Exascale Computing: Science Prospects and Application Requirements

low-storage version, with most images removed

Doug Kothe

Sean Ahern, Sadaf Alam, Mark Fahey, Rebecca Hartman-Baker, Richard Barrett, Ricky Kendall, Bronson Messer, Richard Mills, Ramanan Sankaran, Arnold Tharrington, James White III (Trey)

Research sponsored by the Mathematical, Information, and Computational Sciences Division, Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC.

Oak Ridge National Laboratory

CUG 2008

National Center for Computational Sciences

Build on Town Hall report from DOE

http://www.er.doe.gov/ASCR/ProgramDocuments/TownHall.pdf

Interviewed computational scientists

- Pratul Agarwal
- Valmor de Almeida
- Don Batchelor
- Jeff Candy
- Jackie Chen
- David Dean
- John Drake
- Tom Evans
- Robert Harrison
- Fred Jaeger
- Lei-Quan Lee
- Wei-li Lee
- Peter Lichtner
- Phil Locascio
- Anthony Mezzacappa
- Tommaso Roscilde
- Benoit Roux
- Thomas Schulthess
- William Tang
- Ed Uberbacher
- Patrick Worley

Exascale findings

- Science prospects
  - Materials science
  - Earth science
  - Energy assurance
  - Fundamental science
- Requirements
  - Model and algorithm
  - Hardware
  - I/O
- Research and development needs

Materials science

- First-principles design of materials
  - Catalysts for energy production
  - Nano-particles for data storage and energy storage
  - High-temperature superconductors
- Predict behavior of aqueous environments (biological systems)

Earth science

- Direct simulation of physical and biochemical processes in climate
- Cloud-resolving atmospheres
- Decadal climate prediction
  - Regional impacts
  - Extreme-event statistics
- Socioeconomic feedbacks in climate
- Kilometer-scale basin simulations of supercritical CO_2 sequestration

Energy assurance

- Biomass recalcitrance (biofuels)
  - Plant cell-wall simulations of 100M atoms for milliseconds
- Closed fuel cycle for fission
- Whole-device model of ITER
- Biofuel combustion and emissions
- Optimal separating agents for nuclear material

Fundamental science

- Nucleosynthesis, gravity waves, and neutrino signatures of core-collapse supernovae
- Direct time-dependent simulation of nuclear fission and fusion processes
- Design and optimization of particle accelerators

Exascale findings

- Science prospects
  - Materials science
  - Earth science
  - Energy assurance
  - Fundamental science
- Requirements
  - Model and algorithm
  - Hardware
  - I/O
- Research and development needs

Model and algorithm requirements Colella's "7 Dwarfs*"

- Structured grids
- Unstructured grids
- Fast Fourier transforms (FFTs)
- Dense linear algebra
- Sparse linear algebra
- Particles
- Monte Carlo

** Dwarf population has now grown to 13, though new generation has arguable relevance to HPC.*

Current requirements

Application	Structured	Unstructured	FFT	Dense	Sparse	Particles	Monte Carlo
Molecular		X	X	X		X	
Nanoscience	X			X		X	X
Climate	X		X		X	X	
Environment	X	X			X		
Combustion	X						
Fusion	X		X	X	X	X	X
Nuc. energy		X		X	X		
Astrophysics	X	X		X	X	X	
Nuc. physics				X			
Accelerator		X			X		
QCD	X						X
#X	7	5	3	6	6	5	3

Exascale requirements

Application	Structured	Unstructured	FFT	Dense	Sparse	Particles	Monte Carlo
Molecular		X	X	X		X	X
Nanoscience	X			X		X	X
Climate	X		X		X	X	X
Environment	X	X			X	X	X
Combustion	X			X		X	
Fusion	X	X	X	X	X	X	X
Nuc. energy		X		X	X		
Astrophysics	X	X		X	X	X	
Nuc. physics				X			
Accelerator		X			X		
QCD	X						X
#X	7	6	3	7	6	7	6

Exascale requirements

Application	Structured	Unstructured	FFT	Dense	Sparse	Particles	Monte Carlo
Molecular		X	X	X		X	X
Nanoscience	X			X		X	X
Climate	X		X		X	X	X
Environment	X	X			X	X	X
Combustion	X			X		X	
Fusion	X	X	X	X	X	X	X
Nuc. energy		X		X	X		
Astrophysics	X	X		X	X	X	
Nuc. physics				X			
Accelerator		X			X		
QCD	X						X
#X	7	6	3	7	6	7	6

