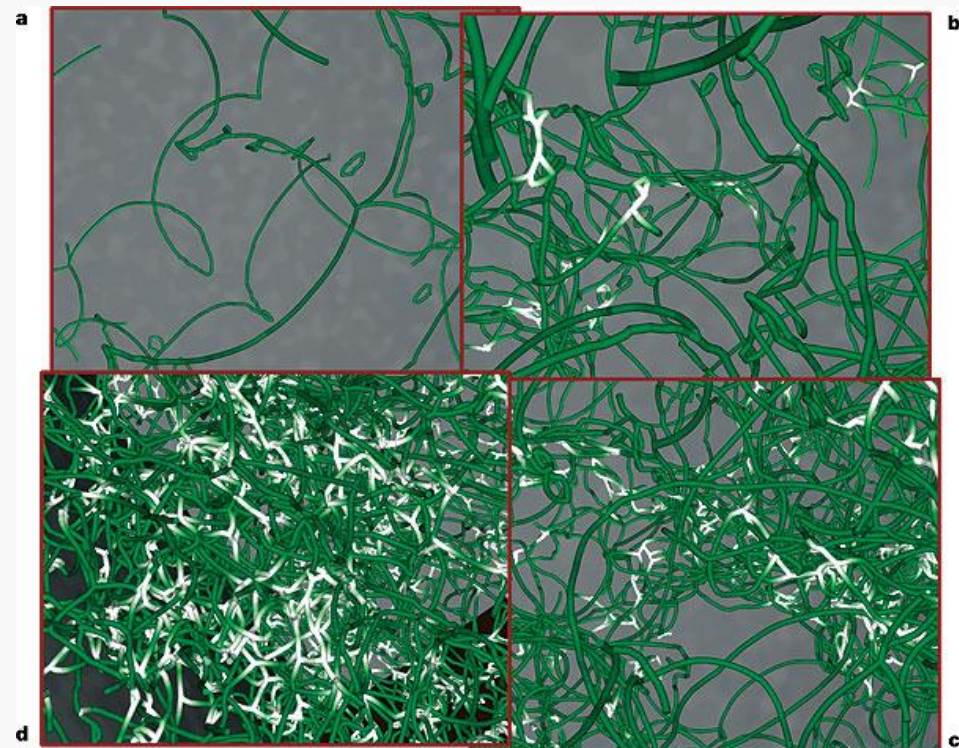
$$E = E_0(\rho) + \frac{1}{2} \sum_{ij} \phi_{ij}(\rho) + \frac{1}{6} \sum_{ijk} \phi_{ijk}(\rho) + \frac{1}{24} \sum_{ijkl} \phi_{ijkl}(\rho)$$

- EAM potential: cheap potential, cost to identify and avoid redundant computation is not worthwhile
- Parity error recovery: periodically store system state (atom positions and velocities) in memory, and restore if an unrecoverable parity error is detected



State-of-the-art dislocation dynamics: ParaDiS

- Parallel Dislocation Simulator
- Discretize dislocations (line defects) into line segments
- Long-range interaction → fast multipole method
- >80% of time force evaluation
- Highly heterogeneous distribution of segments → novel dynamic load balancing algorithm



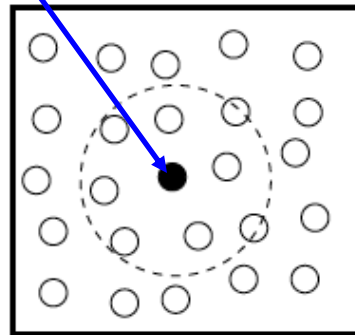
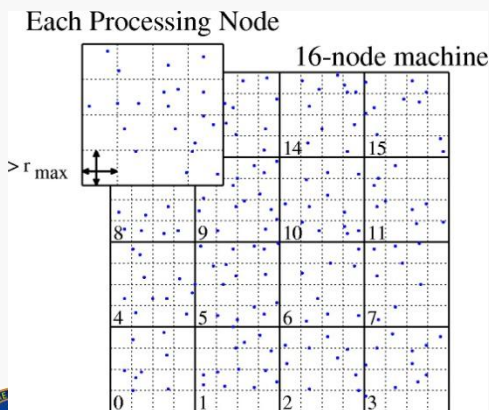
State-of-the-art molecular dynamics: SPaSM

Compute force

Molecular
Dynamics
Timestep

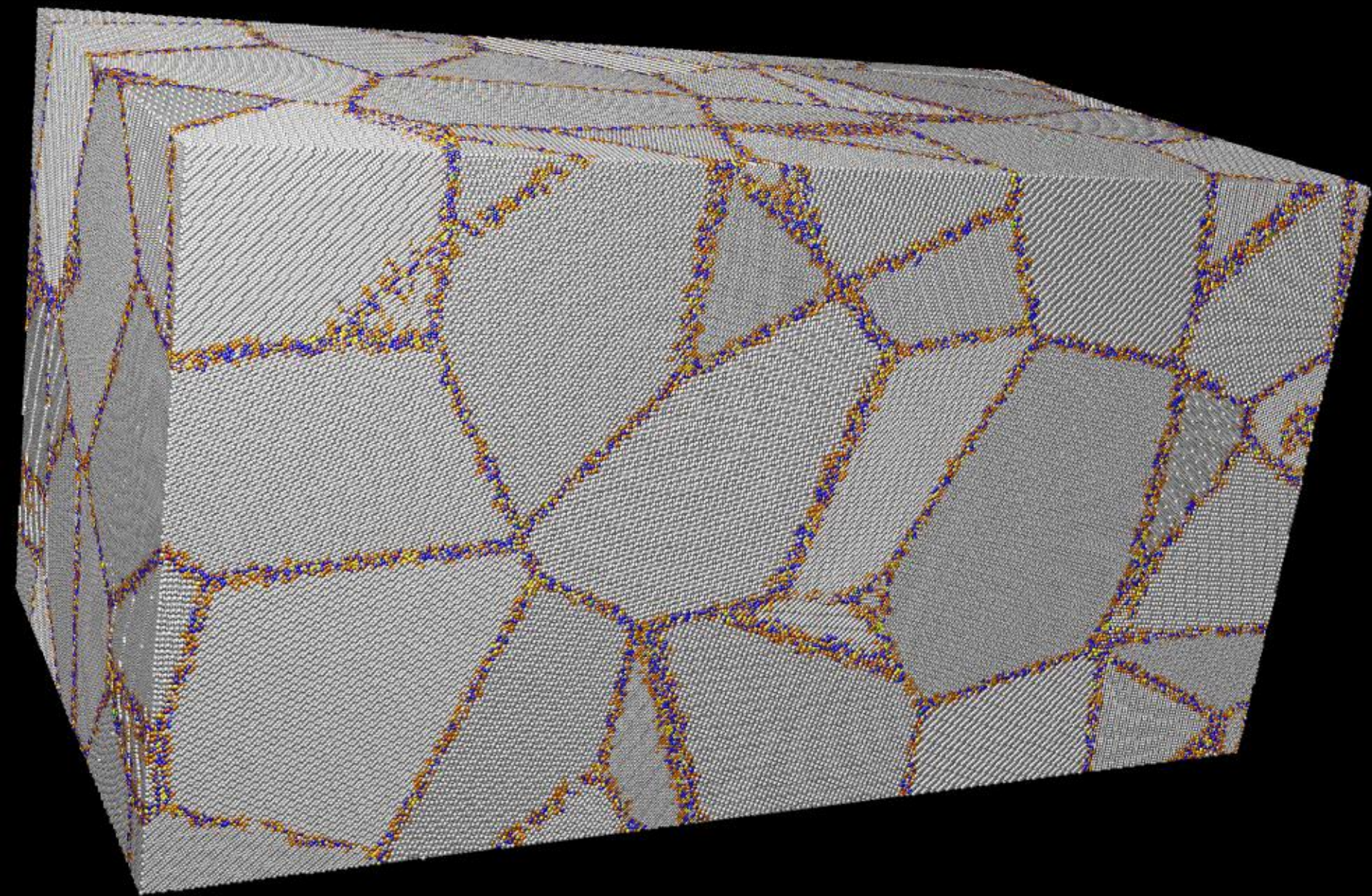
Advance
particles

$$\mathbf{F}_i(t) = m_i \ddot{\mathbf{r}}_i(t) = -\frac{\partial \Phi}{\partial \mathbf{r}_i}$$



- **Scalable Parallel Short-range Molecular dynamics**
- Finite-range (r_{max}) interactions $\Rightarrow O(N)$ computational scaling
- Spatial decomposition on shared and distributed memory architectures
- 1993 Gordon Bell Prize (CM-5)*
- 1998 Gordon Bell Prize (Avalon)
- 2005, 2008 GB Finalist (BG/L, RR)
- First (only) trillion-atom simulation
- Object-oriented scripting language with parallel in situ visualization and analysis libraries (runtime "steering")

*David Beazley, Peter Lomdahl,
Niels Grønbech-Jensen, Pablo Tamayo

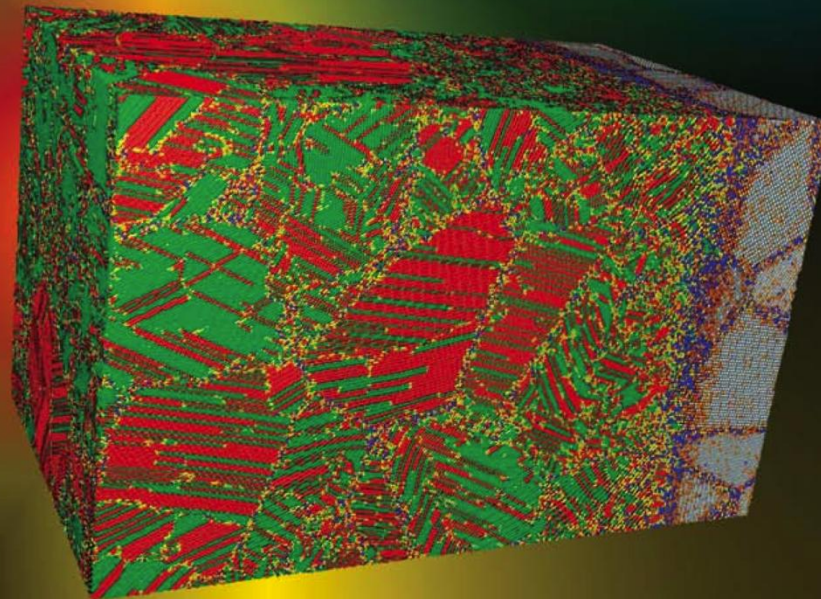


MRS Bulletin

December 2010 Vol. 35 No. 12
www.mrs.org/bulletin

MRS MATERIALS RESEARCH SOCIETY
Advancing materials. Improving the quality of life.

Structural metals at extremes



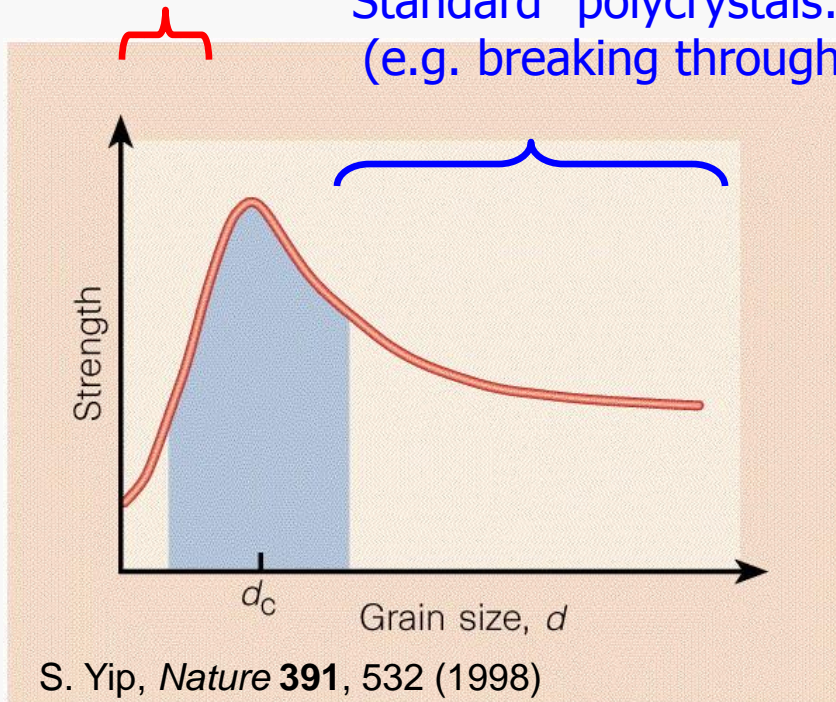
ALSO IN THIS ISSUE

**Materials for organic
and hybrid inorganic/
organic electronics**

Beyond the "Hall-Petch Peak": Polycrystal Plasticity and Mechanical Response

Nanocrystalline metals: grain boundary (GB) sliding dominates

"Standard" polycrystals: dislocation pileups within grains
(e.g. breaking through GBs or other obstacles) dominates



The predicted d_c ranges from 2-3 nm (Ni, Fe) to 20 nm (Cu) for most metals.

