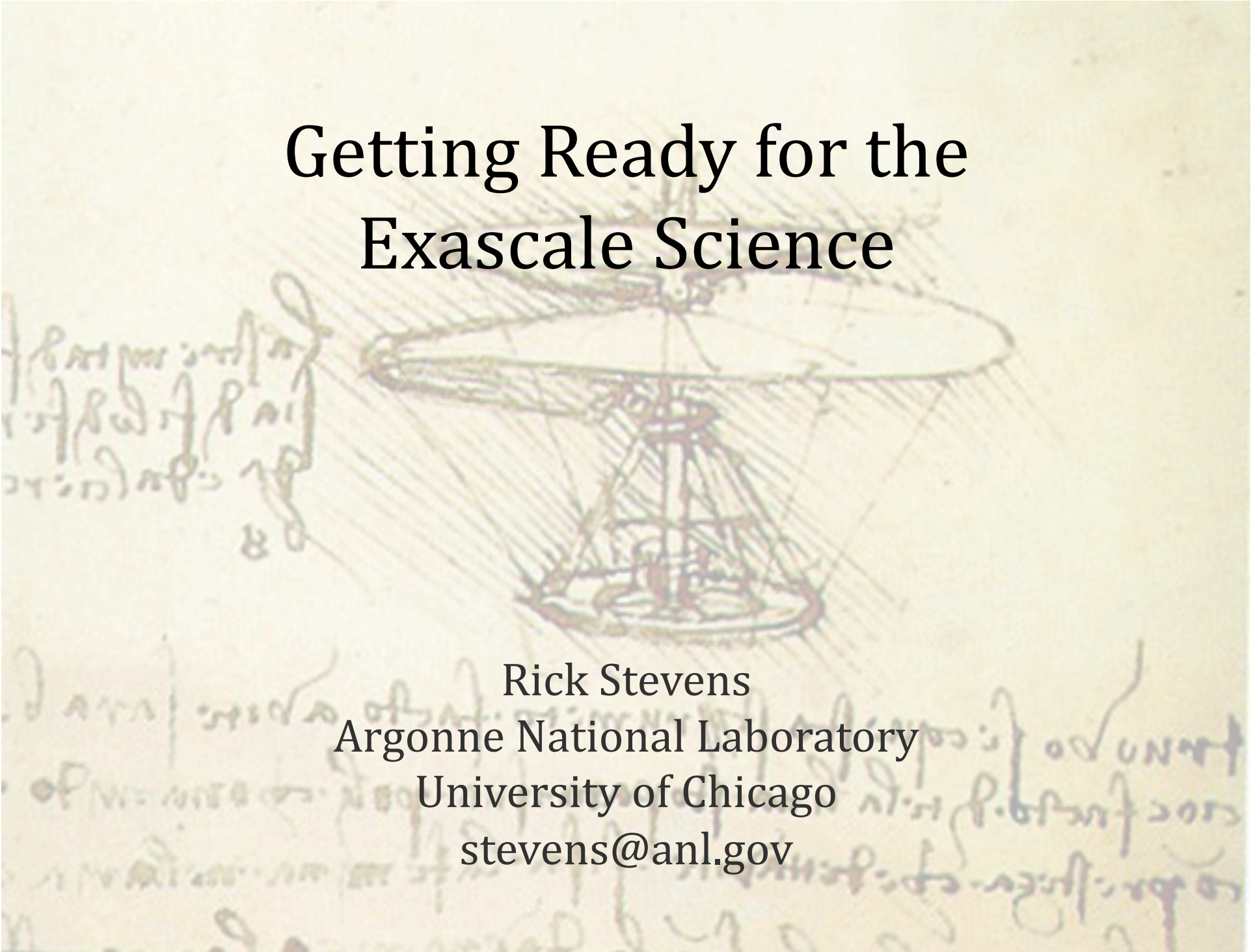


Getting Ready for the Exascale Science



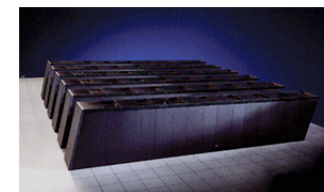
Rick Stevens
Argonne National Laboratory
University of Chicago
stevens@anl.gov

Outline

- What we are doing at ANL
 - BG/P and DOE's Incite Program for allocating resources
- Potential paths to Exascale Systems
 - How feasible are Exascale Systems?
 - What will they look like?
- Issues with heirloom and legacy codes
 - How large is the body of code that is important?
 - What are strategies for addressing migration?
- Driving the development of next generation systems with E3 applications
 - We will need to sustain large-scale investments to make Exascale systems possible, how do we build the case?

ASCR High Performance and Leadership Computing Facilities

- NERSC
 - 104 teraflop Cray XT4 with approximately 9,600 dual core processors; **will upgrade to approximately 360 teraflops with quad core in Summer, 2008**
 - 6.7 teraflop IBM Power 5 (Bassi) with 888 processors, 3.5 terabytes aggregate memory
 - 3.1 teraflop LinuxNetworx Opteron cluster (Jacquard) with 712 processors, 2.1 terabytes aggregate memory
- LCF at Oak Ridge
 - 263 teraflop Cray XT4 (Jaguar) with 7,832 quad core 2.1 GHz AMD Opteron processor nodes, 46 terabytes aggregate memory
 - 18.5 teraflop Cray X1E (Phoenix) with 1,024 multi-streaming vector processors
 - **Delivery of 1 Petaflop Cray Baker expected in late 2008**
- Argonne LCF
 - 5.7 teraflop IBM Blue Gene/L (BGL) with 2,048 PPC processors
 - 100 teraflop IBM Blue Gene/P began operations April 1, 2008
 - **446 teraflop IBM Blue Gene/P upgrade accepted in March, 2008 in transition to operations**



Argonne Leadership Computing Facility

Established 2006. Dedicated to breakthrough science and engineering.