Broad use of all dwarfs

Exascale requirements

Application	Structured	Unstructured	FFT	Dense	Sparse	Particles	Monte Carlo
Molecular		X	X	X		X	X
Nanoscience	X			X		X	X
Climate	X		X		X	X	X
Environment	X	X			X	X	X
Combustion	X			X		X	
Fusion	X	X	X	X	X	X	X
Nuc. energy		X		X	X		
Astrophysics	X	X		X	X	X	
Nuc. physics				X			
Accelerator		X			X		
QCD	X						X
#X	7	6	3	7	6	7	6

None used by all applications

Exascale requirements

Application	Structured	Unstructured	FFT	Dense	Sparse	Particles	Monte Carlo
Molecular		X	X	X		X	X
Nanoscience	X			X		X	X
Climate	X		X		X	X	X
Environment	X	X			X	X	X
Combustion	X			X		X	
Fusion	X	X	X	X	X	X	X
Nuc. energy		X		X	X		
Astrophysics	X	X		X	X	X	
Nuc. physics				X			
Accelerator		X			X		
QCD	X						X
#X	7	6	3	7	6	7	6

Most growth

Suggestions for new dwarfs

- Adaptive mesh refinement
- Implicit nonlinear solvers
- Data assimilation
- Agent-based methods
- Parameter continuation
- Optimization

Current hardware requirements

- 12 hardware categories
- Choose:
  - 4 high priority (green)
  - 4 moderate priority (yellow)
  - 4 low priority (gray)

Current hardware requirements

Attribute	Climate	Astro	Fusion	Chemistry	Combustion	Accelerator	Biology	Materials
Node peak								
MTTI								
WAN BW								
Node memory								
Local storage								
Archival storage								
Memory latency								
Interconnect latency								
Disk latency								
Interconnect BW								
Memory BW								
Disk BW								

Exascale hardware requirements

- How will priorities *change*
- Choose:
  - 4 increasing priority (+)
  - 4 decreasing priority (-)
- Relative to current hardware requirements

Exascale hardware priorities

Attribute	Climate	Astro	Fusion	Chemistry	Combustion	Accelerator	Biology	Materials	sum
Node peak	–	+		+	+	–	–	+	+1
MTTI		+				+		+	+3
WAN BW	–	–	+	+		+	–	–	-1
Node memory	–	+			–	+			0
Local storage		+	–		–				-1
Archival storage			–			–		–	-3
Memory latency	+	–		–	+		+	+	+2
Interconnect latency	+	–		–	–	+	+	+	+1
Disk latency	–		–		–	–	–	–	-6
Interconnect BW	+	+	+	+	+		+		+6
Memory BW	+		+		+		+	+	+5
Disk BW			–	+	–	–	–		-3

Exascale hardware priorities

Attribute	Climate	Astro	Fusion	Chemistry	Combustion	Accelerator	Biology	Materials	sum
Node peak	−	+		+	+	−	−	+	+1
MTTI		+				+		+	+3
WAN BW	−	−	+	+		+	−	−	-1
Node memory	−	+			−	+			0
Local storage		+	−		−				-1
Archival storage			−			−		−	-3
Memory latency	+	−		−	+		+	+	+2
Interconnect latency	+	−		−	−	+	+	+	+1
Disk latency	−		−		−	−	−	−	-6
Interconnect BW	+	+	+	+	+		+		+6
Memory BW	+		+		+		+	+	+5
Disk BW			−	+	−	−	−		-3

Increasing priority

Exascale hardware priorities

Attribute	Climate	Astro	Fusion	Chemistry	Combustion	Accelerator	Biology	Materials	sum
Node peak	–	+		+	+	–	–	+	+1
MTTI		+				+		+	+3
WAN BW	–	–	+	+		+	–	–	-1
Node memory	–	+			–	+			0
Local storage		+	–		–				-1
Archival storage			–			–		–	-3
Memory latency	+	–		–	+		+	+	+2
Interconnect latency	+	–		–	–	+	+	+	+1
Disk latency	–		–		–	–	–	–	-6
Interconnect BW	+	+	+	+	+		+		+6
Memory BW	+		+		+		+	+	+5
Disk BW			–	+	–	–	–		-3

Decreasing priority

What were they thinking?

- About what they want?
- About what they expect?

Exascale hardware priorities

Attribute	Climate	Astro	Fusion	Chemistry	Combustion	Accelerator	Biology	Materials	sum
Node peak	–	+		+	+	–	–	+	+1
MTTI		+				+		+	+3
WAN BW	–	–	+	+		+	–	–	-1
Node memory	–	+			–	+			0
Local storage		+	–		–				-1
Archival storage			–			–		–	-3
Memory latency	+	–		–	+		+	+	+2
Interconnect latency	+	–		–	–	+	+	+	+1
Disk latency	–		–		–	–	–	–	-6
Interconnect BW	+	+	+	+	+		+		+6
Memory BW	+		+		+		+	+	+5
Disk BW			–	+	–	–	–		-3

Decreasing I/O priority?