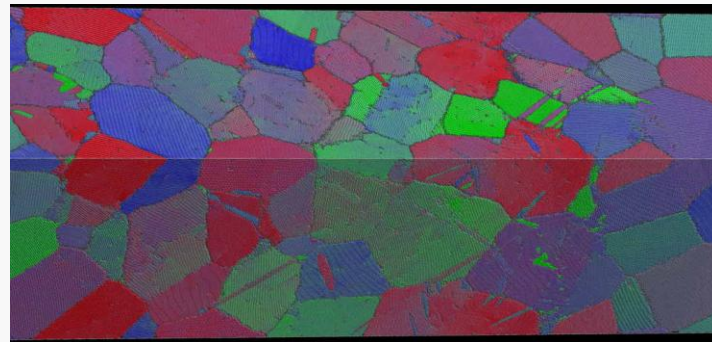
To model a polycrystal exhibiting a "standard" response requires ~ 100 grains with diameter ~ 50 nm

$\Rightarrow 10^8 - 10^9$ atoms

for timescales of ns - μ s

$\Rightarrow 10^6 - 10^9$ timesteps

*Shocked nc-Ta
~100 M atoms
(Prof. Ramon Ravelo)*



Multicore Chips, Co-processors, and Accelerators are Outracing Traditional CPUs

Wired News: <http://wired.com/>

WIRED NEWS

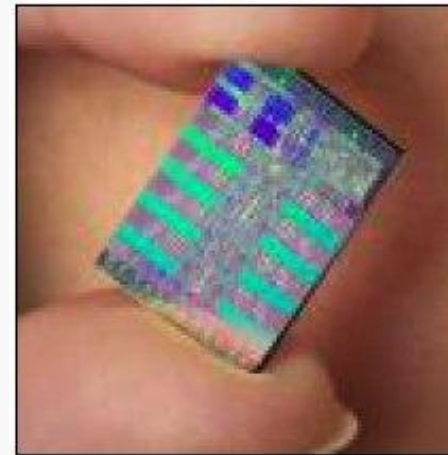
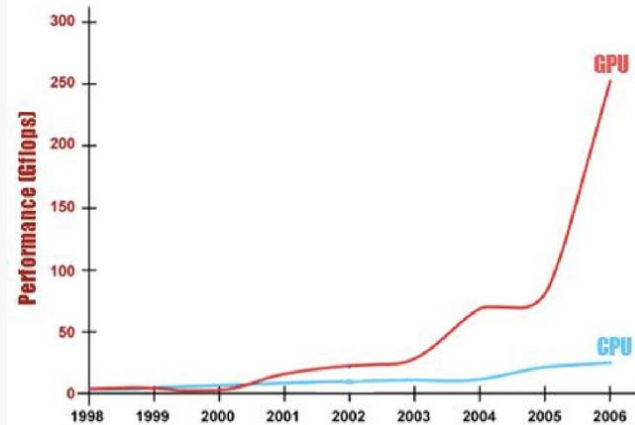
Top Technology Culture Politics Columns Blogs Wired Mag

November 9, 2006 | RSS • PDA • News Archive • Corrections

What's New, GPU?

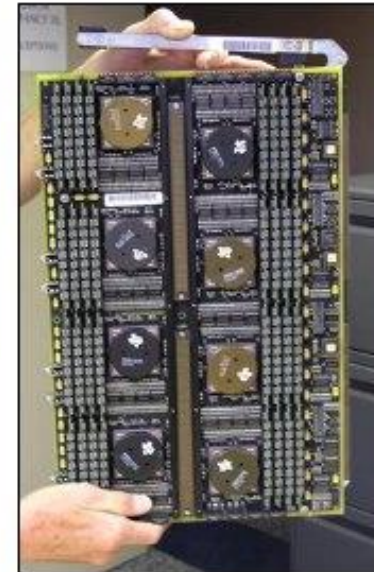
Supercomputing's Next Revolution

The future of high-performance computing lies in graphics chips developed for the consumer video game market. By Paul Tulloch. Nov 9, 2006 | 2:00 AM



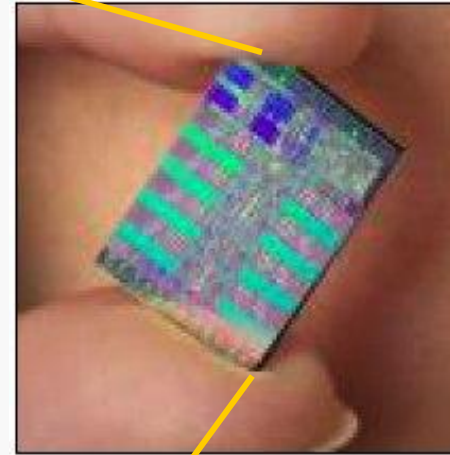
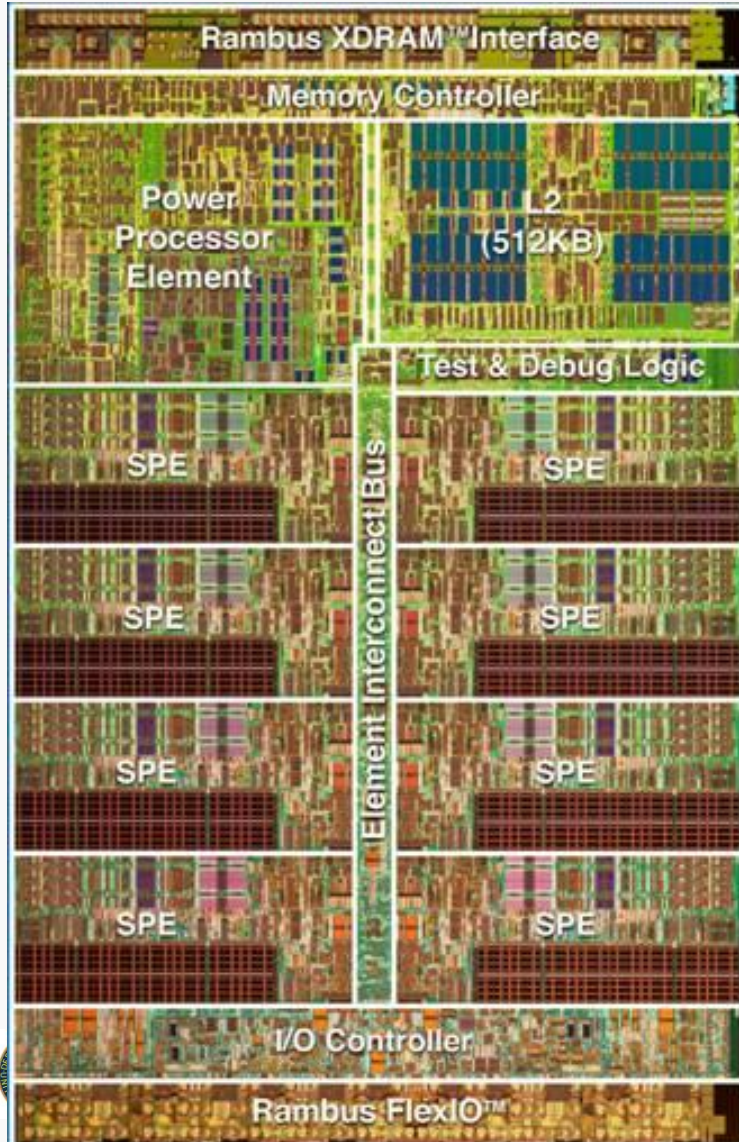
Cell processor (2007, 100 GF)

100x in
14 yrs
8 vector units each



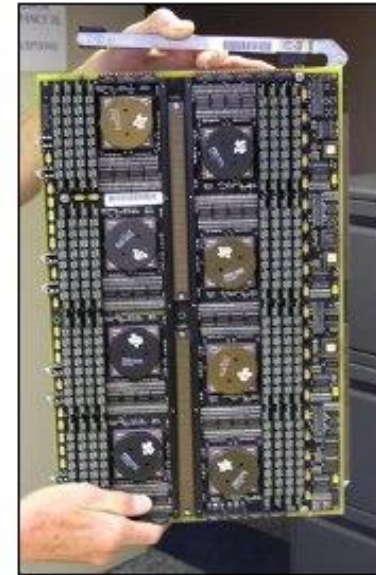
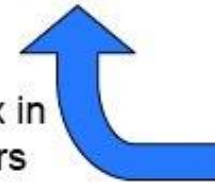
CM-5 board (1994, 1 GF)

Multicore Chips, Co-processors, and Accelerators are Outracing Traditional CPUs



Cell processor (2007, 100 GF)

100x in
14 yrs
8 vector units each



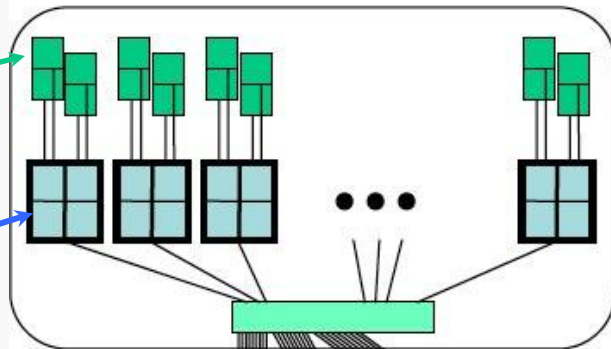
CM-5 board (1994, 1 GF)

Roadrunner, a hybrid Cell-accelerated Opteron cluster, was the 1st PF supercomputer

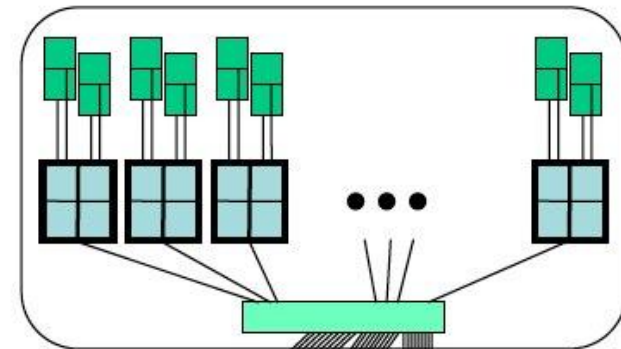
17 Connected Units (CUs)
6,120 dual-core Opteron (+408 I/O)
• 44 TF/s peak
12,240 eDP Cell chips
• 1.26 PF/s peak

“Connected Unit” cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades
connected w/ 4 PCIe x8 links)