- **Computers**

- BGL: 1024 nodes, 2048 cores, 5.7 TF speed, 512GB memory
- Supports development + INCITE

- **2008 INCITE**

- 111 TF Blue Gene/P system
- Fast PB file system
- Many PB tape archive

- **2009 INCITE production**

- 445 TF Blue Gene/P upgrade
- 8PB next generation file system
- 557TF merged system

- **BG/Q R&D proceeding**

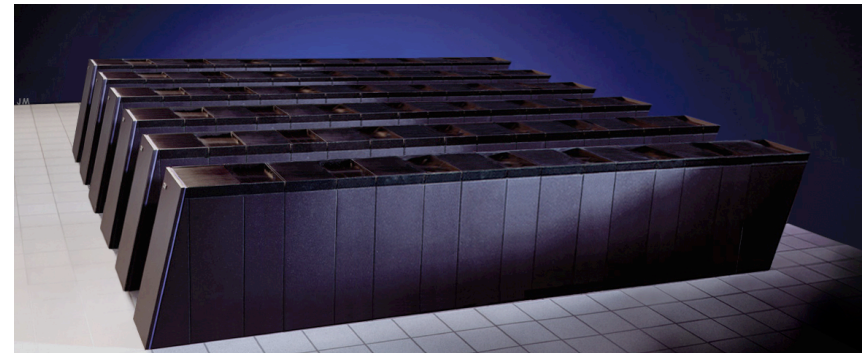
- Frequent design discussions
- Simulations of applications



Blue Gene/L at Argonne

In 2004 DOE selected the ORNL, ANL and PNNL team based on a competitive peer review

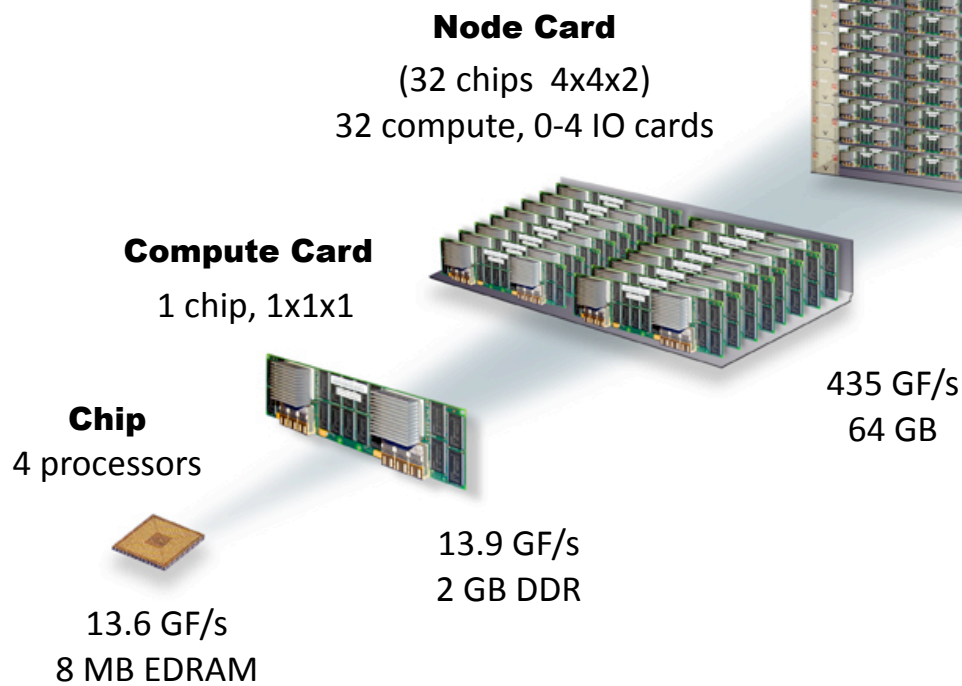
- ORNL to deploy a series of Cray X-series systems
- ANL to deploy a series of IBM Blue Gene systems
- PNNL to contribute software technology



Blue Gene/P Engineering Rendition

Blue Gene/P is an Evolution of BG/L

- Processors + memory + network interfaces are all on the same chip.
- Faster Quad core processors with larger memory
- 5 flavors of network, with faster signaling, lower latency



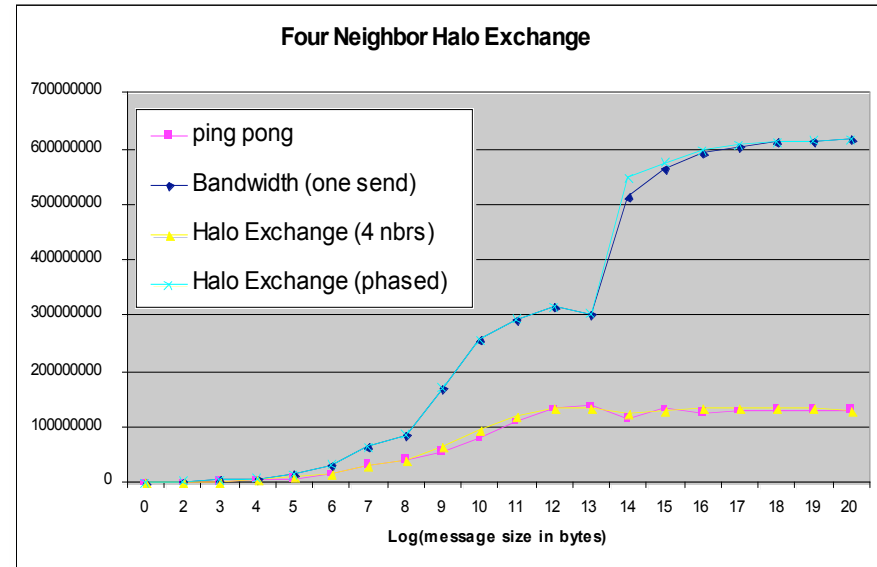
- High packaging density
- High reliability
- Low system power requirements
- XL compilers, ESSL, GPFS, LoadLeveler, HPC Toolkit
- MPI, MPI2, OpenMP, Global Arrays

