NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Challenges on the Path to Exaflop/s Computing

Horst D. Simon

LBNL

October 21, 2007

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Challenges on the Path to Sustained Exaflop/s Computing for Science

Horst D. Simon

LBNL

October 21, 2007

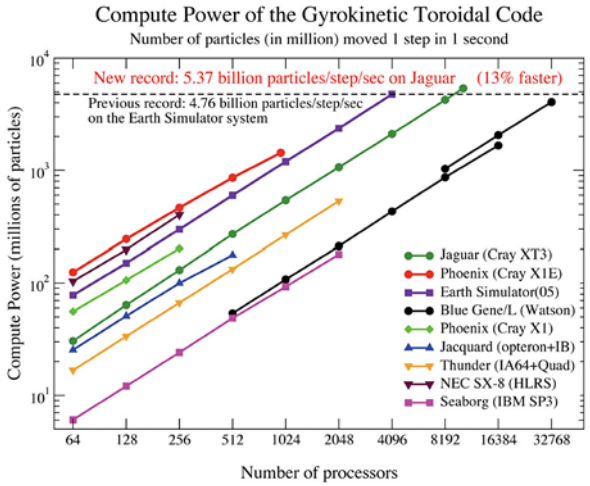
Seven Challenges to Reach Sustained Exaflop/s for Science

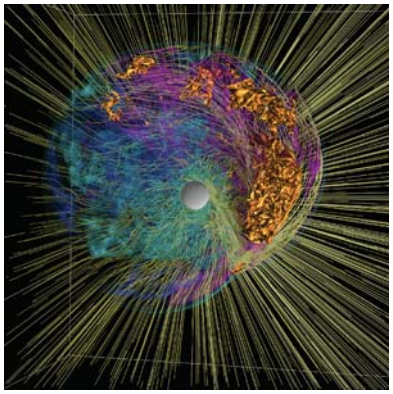
1. There is currently no R&D program that supports production computing beyond Petaflops
   - DARPA HPCS appears to be the last funded program to support high end production computing

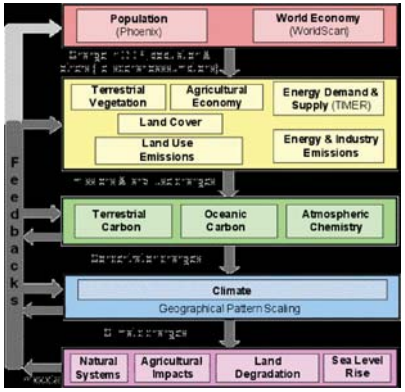
Computational Requirements for Science Are Clear

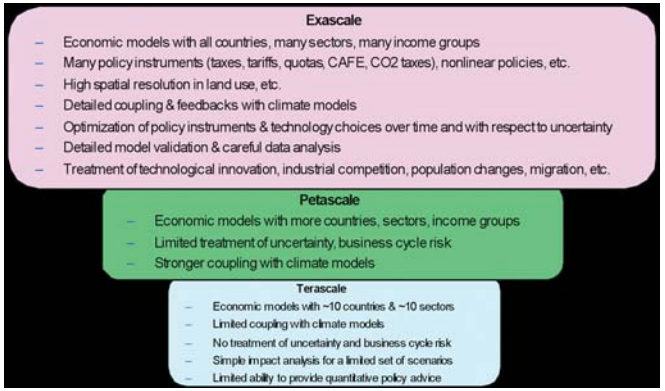
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

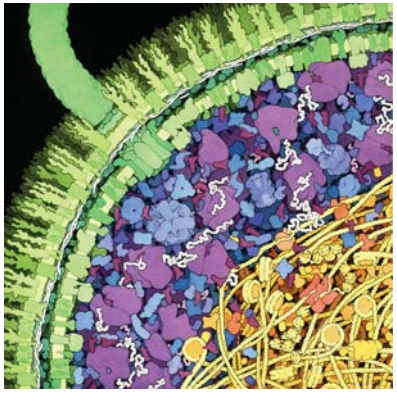
Modeling and Simulation at the Exascale for Energy and the Environment has significant requirements for Exascale

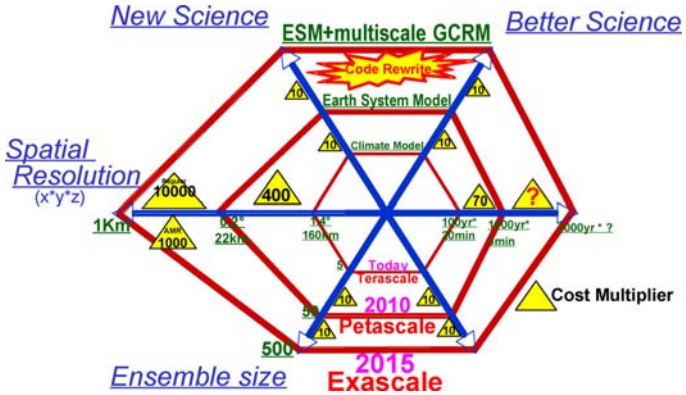












Scientists Need More Than FLOP/s

- **Performance — How fast will a system process work if everything is working well**
- **Effectiveness — What is the likelihood that users can get the system to do their work**
- **Reliability — The system is available to do work and operates correctly all the time**
- **Consistency — How often will the system process users' work as fast as it can**
- **Usability — How easy is it for users to get the system to go as fast as possible**