Dual-Cell
blade



CU clusters



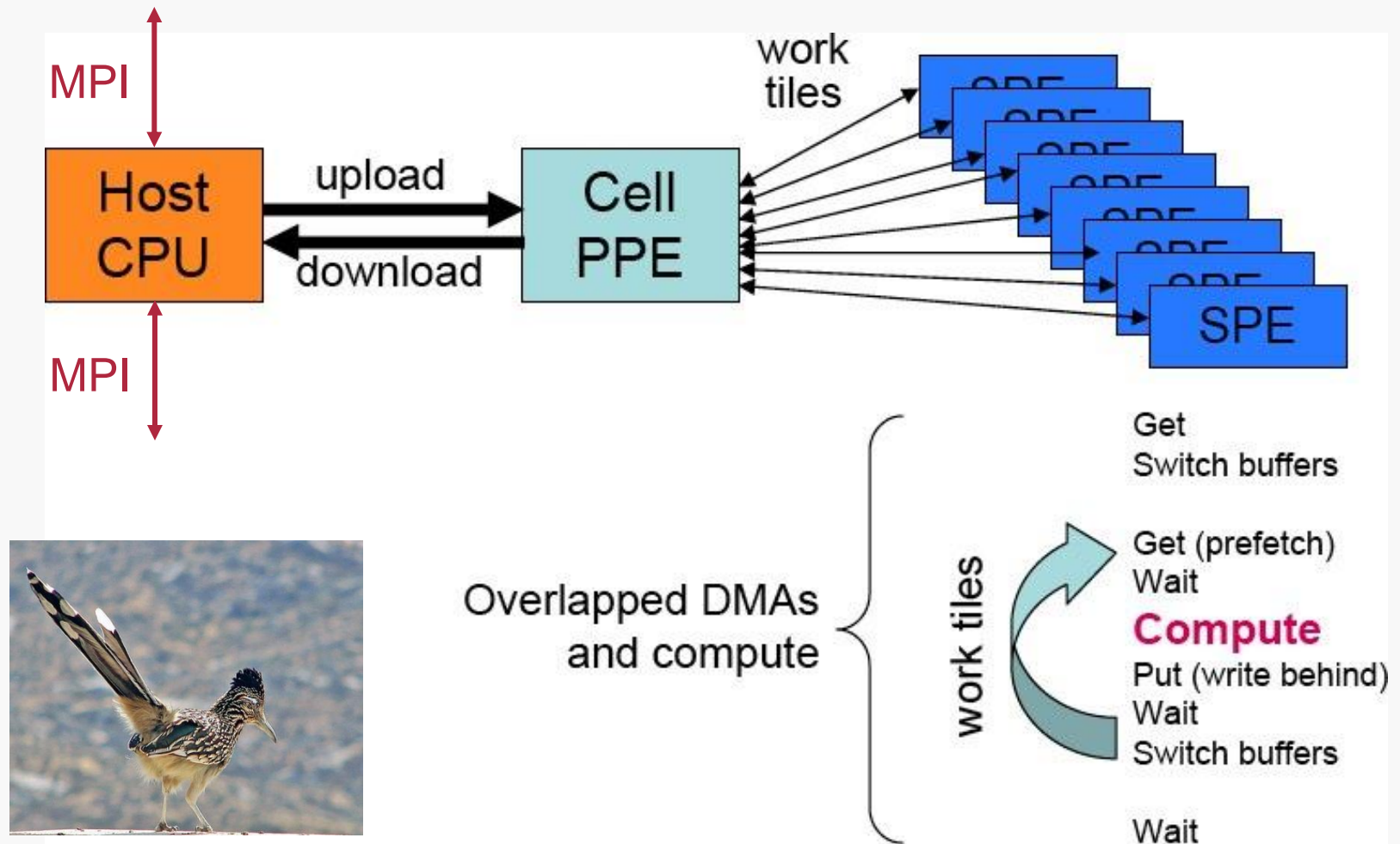
2nd stage InfiniBand 4x DDR interconnect
(18 sets of 12 links to 8 switches)

2nd Gen IB 4X DDR

Opteron
(2x dual-core)

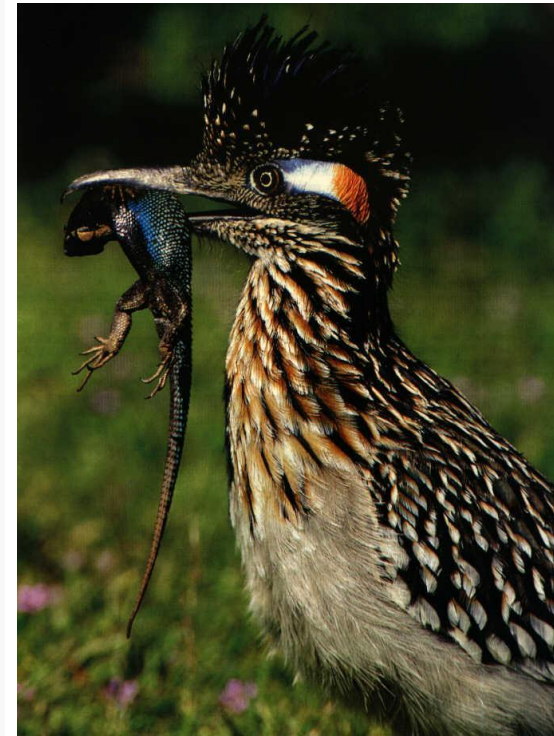


Idealized Parallel Hybrid Programming Model



Real-World Parallel Hybrid Programming

- Three different compilers for 3 CPU types (Opteron, PPC, SPE)
- At least two communication interfaces (some newly invented for Roadrunner....)
- Opteron and Cell (shared PPE/SPE) memory, plus local store/cache(s) on each processor
- PS: the 3 CPU types use different byte orderings (both Big and Little Endian)

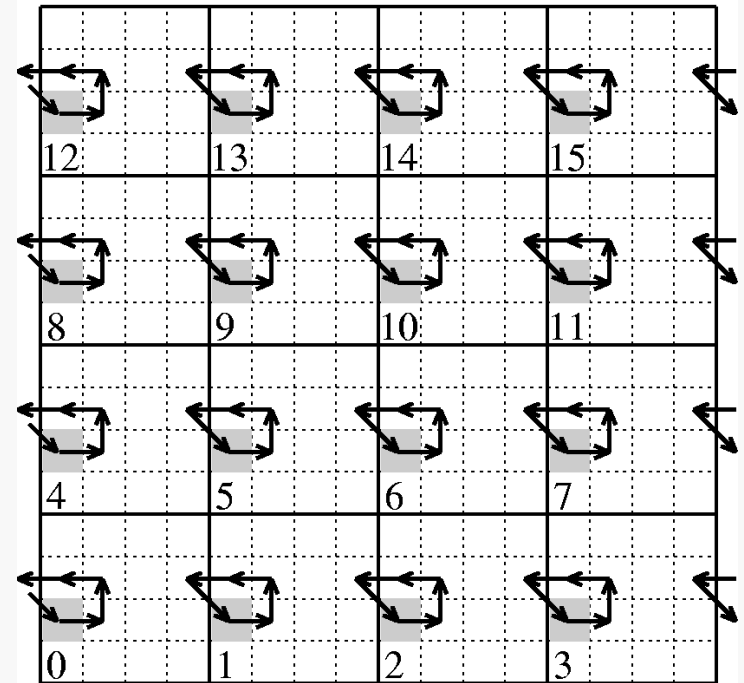


Original SPaSM design considerations

Thinking Machines CM-5 target

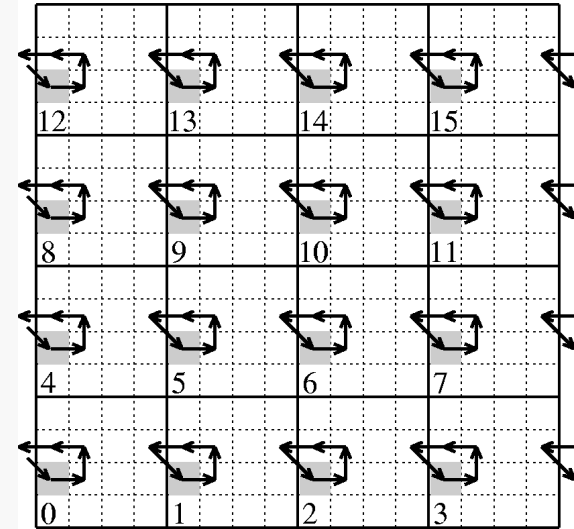
- Memory, computation expensive (32 MB / SPARC-based node)
- Communication fairly cheap
 - CM-5 fat tree network
 - Cray T3D 3D torus
(also for IBM BlueGene/L)

∴ Fine-grain parallelism with each MPI process advancing through subdomains in lockstep, buffering only one off-CPU cell using `MPI_send_and_receive()`

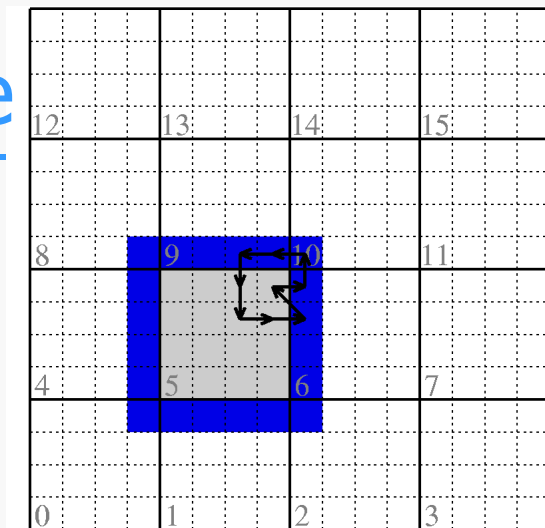


Force calculation pseudocode

```
for each subdomain  $i$ :
    compute self-interactions ( $i, i$ )
    for each neighboring subdomain  $j$  in half-path:
        if half-path crosses processor boundary:
            MPI_send_and_receive()
        compute interactions ( $i, j$ ) = ( $j, i$ )
    end for
end for
```



Full-ghost cell buffering* scheme



***aka “halo exchange”**

```
get ghost cells from neighboring processors:
```

```
    MPI_send_and_receive()
```

```
for each subdomain i:
```

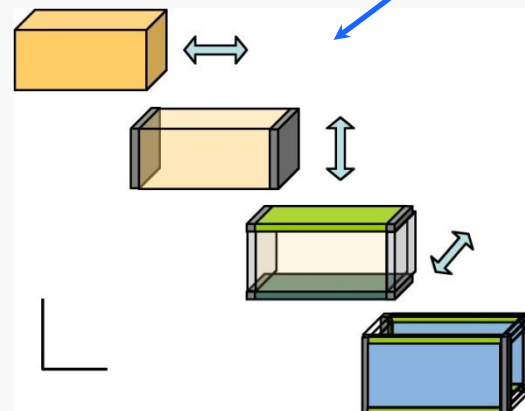
```
    compute self-interactions (i, i)
```

```
    for each neighbor j in full path:
```

```
        compute interactions (i, j)
```

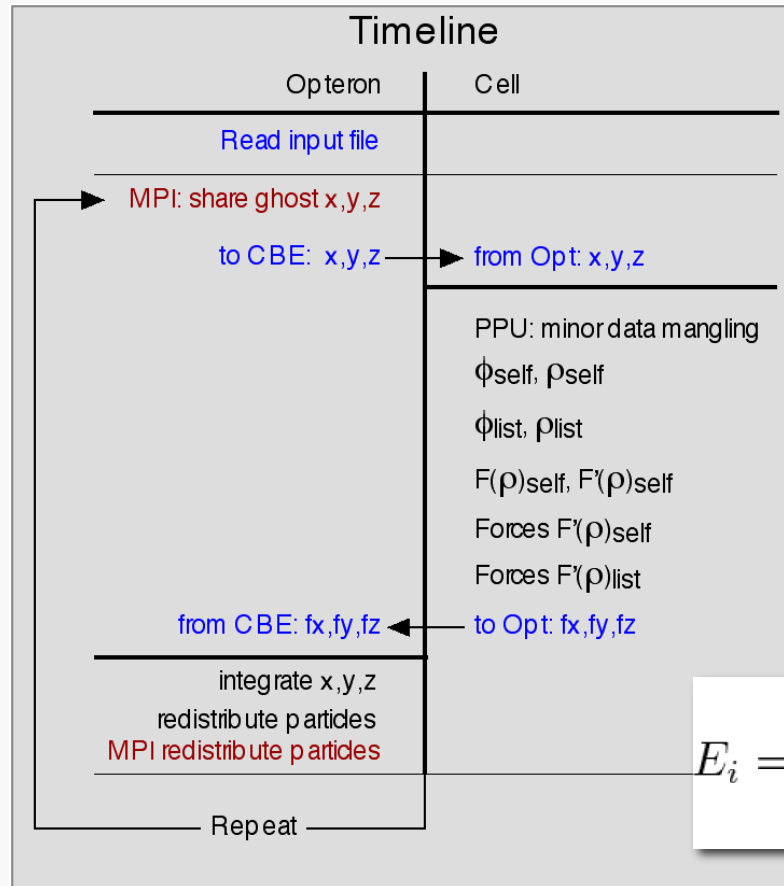
```
    end for
```

```
end for
```

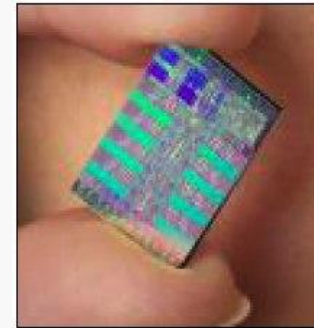


The conservative ("evolutionary") port of SPaSM achieved a speedup of only ~2.5x

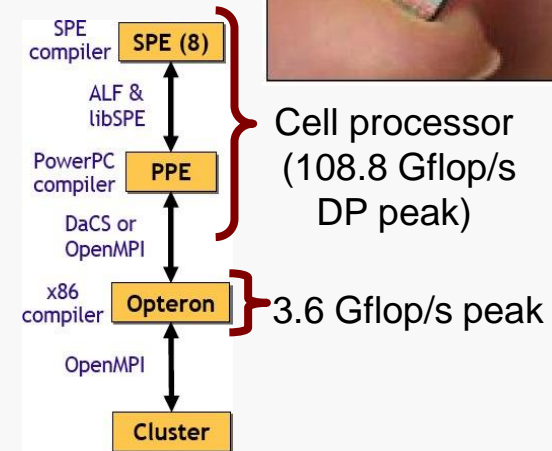
SPUs compute forces, then sit idle as Opterons update positions/velocities (and vice-versa)