IBM Confidential

Blue Gene community knowledge base is preserved

Some Good Features of Blue Gene

- Multiple links may be used concurrently
 - Bandwidth nearly 5x simple “pingpong” measurements
- Special network for collective operations such as Allreduce
 - Vital (as we will see) for scaling to large numbers of processors
- Low “dimensionless” message latency
- Low relative latency to memory
 - Good for unstructured calculations
- BG/P improves
 - Communication/Computation overlap (DMA on torus)
 - MPI-I/O performance



Smaller is Better

	s/f	r/f	s/r	Reduce	Reduce for 1PF
BG/P	2110	9	233	12us	12us
BG/P (one link)	2110	42	50	12us	12us
XT3	7920	10	760	2slog p	176us
Generic Cluster	13500	34	397	2slog p	316us
Power5 SP	3200	6	529	2slog p	41us

Communication Needs of the “Seven Dwarves”

These seven algorithms taken from “Defining Software Requirements for Scientific Computing”, Phillip Colella, 2004

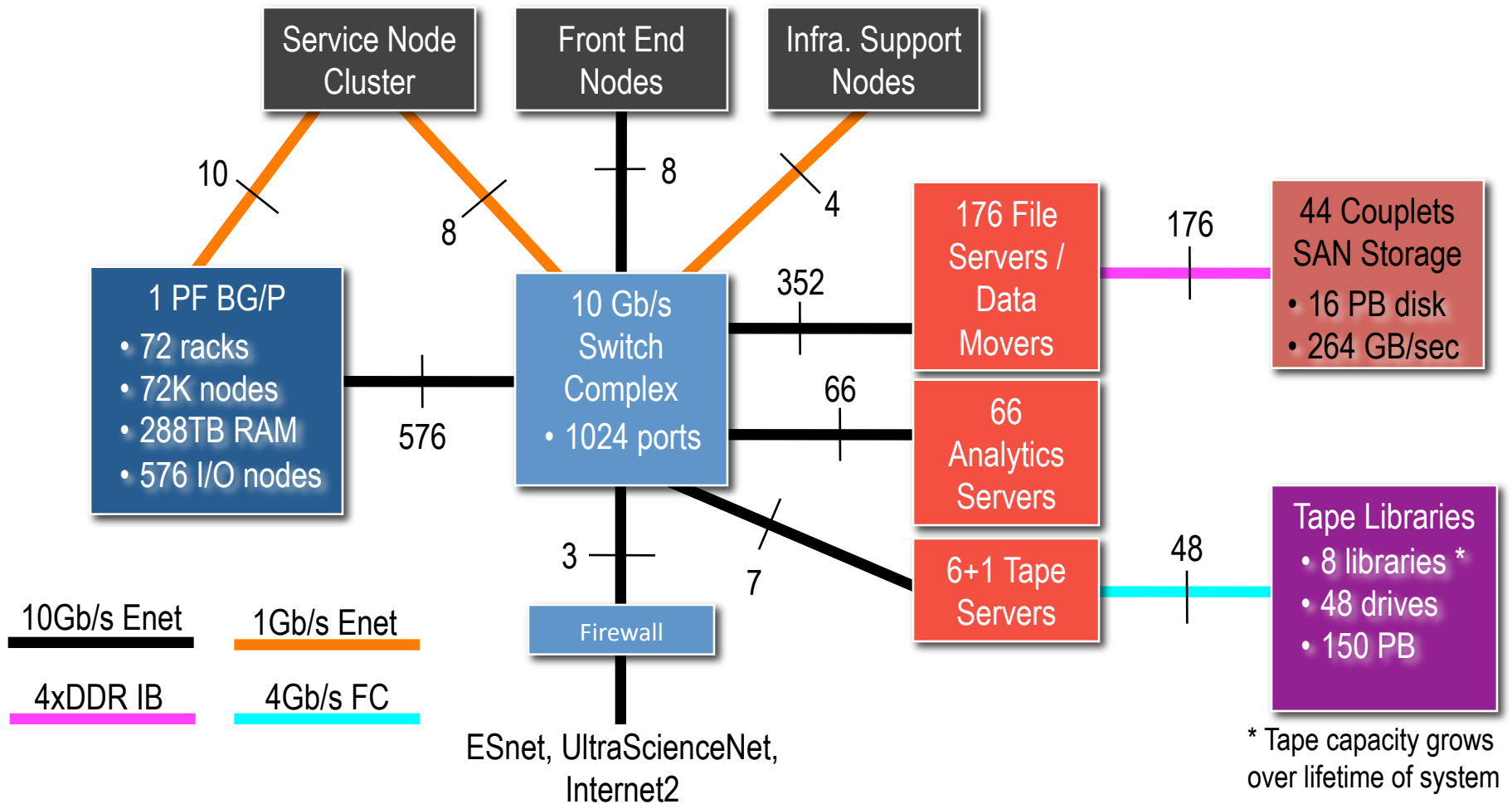
1. Molecular dynamics (mat)
2. Electronic structure
3. Reactor analysis/CFD
4. Fuel design (mat)
5. Reprocessing (chm)
6. Repository optimizations
7. Molecular dynamics (bio)
8. Genome analysis
9. QMC
10. QCD
11. Astrophysics