Seven Challenges to Reach Sustained Exaflop/s for Science

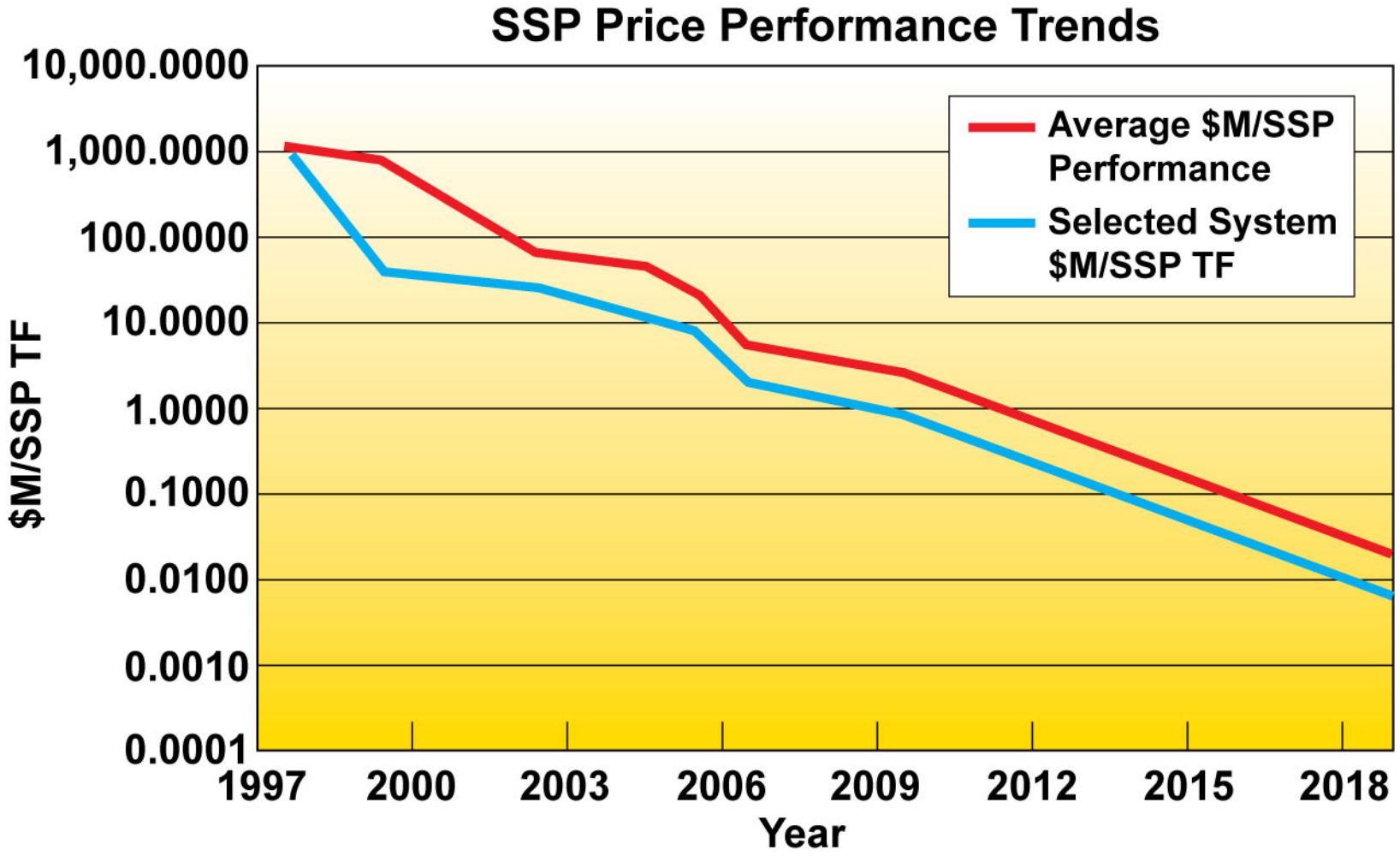
2. The community cannot afford Exaflop/s computing until 2019 at current budget levels