~2.5x faster than original code on base Opterons



Cell processor (108.8 Gflop/s DP peak)



$$E_i = \frac{1}{2} \sum_j \phi(r_{ij}) + F \left[\sum_j \rho_j(r_{ij}) \right]$$



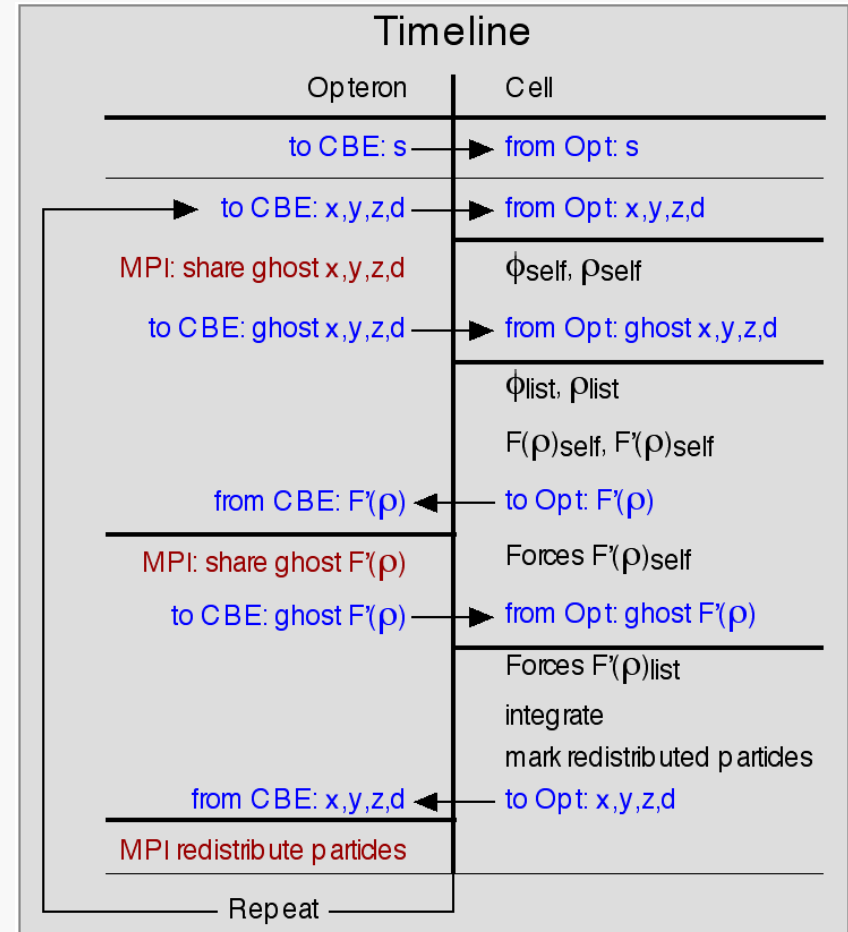
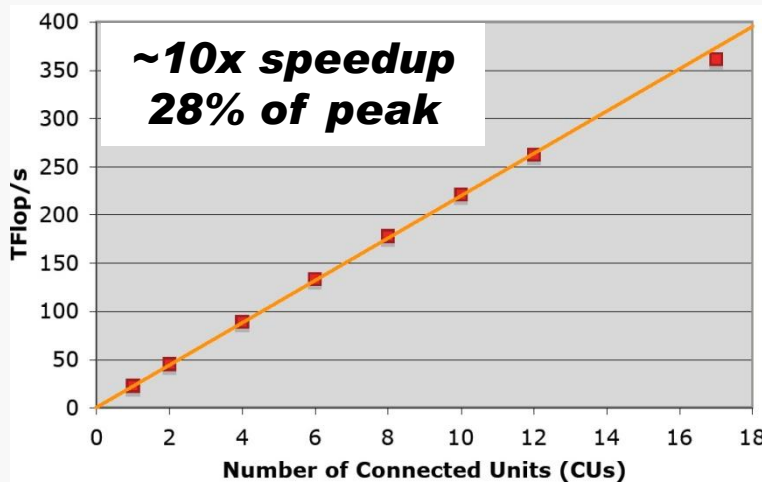
The Cell-centric (“revolutionary”) redesign achieved 369 TFlop/s in June `08

Cell SPUs:

- Own data and all compute-intensive parts
- Have minimal idle time
- Kernel runs at ~50% of theoretical peak

Opteron:

- Own all off-node communication
- Lots of dead time – can be used for analysis, viz, and I/O (checkpointing)



T. C. Germann, K. Kadau, and S. Swaminarayan, “369 Tflop/s Molecular Dynamics Simulations on the Petaflop Hybrid Supercomputer ‘Roadrunner’,” *Concurrency and Computation: Practice and Experience* 21, 2143-2159 (2009).



WHY MULTISCALE (FROM A MATERIALS SCIENCE PERSPECTIVE)

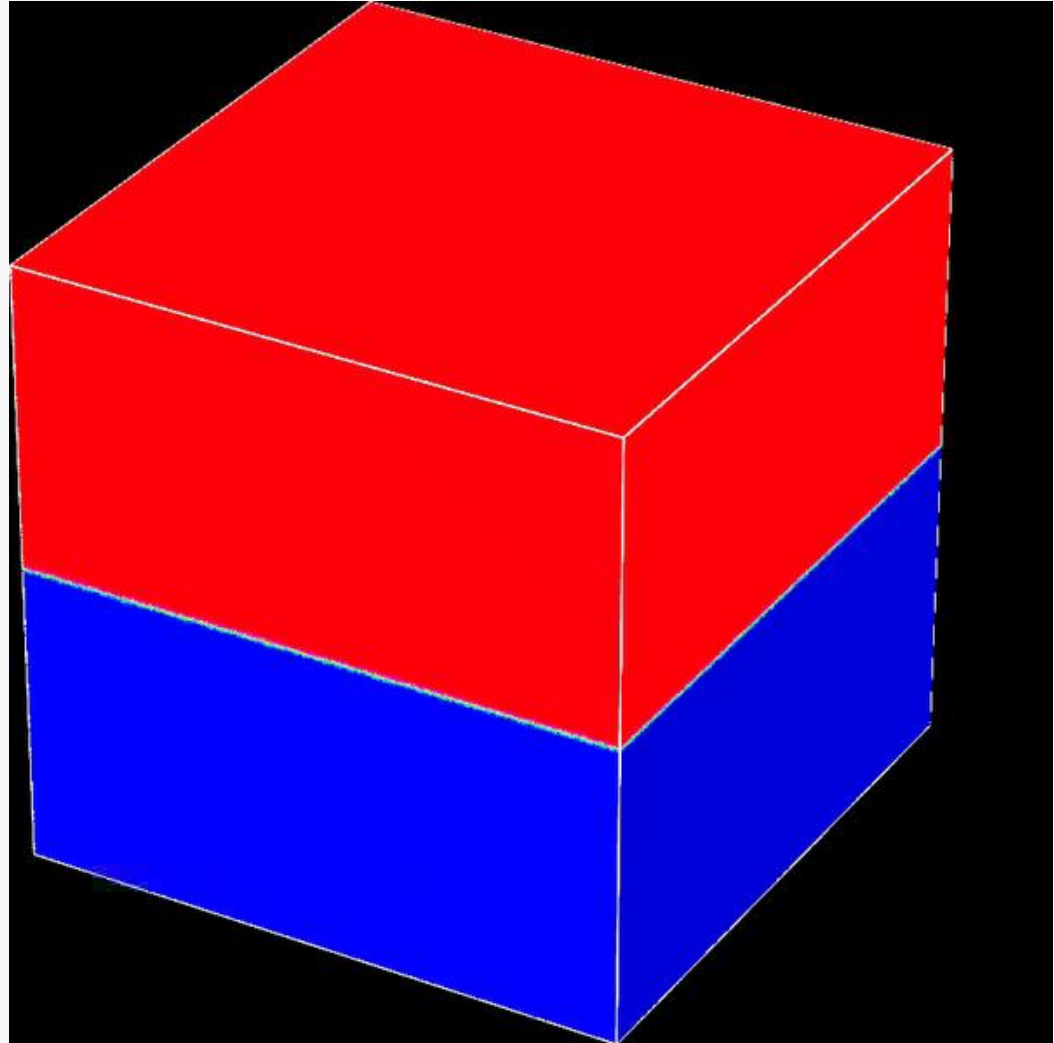


As system sizes increase, atomistic resolution is necessary in a diminishing volume fraction

of atoms (computational effort) scales with volume, while features of interest often scale with area

Examples include:

- Interface instabilities
 - Example: 7.4 billion atom Rayleigh-Taylor (Jan `06 BlueGene/L)
- Shock fronts
- Phase boundaries
 - Product phase nucleation and growth within a parent phase



PHILOSOPHICAL TRANSACTIONS

— OF —
THE ROYAL
SOCIETY

A

MATHEMATICAL, PHYSICAL
& ENGINEERING SCIENCES

ISSN 1364-503X

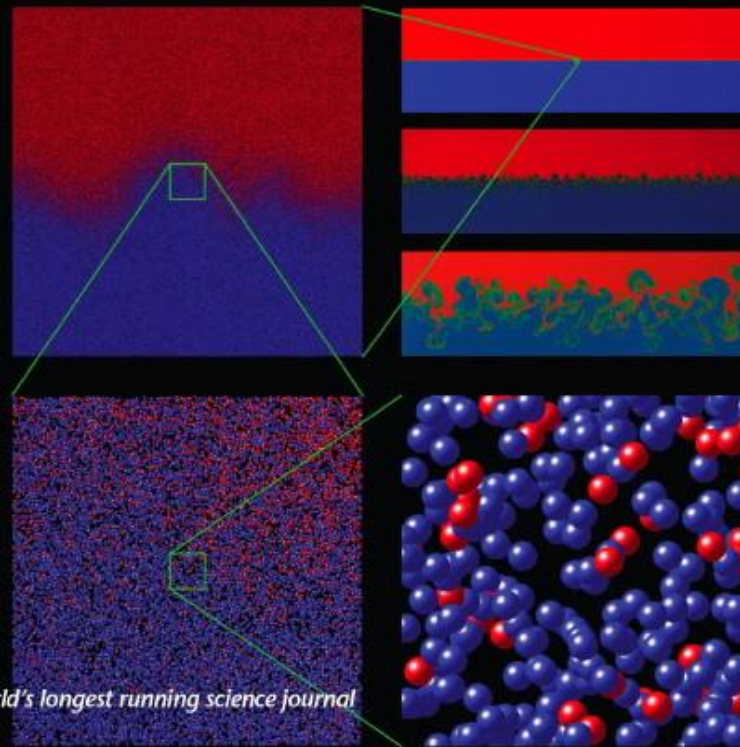
volume 368

number 1916

pages 1537–1828

Turbulent mixing and beyond

*Papers of a Theme Issue compiled and edited by Snezhana I. Abarzhi and
Katepalli R. Sreenivasan*