Decreasing I/O priorities

- I/O doesn't need to keep up with other hardware improvements?
(much evidence to the contrary)
- Or I/O isn't *expected* to keep up (even though it may need to)?

Disruptive hardware technologies

- 3D chips and memory
- Optical processor connections
- Optical networks
- Customized processors
- Improved packaging
  - On chip, on node board, within cabinets

I/O imbalance

Exascale I/O requirements

- Two categories
  - Output of restart files and analysis files
  - Postprocessing for analysis and visualization
- Consider
  - 1 EF computer
  - 100 PB memory
  - Restart and analysis data = 20% of memory
  - Write data once per hour
  - I/O should take 10% or less of runtime

Exascale I/O requirements

- Disk bandwidth
  - 50 TB/s
  - 5 TB/s if asynchronous, overlapping with compute
- Disk capacity
  - 6 EB for 3 weeks of data
- Archive bandwidth
  - 1 TB/s write
  - 2 TB/s read (to speed up analysis)

Exascale analysis requirements

- Memory of analysis system
  - Assume we need 1/100 of all data from the run
  - Assume another 1/100 from out of core and streaming
  - 200 TB
- Memory of analysis system (another way)
  - One full time step, 10% of memory, 10 PB
  - Some say it's more like 2.5%, 2.5 PB
- Shared memory?
- Better network latency?

Reducing I/O requirements

- Recompute instead of store
- Checkpoint in memory
- Analyze data during computation
- Overlap I/O and computation

Exascale findings

- Science prospects
  - Materials science
  - Earth science
  - Energy assurance
  - Fundamental science
- Requirements
  - Model and algorithm
  - Hardware
  - I/O
- Research and development needs

R&D needs

- Automated diagnostics
- Hardware latency
- Hierarchical algorithms
- Parallel programming models
- Accelerated time integration
- Model coupling
- Solver technology
- Maintaining current libraries

Automated diagnostics

- Aggressive automation of diagnostic instrumentation, collection, analysis
- Drivers
  - Performance analysis
  - Application verification
  - Software debugging
  - Hardware-fault detection and correction
  - Failure prediction and avoidance
  - System tuning
  - Requirements analysis

Hardware latency

- Expect improvement: aggregate computation rate, parallelism, bandwidth
- Not so much: hardware latency
- Software strategies to mitigate high latency
- Fast synchronization mechanisms
  - On chip, in memory, or over networks
- Smart networks
  - Accelerate or offload latency-sensitive operations
  - Example: semi-global floating-point reductions

Hierarchical algorithms

- Stagnant latencies → memory hierarchies
- Heterogeneous computing
  → process hierarchies
- Fault tolerance
  → redundancy higher in each hierarchy
- Need hierarchy-aware algorithms
  - Recompute versus load/store
  - Fine-scale hybrid task and data parallelism
  - In-memory checkpointing

Parallel programming models

- Current models target one level of memory hierarchy at a time
  - Source language for instruction-level parallelism
  - OpenMP for intra-node parallelism
  - MPI for inter-node parallelism
  - New levels?
- More coupling of complex models
  - Arbitrary hierarchies of task and data parallelism
- Latency stagnation
  - Minimize synchronization, maximize asynchrony
- New programming model?
  - Easily allow arbitrary number of levels of hierarchy
  - Map hierarchy to hardware at runtime (dynamically?)

Accelerated time integration

- Many applications need more time steps
- Single-process performance stagnating
- Increasing resolution shrinks time steps
- Parallelism doesn’t help (time is serial)
- See presentation tomorrow
  “Accelerating Time Integration”
  Session 12A, this room, 11:15 AM

Model coupling

- Models coupled into more-complete, more-complex models
- Implement, verify, and validate coupling
- Upscaling, downscaling, nonlinear solving
- Uncertainty analysis, sensitivity analysis
- Data assimilation
  - Growing volume of data from satellites and sensors

Solver technology

- More physical processes
- Coupled strongly and nonlinearly
- Latency stagnation → local preconditioners
- Trade flops for memory operations
  → (hierarchical) block algorithms
- Tune advanced algorithms for hierarchies

Maintaining current libraries

- BLAS, MPI, and everything else
- Tune and update for new architectures
- Critical for usability

More information

nccs.gov → Media Center → NCCS Reports

http://www.nccs.gov/media-center/nccs-reports/