How We Estimate Future Computational System Costs

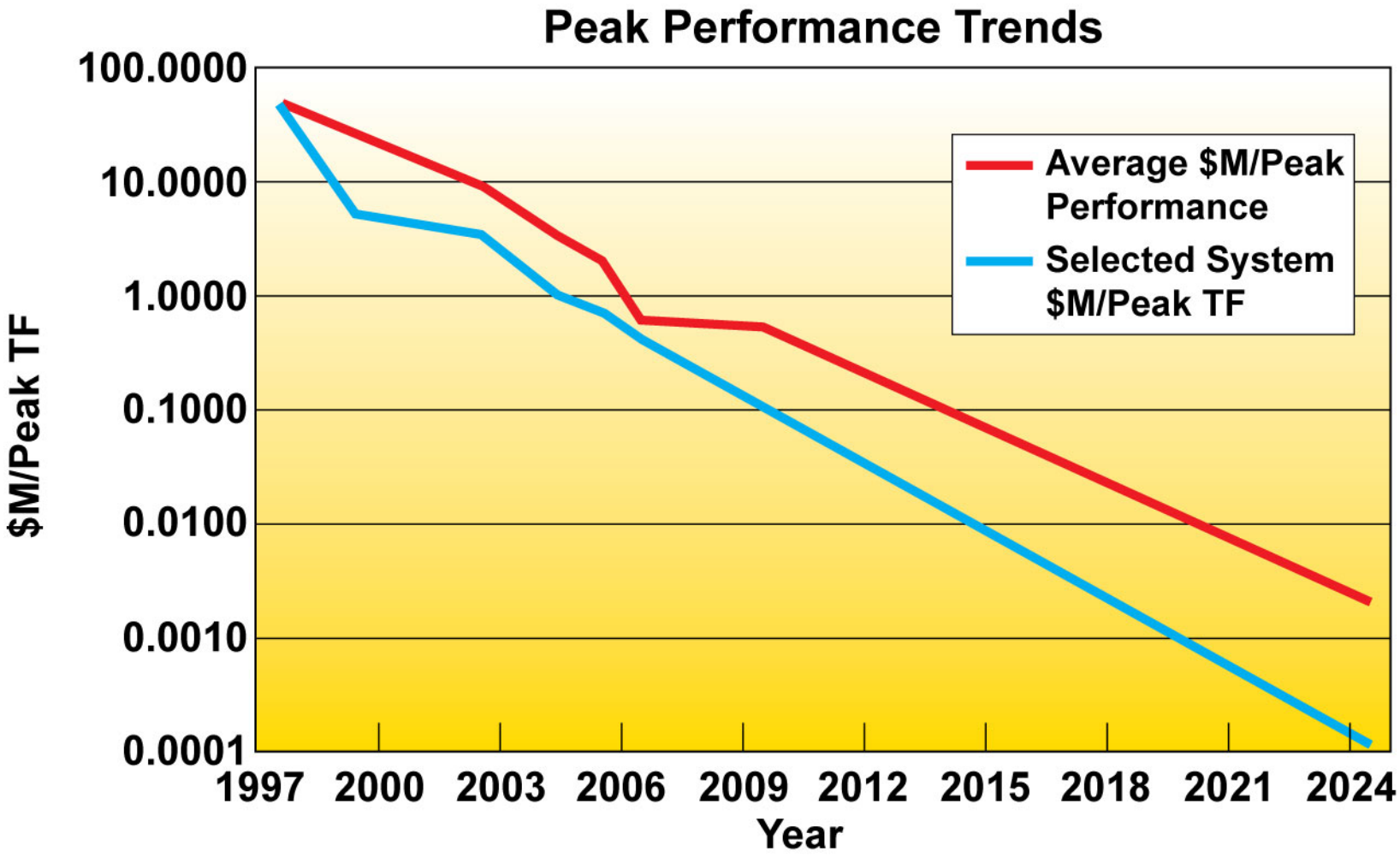
Costs for Exaflop/s in FY 2016

- 10 years of performance (based on the NERSC-5 SSP) and pricing data (from bids)
- Performance is based on the NERSC SSP
  - Composite of the performance of a selection of applications that represent the overall workload
    - AY07 ERCAP indicated almost 900 applications used at NERSC
    - SSP used seven representative and major applications
- Cost data
  - System and TCO over three years based on bids NERSC receives
- SSP-4 (used for NERSC-5) is correlated with other SSP versions
  - SSP usually underestimated real performance overall
  - Averages 10-12% of peak over T3E to XT4 but has improved with time 4% to >16%
- Also have peak performance data