The world's longest running science journal

CELEBRATE
350 YEARS
Se



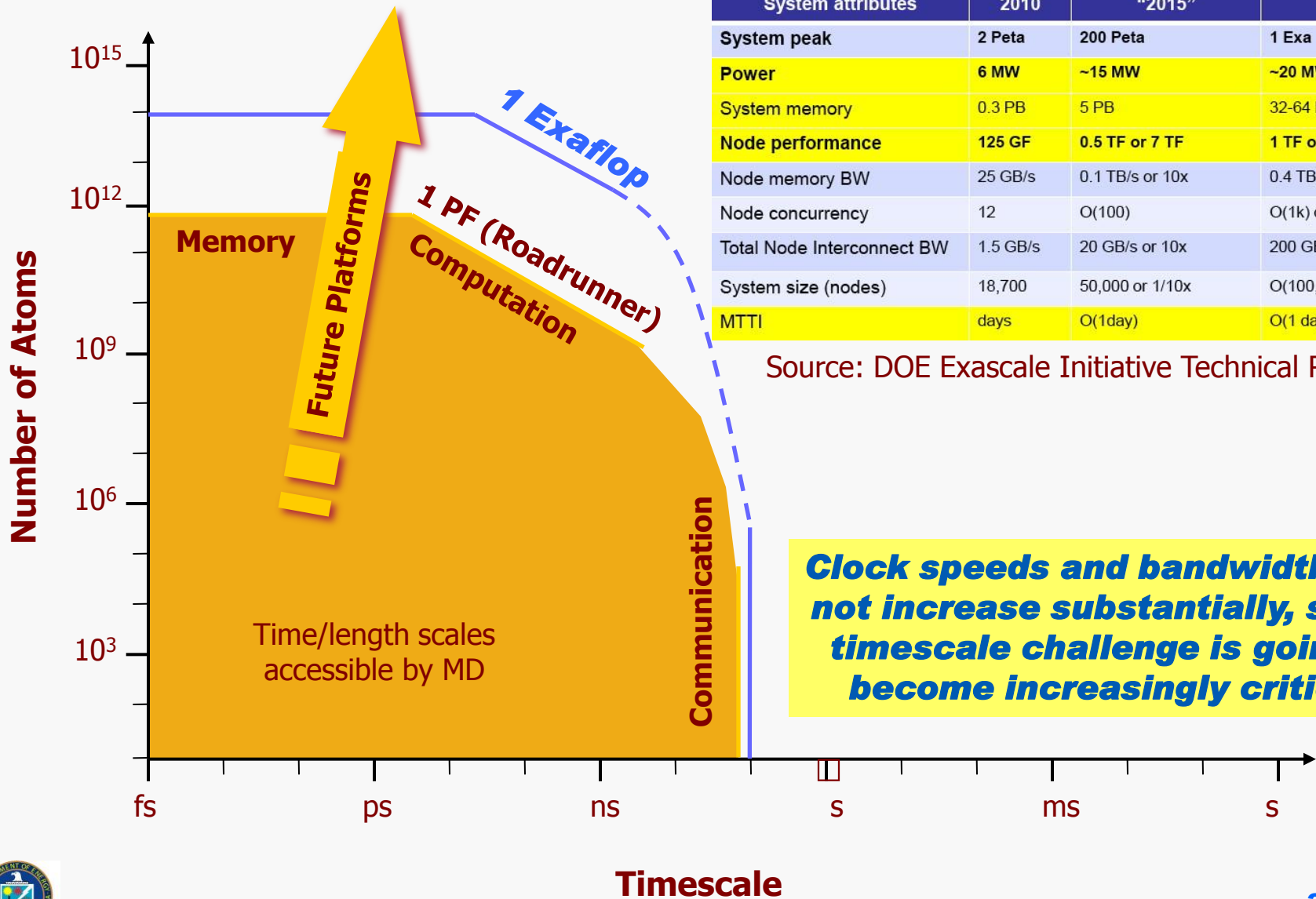
Royal Society **Publishing**

*Informing the science
of the future*

See further with the Royal Society in 2010 – celebrate 350 years

13 April 2010

This decade will see an evolution from petascale (10^{15} Flop/s) to exascale (10^{18} Flop/s) supercomputers



System attributes	2010	"2015"	"2018"
System peak	2 Peta	200 Peta	1 Exa
Power	6 MW	~15 MW	~20 MW
System memory	0.3 PB	5 PB	32-64 PB
Node performance	125 GF	0.5 TF or 7 TF	1 TF or 10x
Node memory BW	25 GB/s	0.1 TB/s or 10x	0.4 TB/s or 10x
Node concurrency	12	O(100)	O(1k) or 10x
Total Node Interconnect BW	1.5 GB/s	20 GB/s or 10x	200 GB/s or 10x
System size (nodes)	18,700	50,000 or 1/10x	O(100,000) or 1/10 x
MTTI	days	O(1day)	O(1 day)

Source: DOE Exascale Initiative Technical Roadmap

Clock speeds and bandwidths will not increase substantially, so the timescale challenge is going to become increasingly critical.



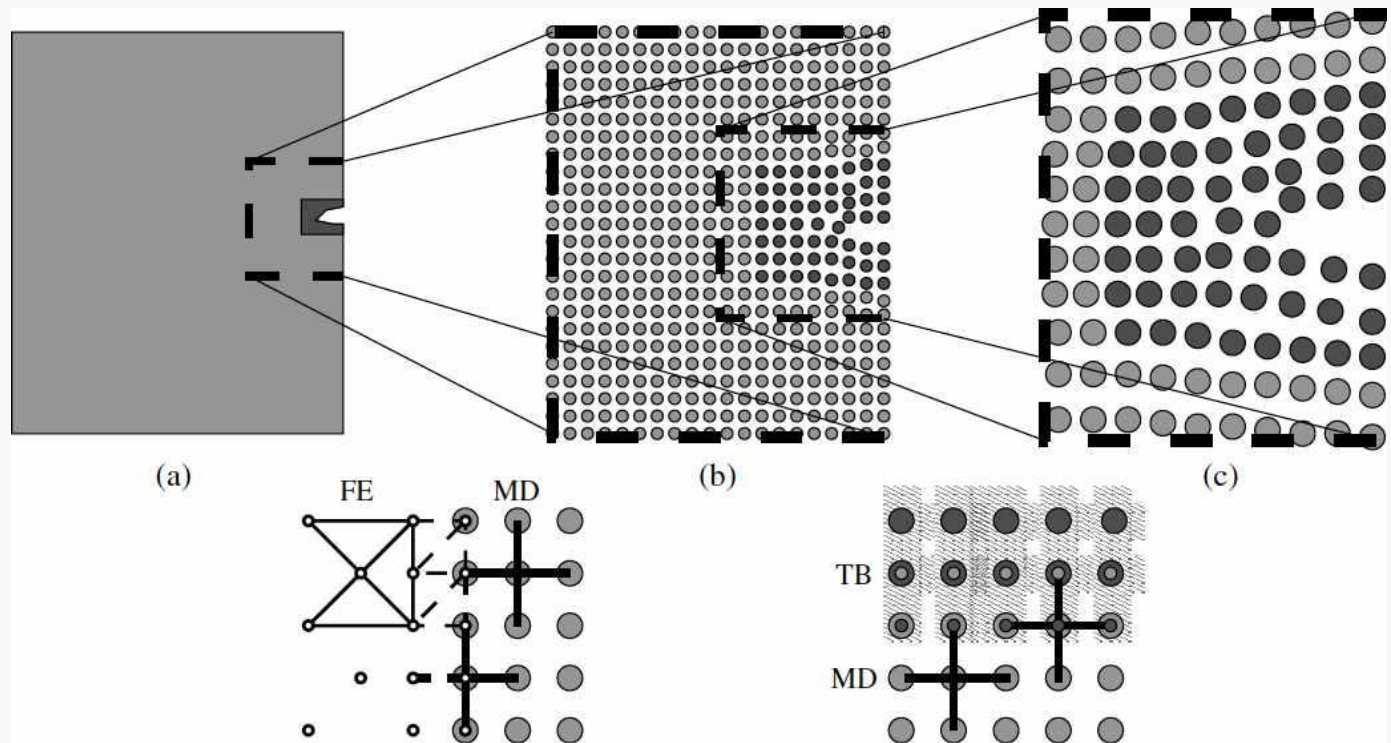
Concurrent multiscale techniques

- Sequential multiscale techniques require a separation of length and time scales, and prior knowledge of relevant physical processes.
- For many systems, the physics is inherently multiscale, with a strong coupling between the behavior occurring at different length/time scales.
- In such cases, it is no longer possible to integrate out degrees of freedom via approximate models as one moves from finer to coarser scales.
- “Multiscale models are also useful for gaining physical insight ... [and] can be an effective way to facilitate the reduction and analysis of data, which sometimes can be overwhelming.”
- G. Lu and E. Kaxiras, “Overview of Multiscale Simulations of Materials,” in *Handbook of Theoretical and Computational Nanotechnology*, M. Rieth and W. Schommers, eds (American Scientific Publishers, 2005).



Concurrent multiscale techniques

- “Onion” methods: finer length scale model regions are embedded within coarser scale regions
- The primary challenge is to maintain consistency between scales, with rigorous “handshaking” in overlap regions



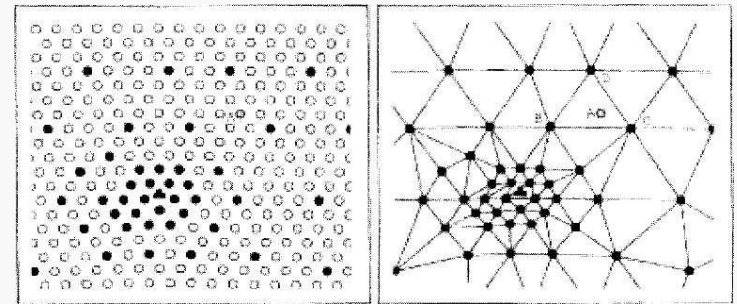
G. Lu and E. Kaxiras, “Overview of Multiscale Simulations of Materials,” in *Handbook of Theoretical and Computational Nanotechnology*, M. Rieth and W. Schommers, eds (American Scientific Publishers, 2005).

STATE-OF-THE-ART IN CONCURRENT MULTISCALE MATERIALS MODELING



Quasicontinuum (QC) method

- In regions of smoothly varying displacement (i.e., linear elastic deformation), full atomistic detail is replaced by representative atoms (“repatoms”).
- These provide the constitutive response for a finite element continuum model.



- Adaptive refinement to adjust repatom placement as needed during a dynamic simulation.
- Original development and vast majority of applications have been $T=0$; including finite temperature is challenging.

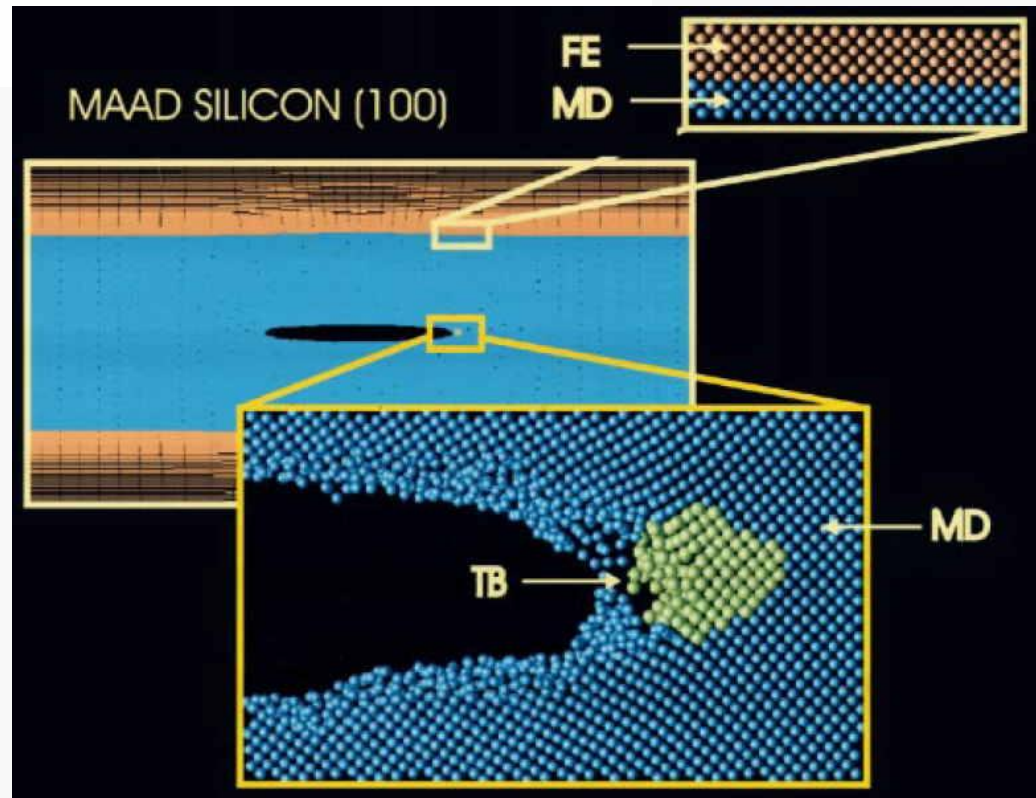
MAAD: Macroscopic, atomistic, *ab initio* dynamics

$$H_{\text{Tot}} = H_{\text{FE}}(\{\mathbf{u}, \dot{\mathbf{u}}\} \in \text{FE}) + H_{\text{FE/MD}}(\{\mathbf{u}, \dot{\mathbf{u}}, \mathbf{r}, \dot{\mathbf{r}}\} \in \text{FE/MD}) \\ + H_{\text{MD}}(\{\mathbf{r}, \dot{\mathbf{r}}\} \in \text{MD}) + H_{\text{MD/TB}}(\{\mathbf{r}, \dot{\mathbf{r}}\} \in \text{MD/TB}) \\ + H_{\text{TB}}(\{\mathbf{r}, \dot{\mathbf{r}}\} \in \text{TB}).$$

Geometric decomposition into 5 overlapping dynamic regions:

- Continuum finite element (FE)
- Atomistic molecular dynamics (MD)
- Quantum tight binding (TB)
- FE/MD "handshaking"
- MD/TB "handshaking"

Pseudohydrogen atoms to terminate silicon dangling bonds in TB regions



J.Q. Broughton, F.F. Abraham, N. Bernstein, and E. Kaxiras, "Concurrent coupling of length scales: Methodology and application," *Phys. Rev. B* **60**, 2391 (1999)