	Tree/Combine		Torus
Algorithm	Scatter/Gather	Reduce/Scan	Send/Recv
Structured Grids 3, 5, 6, 11	Optional	X _{LB}	X
Unstructured Grids 3, 4, 5, 6, 11		X _{LB}	X
FFT 1, 2, 3, 4, 7, 9	Optional		X
Dense Linear Algebra 2, 3, 5	Not Limiting	Not Limiting	X
Sparse Linear Algebra 2, 3, 5, 6, 8, 11		X	X
Particles N-Body 1, 7, 11	Optional	X	X
Monte Carlo 4, 9		*	X

**Blue Gene
Advantage**

Legend: Optional – Algorithm can exploit to achieve better scalability and performance. Not Limiting – algorithm performance insensitive to performance of this kind of communication. X – algorithm performance is sensitive to this kind of communication. X_{LB} – For grid algorithms, operations may be used for load balancing and convergence testing

Argonne Petascale System Architecture



In the BG/P generation like BG/L the I/O Architecture is not tightly coupled to the compute fabric!

Theory and Computational Sciences Building



TCS Conceptual Design

- **A superb work and collaboration environment for computer and computational sciences**
 - 3rd party design/build project
 - 2009 beneficial occupancy
 - 200,000 sq.ft., 600+ staff
 - Open conference center
 - Research Labs
 - Argonne's library
- **Supercomputer Support Facility**
 - Designed to support leadership systems (shape, power, weight, cooling, access, upgrades, etc.)
 - 20,000 sq.ft. initial space
 - Expandable to 40,000+ sq.ft.

Argonne Theory and Computing Sciences Building



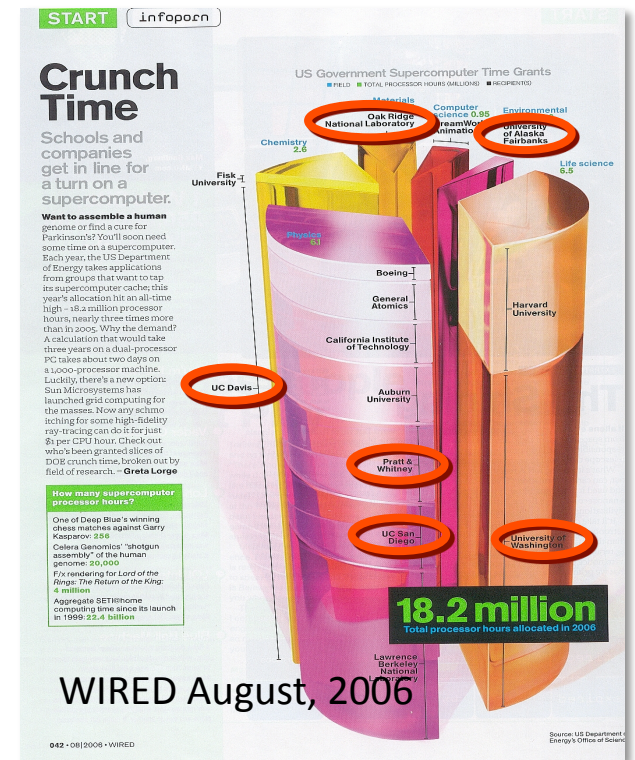
A 200,000 sq ft creative space to do science, Coming Summer 2009

DOE INCITE Program

Innovative and Novel Computational Impact on Theory and Experiment

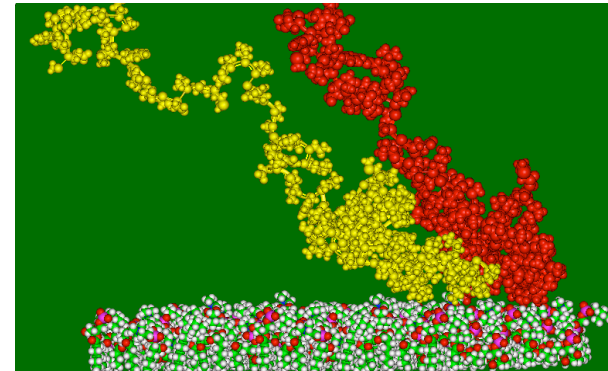
Since
2004

- Solicits large computationally intensive research projects
 - To enable high-impact scientific advances
- Open to all scientific researchers and organizations
 - Scientific Discipline Peer Review
 - Computational Readiness Review
- Provides large computer time & data storage allocations
 - To a small number of projects for 1-3 years
 - Academic, Federal Lab and Industry, with DOE or other support
- Primary vehicle for selecting Leadership Science Projects for the Leadership Computing Facilities
- Call current open Due August 11, 2008

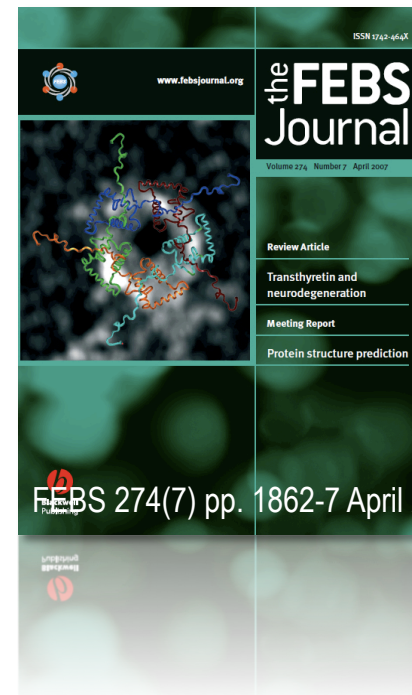


Modeling of Protofibril Structures Provides Insight into Molecular Basis of Parkinson's Disease