10 Years of Sustained Price Performance Information



10 Years of Peak Price Performance Information



Costs for Exaflop/s in FY 2016

- It will take twelve years to move from the first peak terascale system (ASCI Red in 1997) to the first peak petascale system (hopefully 2009)
  - Achieving exascale six years after petascale is possible — but costly
- An exaflop/s peak system in service for FY 2016 will cost ~$750M
  - The same data projects a petaflop/s **sustained** system installed in FY 2011 would cost ~$190M (matches $208M NSF contracted)
  - Aligned to the costs of the LCFs
  - A **sustained** exaflop/s for the Office of Science workload would cost $3.7B
- An exaflop/s **sustained** system in service in FY 2019 will cost ~$800M
  - To be in service in FY 2019, purchase and installation would be in second half of FY 2018
- A peak exaflop/s system in this time frame is $180M

Seven Challenges to Reach Sustained Exaflop/s for Science (3)

3. The community cannot afford the power requirements for Exaflop/s computing

New Design Constraint: *POWER*

- Transistors still getting smaller
  - Moore's Law is alive and well
- But Denard scaling is dead!
  - No power efficiency improvements with smaller transistors
  - No clock frequency scaling with smaller transistors
  - All "magical improvement of silicon goodness" has ended
- Traditional methods for extracting more performance are well-mined
  - Cannot expect exotic architectures to save us from the "power wall"

Power Demands Threaten to Limit the Future Growth of Computational Science

- LBNL Study for Climate Requirements in 2008
  - Extrapolation of Blue Gene and AMD design trends
  - Estimate: 20 MW for BG and 179 MW for AMD
- DOE E3 Report
  - Extrapolation of existing design trends
  - Estimate: 130 MW
- DARPA Study
  - More detailed assessment of component technologies
  - Estimate: 20 MW just for memory alone, 60 MW aggregate so far

Path to Power Efficiency
Reducing Waste in Computing

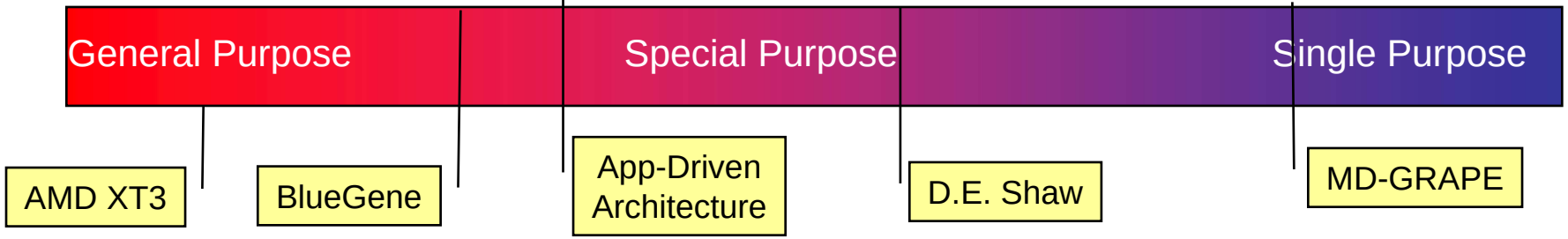
- **Examine methodology of low-power embedded computing market**
  - **optimized for low power, low cost, and high computational efficiency**

"Years of research in low-power embedded computing have shown only one design technique to reduce power: ***reduce waste.****"*

—Mark Horowitz, Stanford University & Rambus Inc.

- **Sources of Waste**
  - **Wasted transistors (surface area)**
  - **Wasted computation (useless work/speculation/stalls)**
  - **Wasted bandwidth (data movement)**
  - **Designing for serial performance**

Customization Continuum



- **Application-driven architecture does not NOT necessitate a special purpose machine!**
- **D.E. Shaw System: Semicustom design with some custom elements**
  - **Uses fully programmable cores with full-custom co-processors to achieve efficiency**
  - **Simulate 100x–1000x longer timescales than ANY feasible HPC system using 20 kilowatts**
  - **Programmability broadens application reach (but narrower than our approach)**
- **MD-Grape: Full custom ASIC design**
  - **1 petaflop performance for one application using 26 kilowatts**
  - **Cost $9M from concept to implementation**
- **Application-Driven Architecture (Climate Simulator): Semicustom design**
  - **Highly programmable core architecture using C/C++/Fortran**
  - **100x better power efficiency is modest compared to demonstrated capability of more specialized approaches!**

Climate Strawman System Design In 2008

- Design system around the requirements of the massively parallel application
- Example: kilometer scale climate model application

We examined three different approaches:

- AMD Opteron: Commodity approach, lower efficiency for scientific applications offset by cost efficiencies of mass market
- BlueGene: Generic embedded processor core and customize system-on-chip (SoC) services to improve power efficiency for scientific applications
- Tensilica: Customized embedded CPU as well as SoC provides further power efficiency benefits but maintains programmability

Processor	Clock	Peak/ Core (Gflops)	Cores/ Socket	Mem/ BW (GB/s)	Network BW (GB/s)	Sockets	Power	Cost 2008
AMD Opteron	2.8GHz	5.6	2	6.4	4.5	890K	179 MW	
IBM BG/P	850MHz	3.4	4	5.5	2.2	740K	20 MW	
Climate computer	650MHz	2.7	32	51.2	34.5	120K	3 MW	$75M