Coarse-grained molecular dynamics (CGMD)

- Addresses difficulties in a smooth transition between atomistic and continuum regions by replacing the continuum FE mesh with a continuum model developed by statistical coarse-graining.
- As the continuum mesh size approaches the atomistic scale, CGMD equations of motion become MD equations.
- Behavior derived solely from MD model
 - no continuum parameters
 - consistent treatment of phonon modes
 - smoother elastic wave propagation between regions
- Designed for finite-temperature dynamics

R.E. Rudd and J.Q. Broughton, "Coarse-grained molecular dynamics: Nonlinear finite elements and finite temperature," *Phys. Rev. B* **72**, 144104 (2005)



Heterogeneous Multiscale Method (HMM)

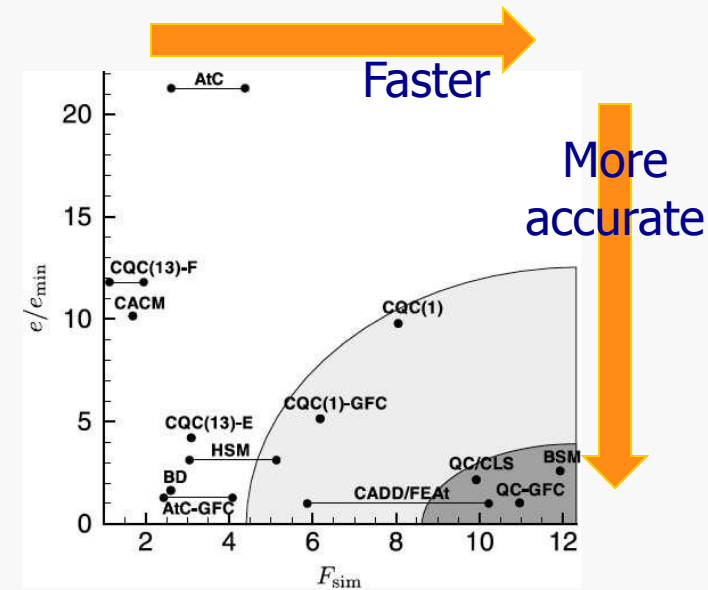
- Energy-based methods with coarse-grained Hamiltonians have several challenges:
 - Time scales between regions are still coupled
 - Matching conditions at boundaries often either cause spurious reflections or are expensive and non-scalable
 - Because of these, finite temperature, dynamical simulations are difficult.
- HMM philosophy is to use microscale models (e.g. MD) to supply missing data (e.g. constitutive laws and kinetic relations) for a macroscale solver (e.g. FEM)
- “Type A problems”: isolated defects treated via adaptive model refinement
- “Type B problems”: on-the-fly computation of constitutive information

X. Li and W. E, “Multiscale modeling of the dynamics of solids at finite temperature,” *J. Mech. Phys. Solids* **53**, 1650 (2005).



Performance and scalability of multiscale material methods needs to be assessed

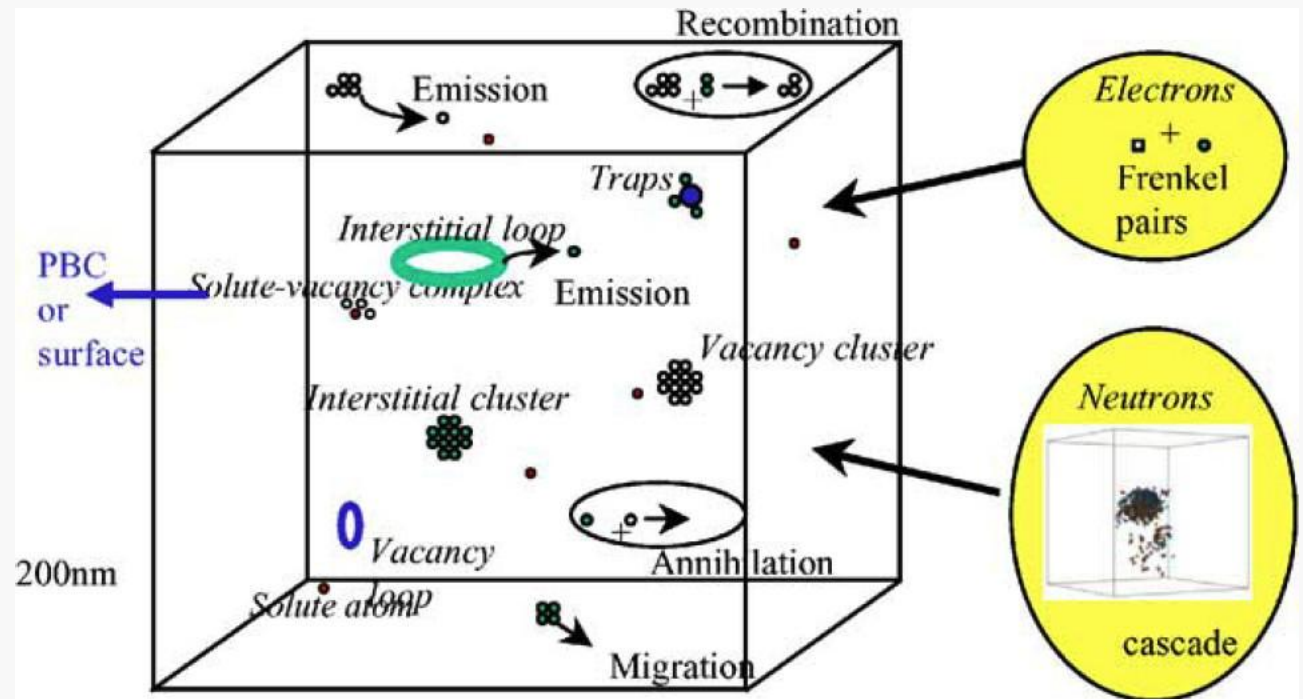
- “Multiscale methods like the ones discussed in this review show much promise to improve the efficiency of atomistic calculations, but they have not yet fully realized this potential. *This is in part because the focus to date has mainly been on development of the methodology as opposed to the large-scale application to materials problems.* ... In order for multiscale methods to compete with, or eventually replace, atomistics it is necessary that the methods be implemented in 3D, parallel codes optimized to the same degree as atomistic packages.”



R. E. Miller and E. B. Tadmor,
“A unified framework and
performance benchmark of
fourteen multiscale
atomistic/continuum coupling
methods,”
*Modelling Simul. Mater. Sci.
Eng.* **17**, 053001(2009)

Object kinetic Monte Carlo

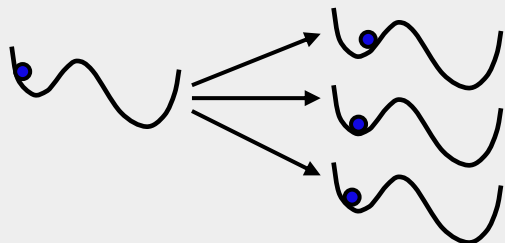
- Rather than a fully *atomistic* representation, one can track the formation, evolution, and interaction of *defects*, such as point defects, point defect clusters, solute atoms, and their sources and sinks (e.g. surfaces, grain boundaries, or dislocations).



C. Domain, C.S. Becquart, and L. Malerba, "Simulation of radiation damage in Fe alloys: an object kinetic Monte Carlo approach," *J. Nucl. Mater.* **335**, 121 (2004).

Accelerated Molecular Dynamics Methods

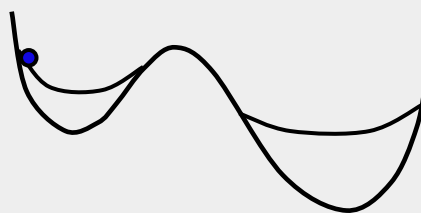
- In many cases, defect evolution pathways may be unknown
- AMD methods let the trajectory find an appropriate escape pathway



Parallel Replica Dynamics

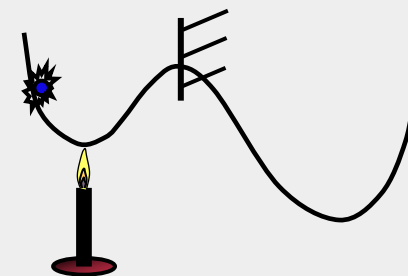
Parallelize time, by running an ensemble of trajectories

Most general technique



Hyperdynamics

Run thermostatted trajectory on a biased potential energy surface ($V+DV$), accumulate statistical "hyper-time"



Temperature Accelerated Dynamics

Increase system temperature to accelerate discovery of escape pathways

Most approximate technique

A.F. Voter, F. Montalenti, and T.C. Germann, "Extending the Time Scale in Atomistic Simulation of Materials," *Annu. Rev. Mater. Res.* **32**, 321 (2002).

WHY MULTISCALE (FROM A COMPUTATIONAL PERSPECTIVE)



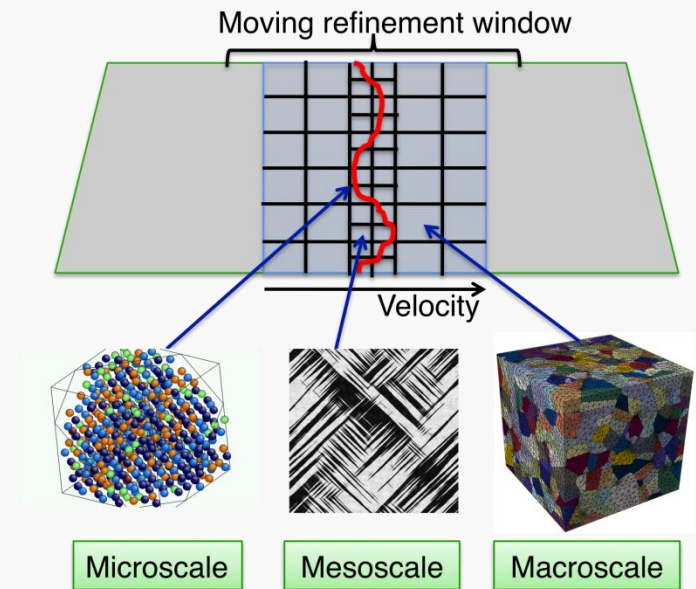
Preparing for exascale: issues to confront

- Computer architectures are becoming increasingly **heterogeneous** and **hierarchical**, with greatly increased flop/byte ratios.
- The algorithms, programming models, and tools that will thrive in this environment must mirror these characteristics.
- SPMD bulk synchronous (10^9 -way) parallelism will no longer be viable.
- Power, energy, and heat dissipation are increasingly important.
- Traditional global checkpoint/restart is becoming impractical.
 - Local flash memory?
- Fault tolerance and resilience
 - Recovering from soft and hard errors, and anticipating faults
 - MPI/application ability to drop or replace nodes
 - The curse of silent errors
- Analysis and visualization
 - *In situ*, e.g. "active storage" using I/O nodes?



Embedded Scale-Bridging Algorithms

- Our goal is to introduce more detailed physics into computational materials science applications in a way which escapes the traditional synchronous SPMD paradigm and exploits the heterogeneity expected in exascale hardware.
- To achieve this, we are developing a UQ-driven adaptive physics refinement approach.
- Coarse-scale simulations dynamically spawn tightly coupled and self-consistent fine-scale simulations as needed.
- This *task-based* approach naturally maps to exascale heterogeneity, concurrency, and resiliency issues.



WHERE ARE WE TRYING TO GO?