- PI: Igor Tsigelny, UCSD
- Parkinson's Disease is the 2nd most common adult neurological disease
- Increased aggregation of *alpha-synuclein* protein is thought to lead to harmful pore-like structures in human membranes
- UCSD - SDSC team used molecular modeling and molecular dynamics simulations in combination with biochemical and ultrastructural analysis to show that *alpha-synuclein* can lead to the formation of pore-like structures on membranes
- Used NAMD and MAPAS on Blue Gene/L at ALCF and SDSC

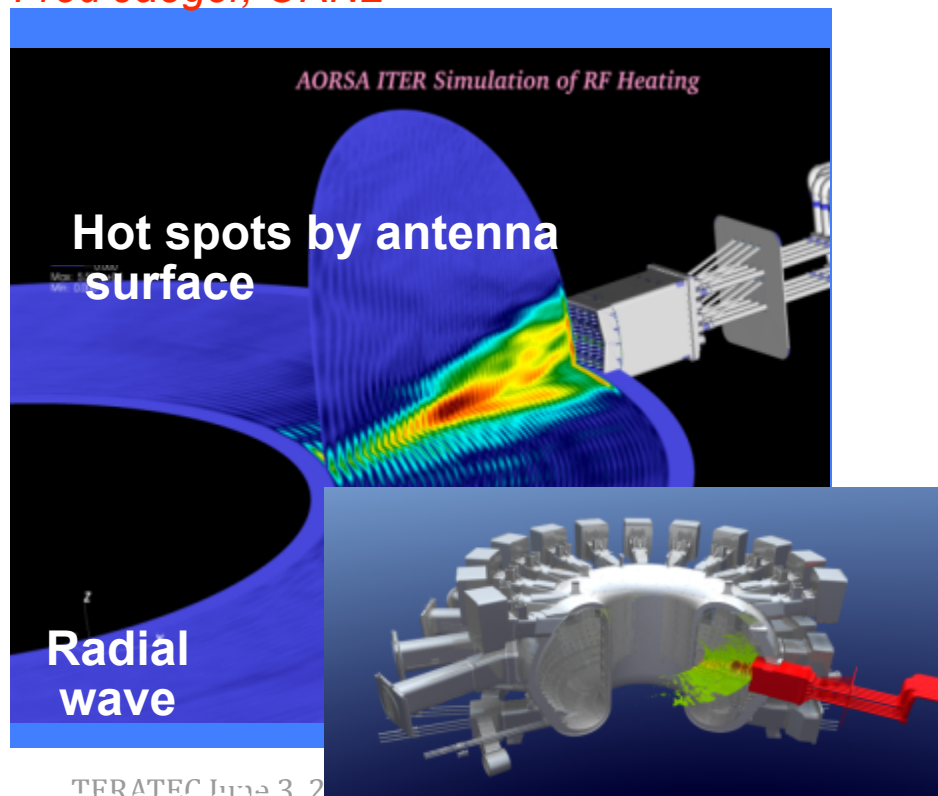


Above - formation of *alpha-synuclein* dimer on a membrane, aggregating toward the pentamer pore structure - below.



Producing New Insights for RF Heating of ITER Plasmas

“Until recently, we were limited to two-dimensional simulations. The larger computer [Jaguar] has allowed us to achieve three-dimensional images and validate the code with observations.” – Fred Jaeger, ORNL



- 3D simulations reveal new insights
 - “Hot spots” near antenna surface
 - “Parasitic” draining of heat to the plasma surface in smaller reactors
- Work pushing the boundaries of the system (22,500 processor cores, 87.5 TF) and demonstrating
 - Radial wave propagation and rapid absorption
 - Efficient plasma heating
- AORSA’s predictive capability can be coupled with Jaguar power to enhance fusion reactor design and operation for an unlimited clean energy source

Fully 3-dimensional simulations of plasma shed new light on the behavior of superheated ionic gas in the multibillion-dollar ITER fusion reactor



Office of Science

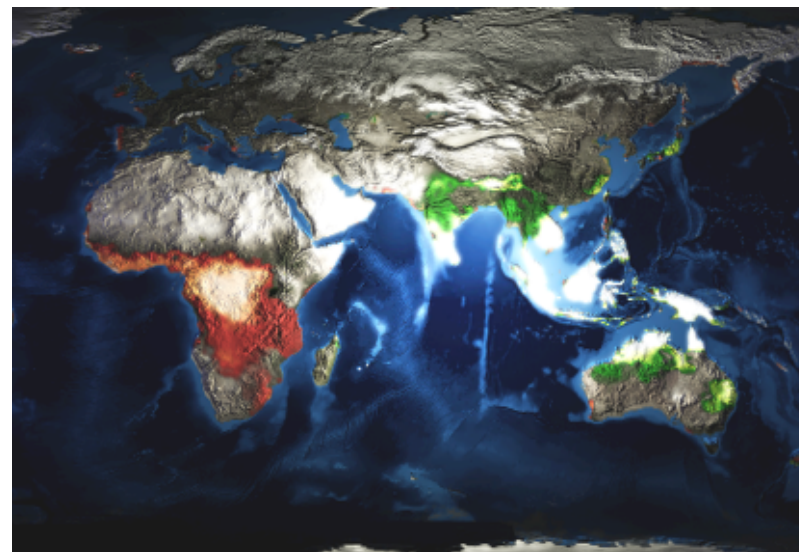
Accelerating Climate Science

- **PI– Warren Washington, NCAR**
- **First-ever control runs of CCSM 3.5 at groundbreaking speed**

“[On Jaguar,] we got 100-year runs in three days. This was a significant upgrade of how we do science with this model. 40 years per day was out of our dreams.”