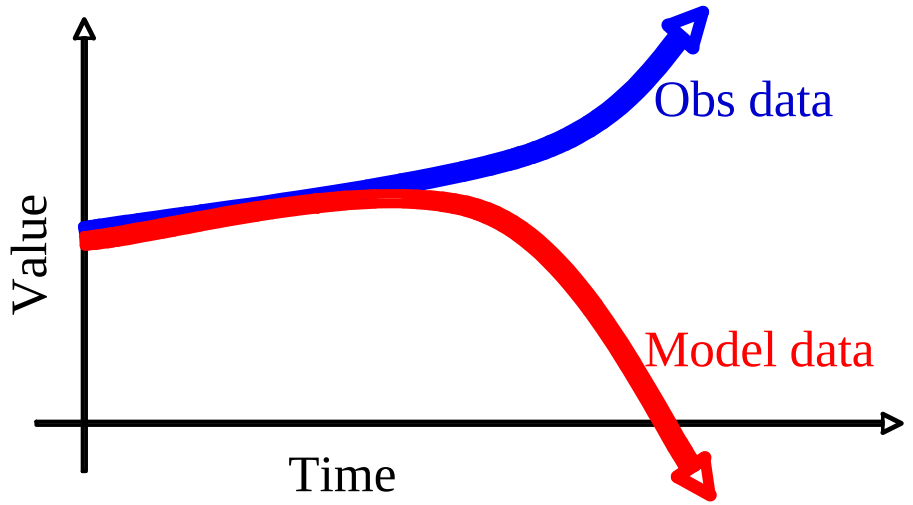
Solve an exascale problem without building an exaflop/s machine!

Seven Challenges to Reach Sustained Exaflop/s for Science (4)

4. **Productivity for science, such as an integrated data environment does not get adequately addressed**

Data Tsunami

- Soon it will no longer be sufficient for NERSC to rely solely on center balance and HPSS to address the massive volumes of data on the horizon
- The volume and complexity of experimental data will overshadow data from simulation
  - LHC
  - ITER
  - JDEM/SNAP
  - PLANCK
  - SciDAC
  - JGI
  - Earth Systems Grid

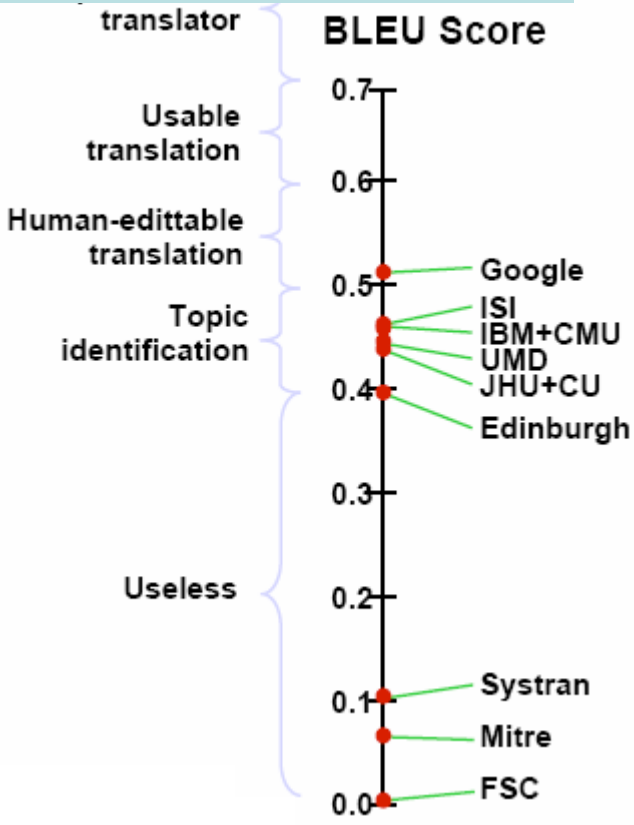


Courtesy L. Buja

Easy Access to Data Accelerates Science

- MapReduce is applicable to science analytics
  - Apply functions on numeric, image, text data
  - E.g., telesopes, genome, simulations,…
- Simple, extensible interface
  - Allows for domain-specific analysis
  - Leverages domain-independent infrastructure
- Efficient use of wide area bandwidth
  - Ship functions to raw data; return filtered information