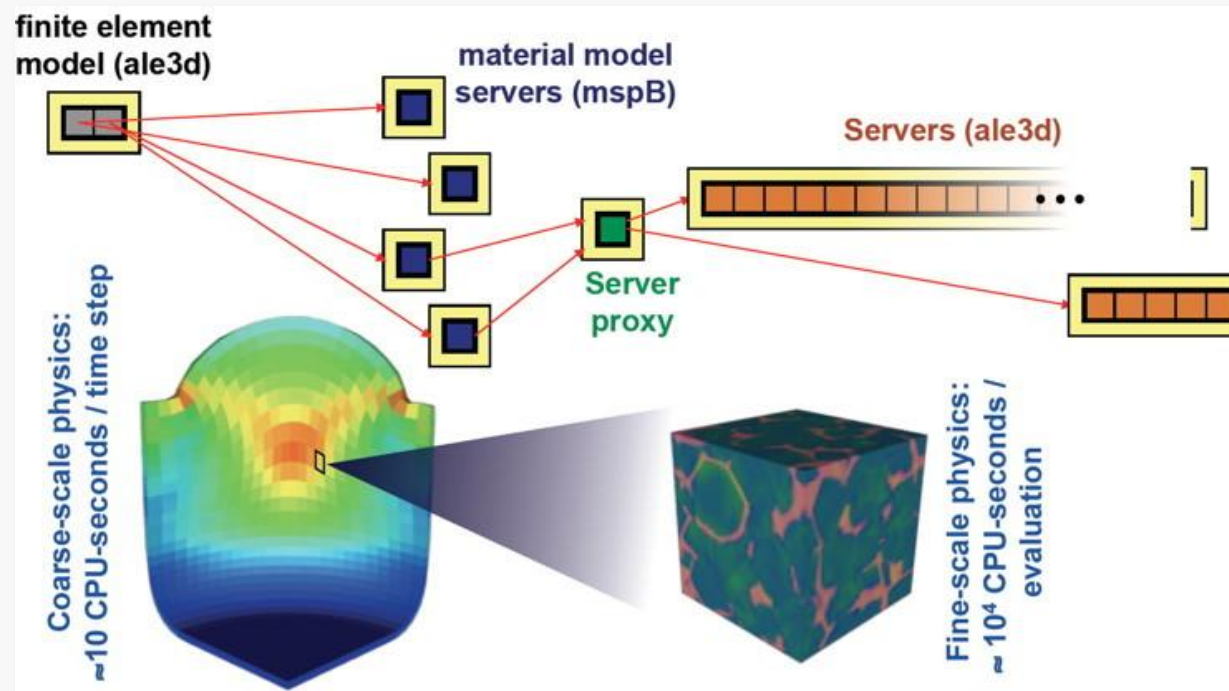
Embedded Scale-Bridging Algorithms

- Scale-bridging algorithms require a consistent two-way algorithmic coupling between temporally evolving distinct spatial levels; they are not "modeling", and not one-way information flow.
- Our focus is on coupling between macro (coarse-scale model) and meso (fine-scale model) scales with all unit physics being deterministic.
- We begin by building off of our adaptive sampling success, but move to the use of temporally evolving mesoscale and spatial adaption.
- Similar concepts apply in the time domain, e.g. using *ab initio* techniques to compute activation energies for a rate theory or kinetic Monte Carlo model ("on-the-fly kMC") applied to radiation damage modeling.



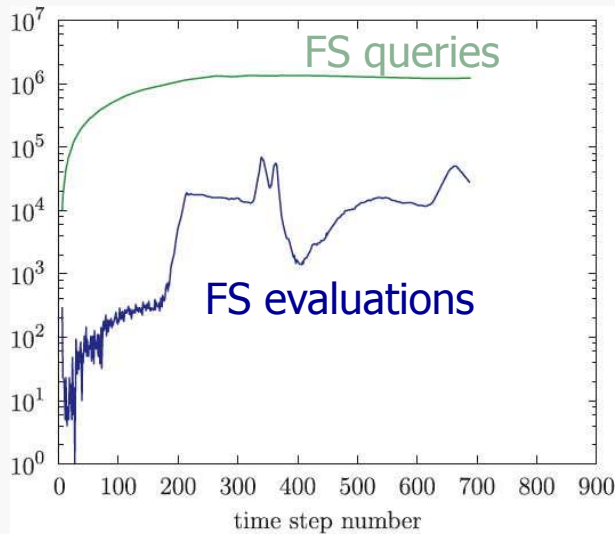
Adaptive sampling techniques have been successfully demonstrated by LLNL

- A coarse-scale model (e.g. FEM) calls a lower length-scale model (e.g. polycrystal plasticity) and stores the response obtained for a given microstructure, each time this model is interrogated
- A microstructure-response database is thus populated
- The fine-scale workload varies dramatically over the coarse-scale spatial and temporal domain
- Dynamic workload balancing in a task parallel context



N. R. Barton, J. Knap, A. Arsenlis, R. Becker, R. D. Hornung, and D. R. Jefferson.
Embedded polycrystal plasticity and adaptive sampling. *Int. J. Plast.* **24**, 242-266 (2008)

"A call to arms"



464 cores: 51x speedup

Table I. Machine configuration for 29 compute nodes (464 cores).

Component	Instances	Processes/instance	Total nodes
CS	1	192	12
ServerProxy	1	1	1
FS	8	32	16

2272 cores: 97x speedup

Table II. Machine configuration for 142 compute nodes (2272 cores).

Component	Instances	Processes/instance	Total nodes
CS	1	192	12
ServerProxy	2	1	2
FS	64	32	128

INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING
Int. J. Numer. Meth. Engng 2011; **86**:744–764
 Published online 1 February 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/nme.3071

A call to arms for task parallelism in multi-scale materials modeling[‡]

Nathan R. Barton^{1,*,*†}, Joel V. Bernier¹, Jaroslaw Knap², Anne J. Sunwoo¹,
 Ellen K. Cerreta³ and Todd J. Turner⁴

¹Lawrence Livermore National Laboratory, Livermore, CA 94550, U.S.A.

²U.S. Army Research Laboratory, Aberdeen Proving Ground, MD 21005, U.S.A.

³Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.

⁴U.S. Air Force Research Laboratory, Wright Patterson AFB, OH 45433, U.S.A.

SUMMARY

Simulations based on multi-scale material models enabled by adaptive sampling have demonstrated speedup factors exceeding an order of magnitude. The use of these methods in parallel computing is hampered by dynamic load imbalance, with load imbalance measurably reducing the achieved speedup. Here we discuss these issues in the context of task parallelism, showing results achieved to date and discussing possibilities for further improvement. In some cases, the task parallelism methods employed to date are able to restore much of the potential wall-clock speedup. The specific application highlighted here focuses on the connection between microstructure and material performance using a polycrystal plasticity-based multi-scale method. However, the parallel load balancing issues are germane to a broad class of multi-scale problems. Copyright © 2011 John Wiley & Sons, Ltd.

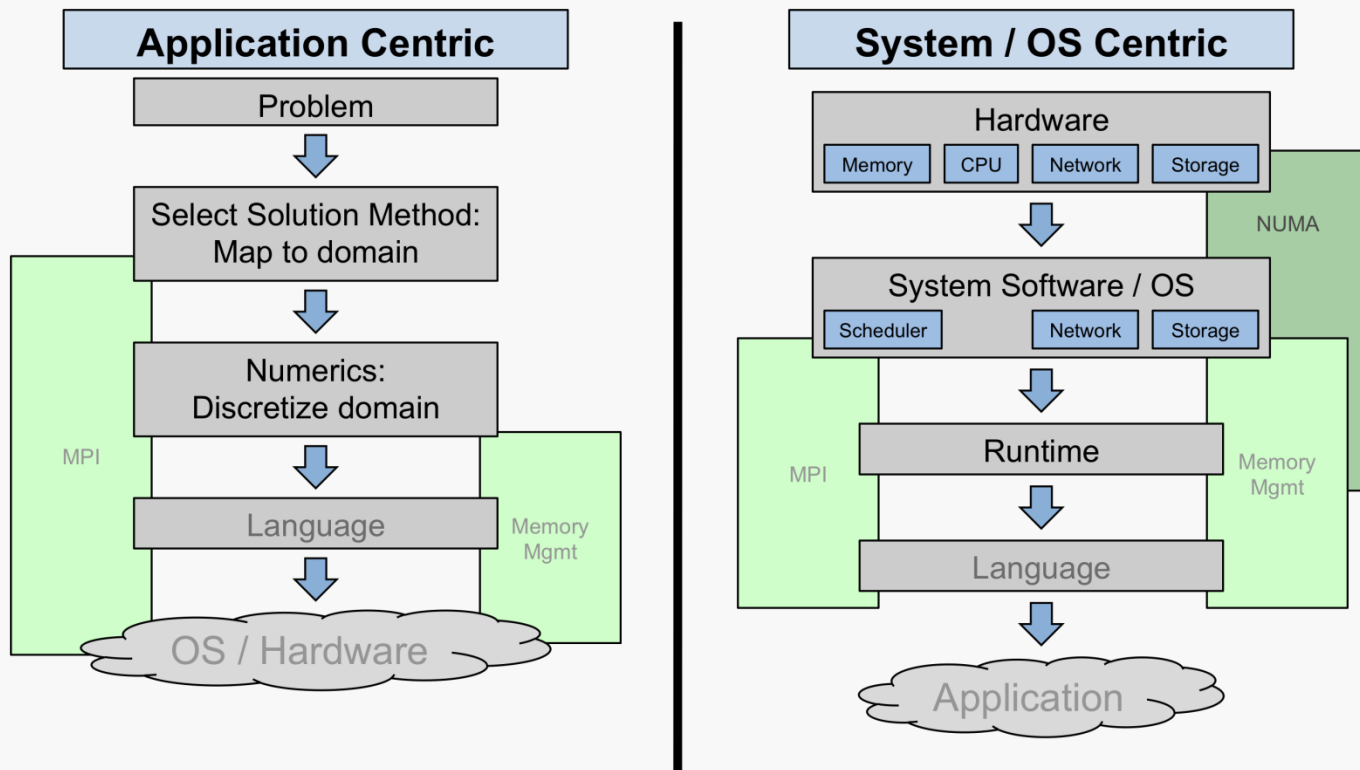


HOW CAN WE GET THERE FROM HERE?



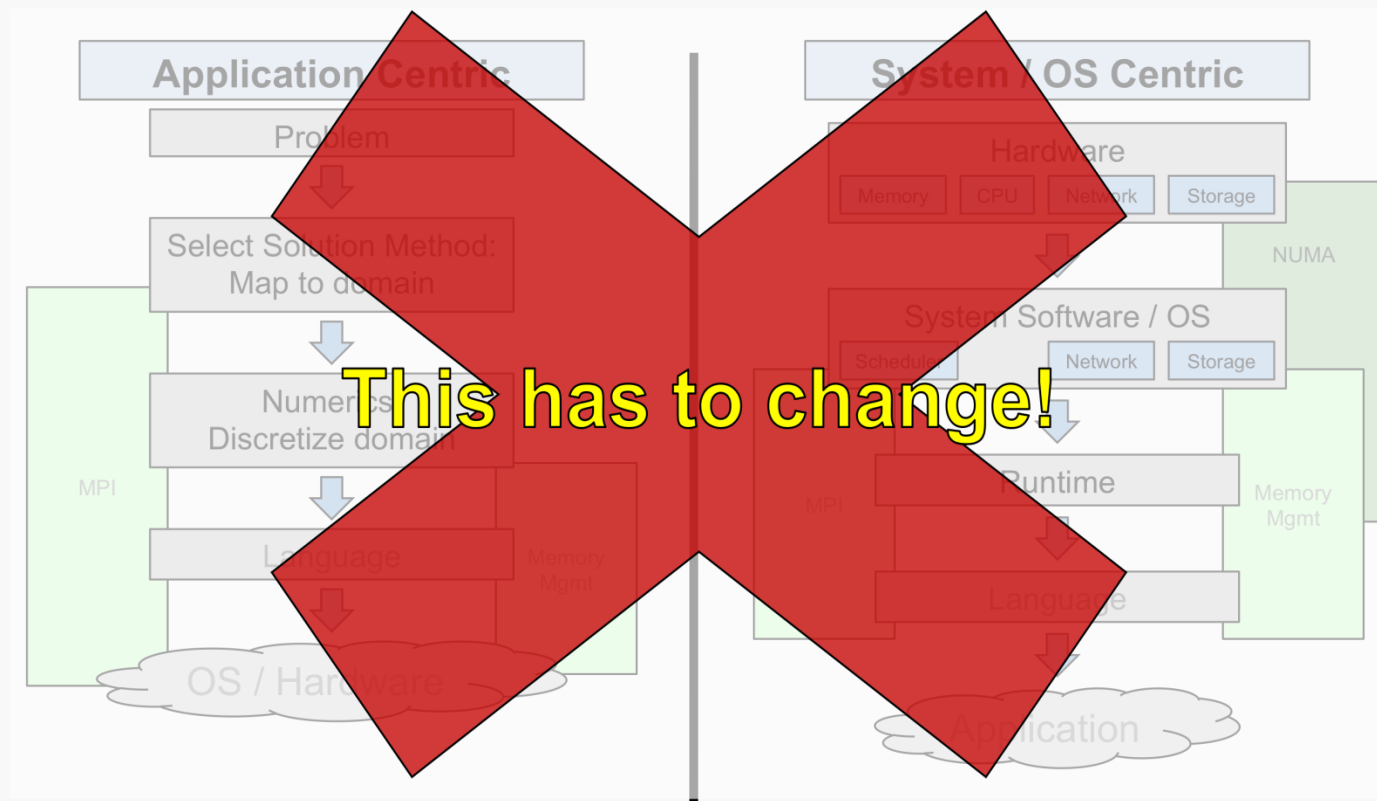
Current science application development strategies are difficult to sustain

An air gap has been encouraged between application developers & system / OS developers and the hardware



Current science application development strategies are difficult to sustain

An air gap has been encouraged between application developers & system / OS developers and the hardware