Peter Gent of NCAR, Chairman of CCSM Scientific Steering Committee, during keynote at CCSM Workshop, June 19, 2007

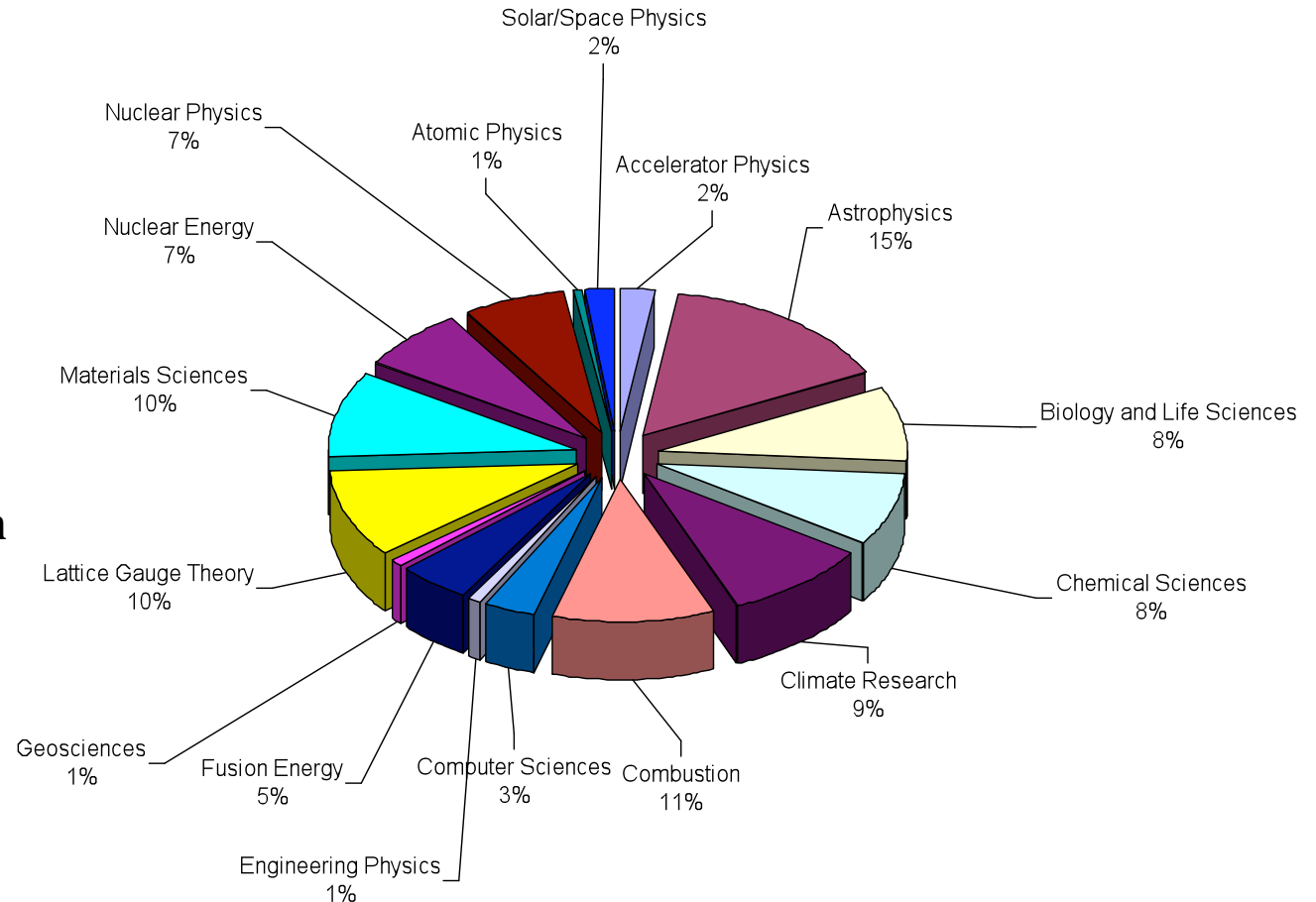
- **Major improvements in CCSM 3.5**
 - Arctic and Antarctic sea ice: Will the Arctic be ice free in summer of 2050?
 - Surface hydrology of land, critical for predictions of drought
- **Positioned to test full carbon-nitrogen cycle**



Instantaneous net ecosystem exchange (NEE): eastern half is in sunlight and the terrestrial ecosystems are taking up carbon (negative NEE, shown in green to bright white). Meanwhile, the sun has not yet risen in the western half of the image where the ecosystems are only respiring (positive NEE, shown in red)

INCITE 2008

- Received 88 new and 24 renewal proposals requesting over 600 Million processor hours
 - 44% from Universities
 - 46% funded from non-DOE sources
- Over 265 Million hours awarded to 55 new and renewal projects