Arabic translation: Google with more data beats others with more specialists



NERSC Data Program Elements

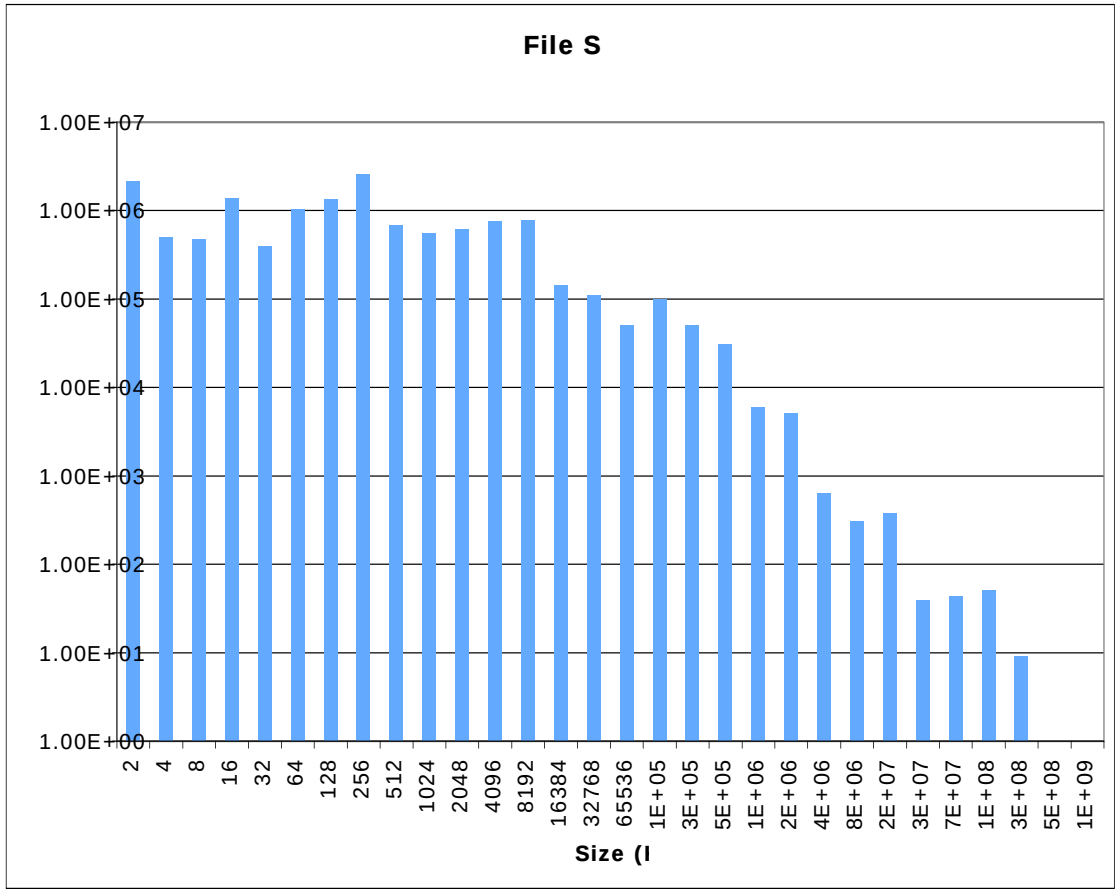
- Next-generation mass storage
- Infrastructure for data
  - Hardware: computational platforms
  - Software for data management, analysis/analytics, interfaces between integrated data components
- Develop or adapt reusable, broad-impact tools
  - Analogous to Google Earth, Microsoft SharePoint
  - Host and adapt SciDAC tools for science community
- User and project expertise
  - Consulting expertise in scientific data management, analytics, visualization, workflow management, etc.

Storage Common Wisdom

- Old
  1. Users have a small number of large files
  2. Files are the lowest level unit of storage
  3. We need to cause users pain to move their files from place to place
  4. Users have all the files they need in each place they compute
  5. One system is sufficient for all the steps a workflow
- New
  1. Large numbers of small files dominate performance
  2. Objects are the lowest unit of storage
  3. It is more productive to systems and users to let systems to manage the placement of files
  4. User's have data in many places and need to move the data frequently - even within a facility
  5. Job steps are best run on systems with the most appropriate balance

NERSC Data Program Elements: Data Storage

- The current ways of storing data will not scale to exascale
  - Global filesystems
  - Archival storage
- Archival storage systems were designed to requirements that are 15 years old
  - Very unclear whether they can stretch to the exascale



Seven Challenges to Reach Sustained Exaflop/s for Science (5)