Co-design as a strategy for advancing science

- Co-design is a process by which a collection of computer science, applied math, and domain science experts work together to enable scientific discovery
- Hardware is changing dramatically
 - Increased concurrency
 - Increased heterogeneity
 - Reduced memory per core
 - 'Business as usual' is not going to work
- Algorithms and methods will have to be rethought / revisited
 - Flops are (almost always) free
 - Memory is at a premium
 - Power is a constraint for large scale systems
 - Requires co-operation between domain scientists, hardware manufacturers, and computer scientists to make progress
- Few domain scientists have the extended expertise 'from hardware to application' to enable applications to run at exascale
- Success on the next generation of machines will require extensive collaboration between domain scientists, computer scientists, and hardware manufacturers

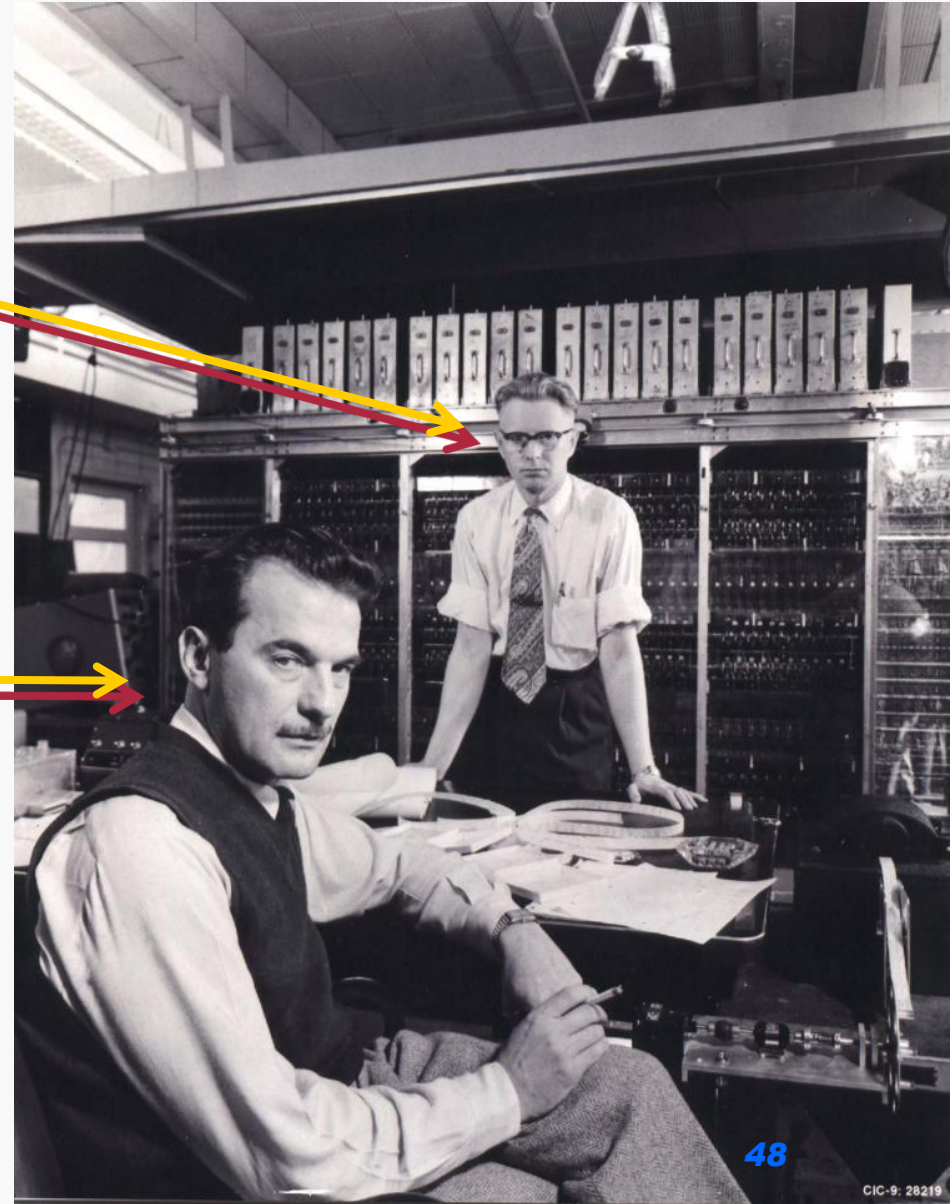


Los Alamos computational co-design, circa 1950

**Hardware architect
(Richardson)**

**Application scientist
(Metropolis)**

H. L. Anderson,
"Metropolis, Monte Carlo, and the MANIAC,"
Los Alamos Science **14**, 96-107 (1986).



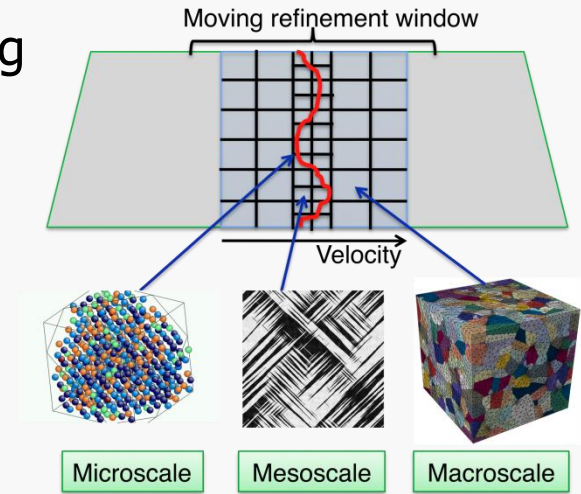
Co-design lessons learnt from Roadrunner

- Roadrunner was a leap into the future
 - First computer to reach a petaflop
 - First *heterogeneous* supercomputer
 - First *accelerated* supercomputer
 - Demonstrated that accelerated supercomputing was possible
 - 96% of compute power concentrated in accelerators
 - Success required domain scientists, applied mathematicians, and computer scientists working together to identify the correct abstractions for domain science, applied mathematics, programming models, and hardware
- Many successes including
 - Large Scale MD
 - Long time MD
 - Roadrunner Universe
 - DNS of turbulence
 - VPIC laser backscatter
 - VPIC Magnetic Reconnection
 - Supernova simulations
 - Phylogenetics of HIV



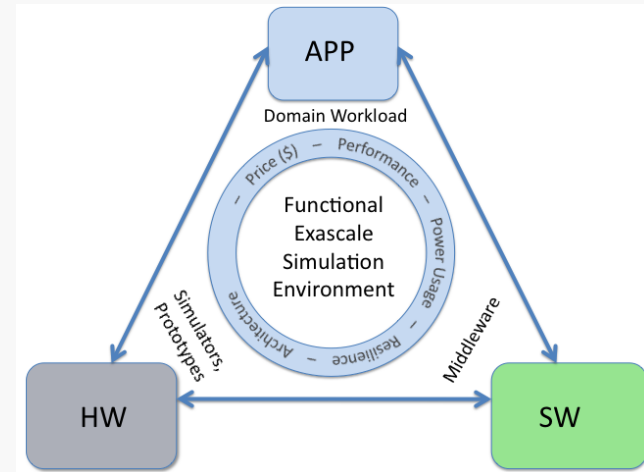
Exascale Co-design Center for Materials in Extreme Environments (ExMatEx)

- ASCR co-design center with a focus on developing a suite of Materials Science codes that will scale from laptops to exascale machines
- Large scale collaboration between national labs, industry and academia
- Goal is to enable more realistic large scale simulations with adaptive physics



ExMatEx Co-Design Project Goals

- Our **goal** is to establish the interrelationship between hardware, middleware (software stack), programming models, and algorithms required to enable *a productive exascale environment* for multiphysics simulations of materials in extreme mechanical and radiation environments.
- We will exploit, rather than avoid, the greatly increased levels of concurrency, heterogeneity, and flop/byte ratios on the upcoming exascale platforms.
- Our **vision** is an uncertainty quantification (UQ)-driven adaptive physics refinement in which meso- and macro-scale materials simulations spawn micro-scale simulations as needed.
 - This *task-based* approach leverages the extensive concurrency and heterogeneity expected at exascale while enabling fault tolerance within applications.
 - The programming models and approaches developed to achieve this will be broadly applicable to a variety of multiscale, multiphysics applications, including astrophysics, climate and weather prediction, structural engineering, plasma physics, and radiation hydrodynamics.



ExMatEx Co-Design Project Objectives

- **Inter-communication of requirements and capabilities between the materials science community and the exascale hardware and software community**
 - Proxy apps communicate the application workload to the hardware architects and system software developers, and are used in models/simulators/emulators to assess performance, power, and resiliency.
 - Exascale capabilities and limitations will be continuously incorporated into the proxy applications through an agile development loop.
 - Single-scale SPMD proxy apps (e.g. molecular dynamics) will be used to assess node-level data structures, performance, memory and power management strategies.
 - System-level data movement, fault management, and load balancing techniques will be evaluated via the asynchronous task-based MPMD scale-bridging proxy apps.
- **Perform trade-off analysis between competing requirements and capabilities in a tightly coupled optimization loop**
 - A three-pronged approach combining:
 - Node- to system-level models and simulators
 - Exascale emulation layer (GREMLIN) to introduce perturbations similar to those expected on future architectures
 - Performance analysis on leadership-class machines
 - Co-optimization of algorithms and architectures for price, performance, power (chiefly memory and data movement), and resilience (P³R)



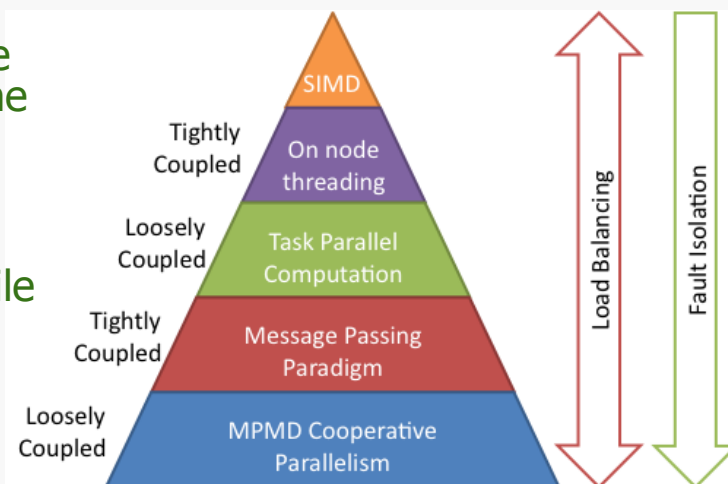
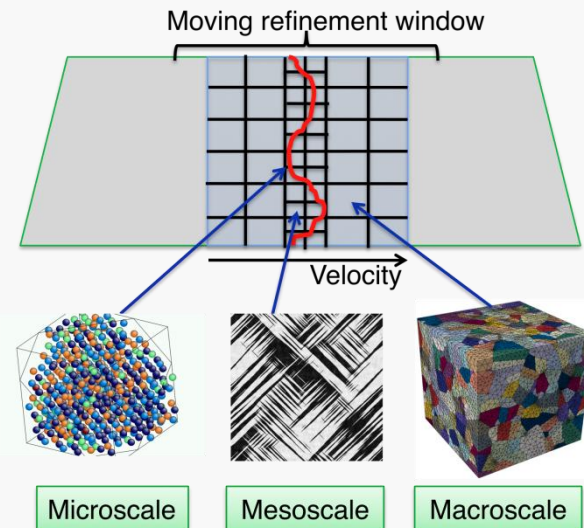
ExMatEx Co-Design Project Objectives

- **Full utilization of exascale concurrency and locality**

- Heterogeneous, hierarchical MPMD algorithms map naturally to anticipated heterogeneous, hierarchical architectures.
- Escape the traditional bulk synchronous SPMD paradigm, improve data locality and reduce I/O burden.

- **Application friendly programming models**

- Must expose hardware capabilities to the application programmer while at the same time hiding the continuous flux and complexity of the underlying hardware through a layer of abstraction that will aid portability.
- Task-based MPMD approach leverages concurrency and heterogeneity at exascale while enabling novel data models, power management, and fault tolerance strategies.



Summary

- Single-scale computational materials science codes have been useful not only for gaining scientific insight, but also as testbeds for exploring new approaches for tackling evolving challenges, including massive (nearly million-way) concurrency, an increased need for fault and power management, and data bottlenecks.
- No longer just “porting code” – the current technology revolution is a tremendous opportunity to fundamentally rethink our applications and algorithms.
- Scale-bridging methods are crucial from both application and computer science perspectives, and map well to the increasingly heterogeneous and hierarchical nature of computer architectures.
- Preparations for the exascale (10^{18} operations/second) era are underway by initiating an early and extensive collaboration between domain scientists, computer scientists, and hardware manufacturers – i.e., computational co-design.