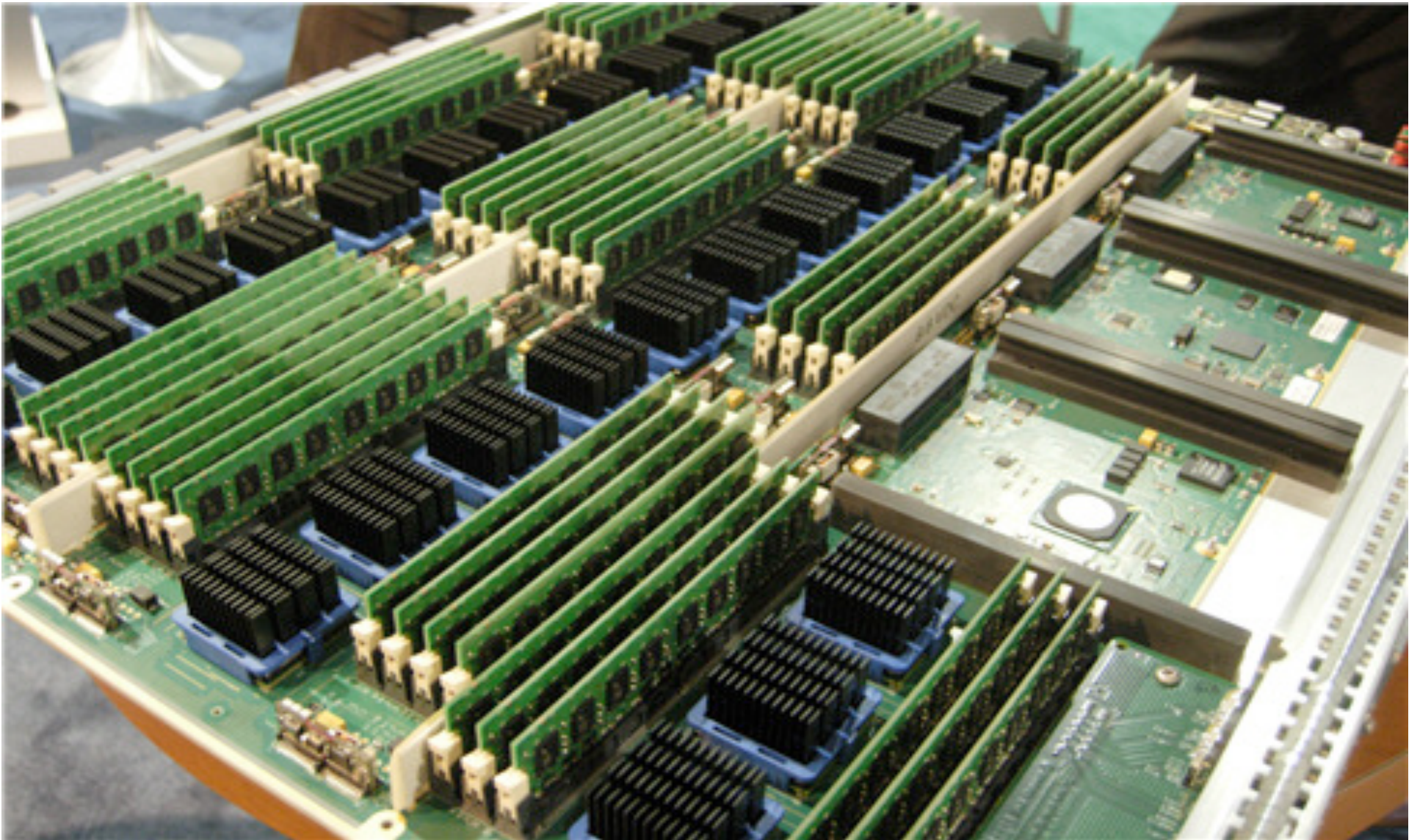


Thierry Poinsot, CERFACS awarded 4M hours on Argonne National Laboratory's IBM Blue Gene/P