5. Application parallelism at 100K way parallel is hard - disconnect between productive science and easy scaling

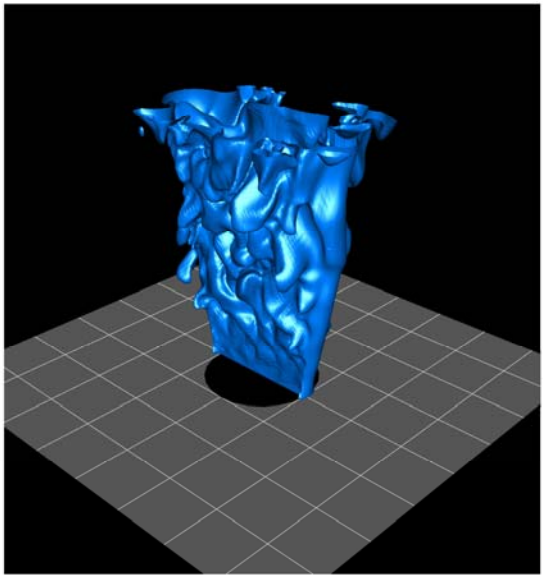
DOE Computational Science Is Diverse

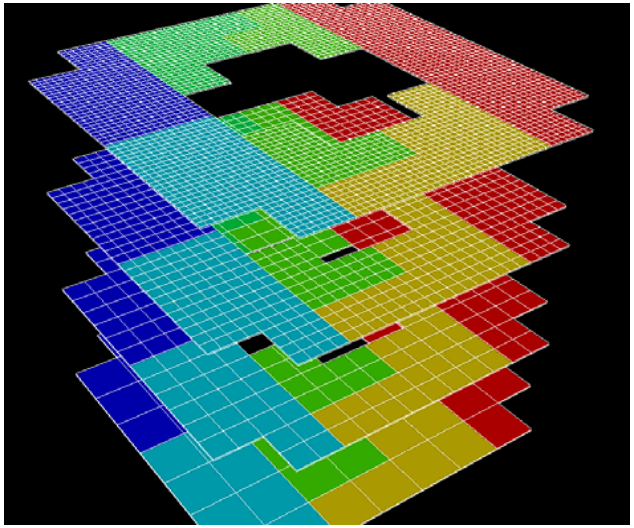
Science areas	Multi-physics, Multi-scale	Dense linear algebra (DLA)	Sparse linear algebra (SLA)	Spectral Methods (FFT)s (SM-FFT)	N-Body Methods (N-Body)	Structured Grids (S-Grids)	Unstructured Grids (U-Grids)	Data Intensive (Map Reduce)
Nanoscience	X	X	X	X	X	X		
Chemistry	X	X	X	X	X			
Fusion	X	X	X			X	X	X
Climate	X		X	X		X	X	X
Combustion	X		X			X	X	X
Astrophysics	X	X	X	X	X	X	X	X
Biology	X	X					X	X
Nuclear		X	X		X			X
System Balance Implications	General Purpose balanced System	High Speed CPU, High Flop/s rate	High Performance Memory	High Interconnect Bisection bandwidth	High Performance Memory	High Speed CPU, High Flop/s rate	Irregular Data and Control Flow	High Storage and Network bandwidth

Algorithm Choices: A Case Study
by Phil Colella, LBNL, and SciDAC APDEC Center

AMR Low-Mach-Number Combustion (LMC) Algorithm

- Discretization of the Ma ! 0 fluid equations with detailed hydrocarbon chemistry and transport. For methane (GRIMech 3.0, EGLib), 60 unknowns per grid point.
- Used to investigate a variety of turbulent flames. Extensions to simulate syngas combustion, nuclear burning in supernovae.
- Computational time dominated by solving ODEs at every grid point for chemistry using an implicit solver. Single-level variable-coefficient elliptic solvers to impose divergence constraint that replaces acoustic wave dynamics. The latter use multigrid-preconditioned BiCGStab.

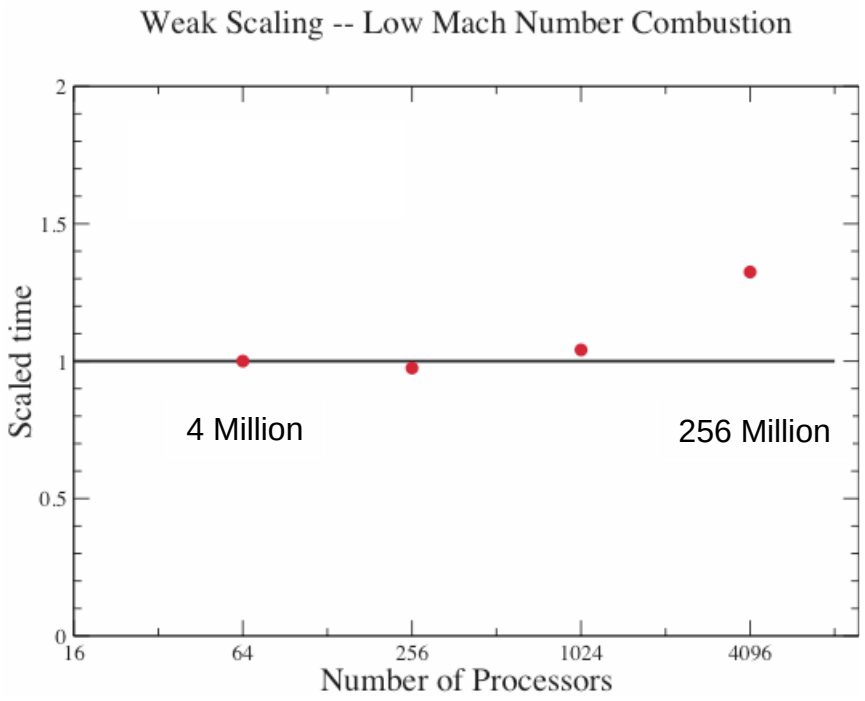




LMC Benchmark

• LMC Replication Benchmark: Single image is a wrinkled flame. Two levels of refinement, factor of 2 each, refinement in time. Total of 4M grid points.

- For $N_{proc} \cdot 1024$, the cost of the computation is dominated by the cost of solving the chemical rate equations. By rebalancing the data for this task on the fly, this part of the computation scales perfectly.
- Variable-coefficient elliptic solvers used here are leading to a loss of weak scaling of the whole application for larger numbers of processors.
- Similar scaling results obtained on a Linux cluster, and on an SGI system (NASA Columbia).



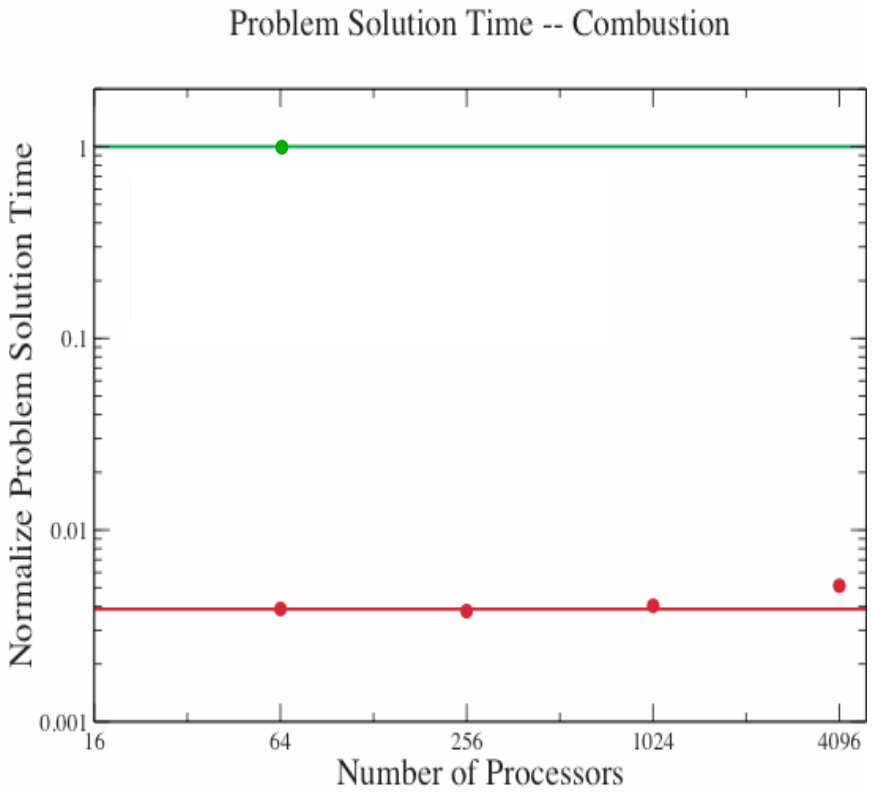
LMC vs. Fully Explicit Method

- Algorithmic choices are driven by science requirements: of the available alternatives, which is going to provide the most scientific output for the least cost ? To address this question, we compared LMC to a version of what has been the standard approach to solving these problems for the last 20 years.

- Fully explicit method for viscous compressible flow on a uniform grid:
  - Explicit stencil operations scale perfectly.
  - Time step is determined by CFL condition for acoustic waves (.02 μsec).
  - For chemistry, use explicit ODE method, subcycle in time as needed.
- LMC:
  - Elimination of acoustic waves leads to a 50X increase in the time step (1 μsec). This comes at the cost of introducing elliptic solvers and the accompanying loss of ideal scaling.
  - AMR provides 10X reduction in the number of grid points over a uniform fine grid with the same resolution.
  - Chemistry ODEs integrated with an implicit solver.

LMC vs. Fully Explicit Method

- Improvement by a factor of 200-250 in time to solution by using LMC over fully explicit method on a uniform grid at the same effective resolution.
- Deviation from scalability is a miniscule effect relative to the difference between the approaches.



Green: fully explicit method.
Red: LMC.

Seven Challenges to Reach Sustained Exaflop/s for Science (6)

6. We still don't know yet how to express parallelism

(see Kathy Yelick's talk tomorrow)

Seven Challenges to Reach Sustained Exaflop/s for Science (7)

7. System software is an afterthought