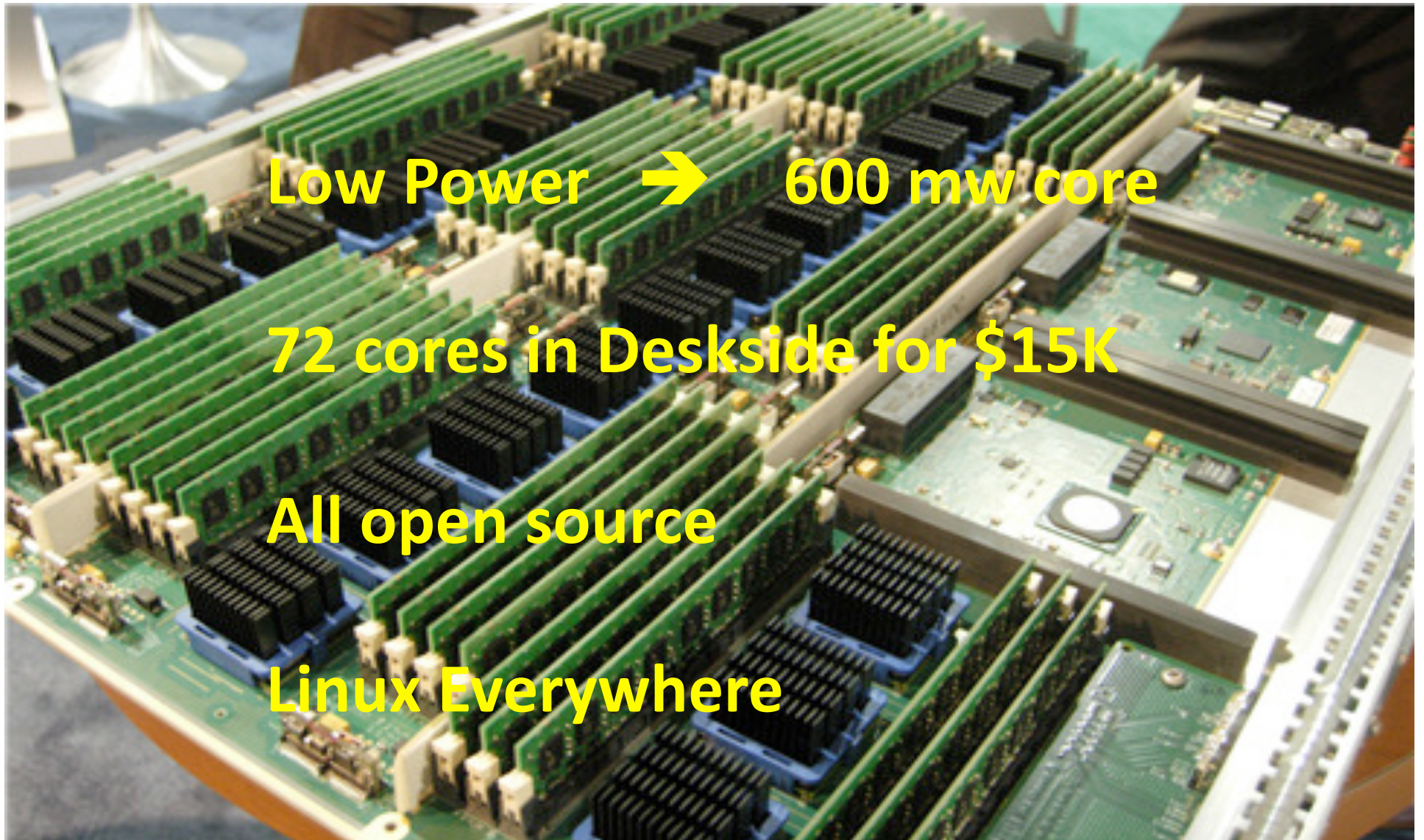
Supercomputing & Cloud Computing

- Two macro architectures dominate large-scale (intentional) computing infrastructures
- Supercomputing type Structures
 - Large-scale integrated coherent systems
 - Managed for high utilization and efficiency
- Emerging cloud type Structures
 - Large-scale loosely coupled, lightly integrated
 - Managed for availability, throughput, reliability

SiCortex Node Board



SiCortex Node Board



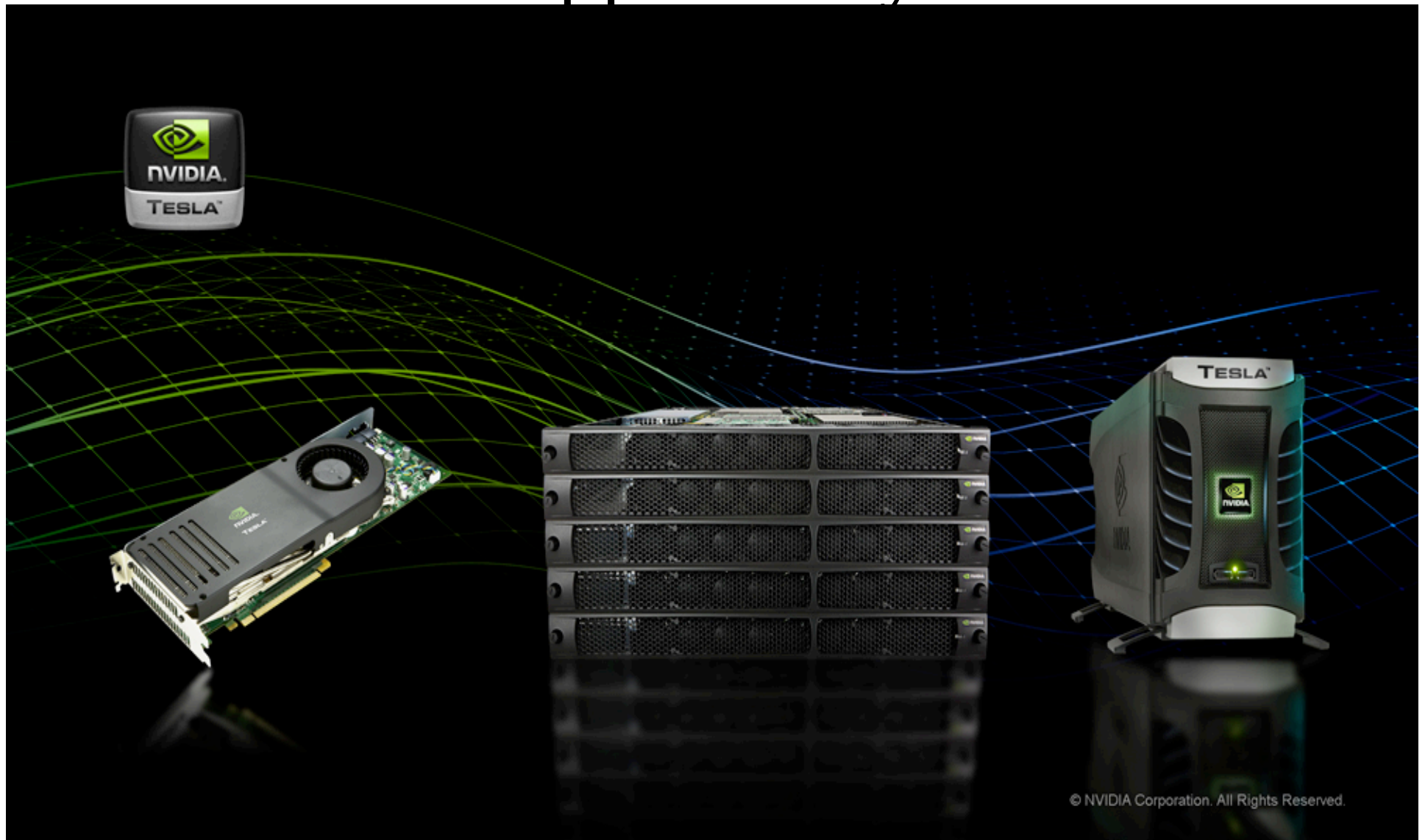
Low Power → 600 mw core

72 cores in Deskside for \$15K

All open source

Linux Everywhere

The NVIDIA Challenge and Opportunity



The NVIDIA Challenge and Opportunity



Potentially Easy Access to Teraflops

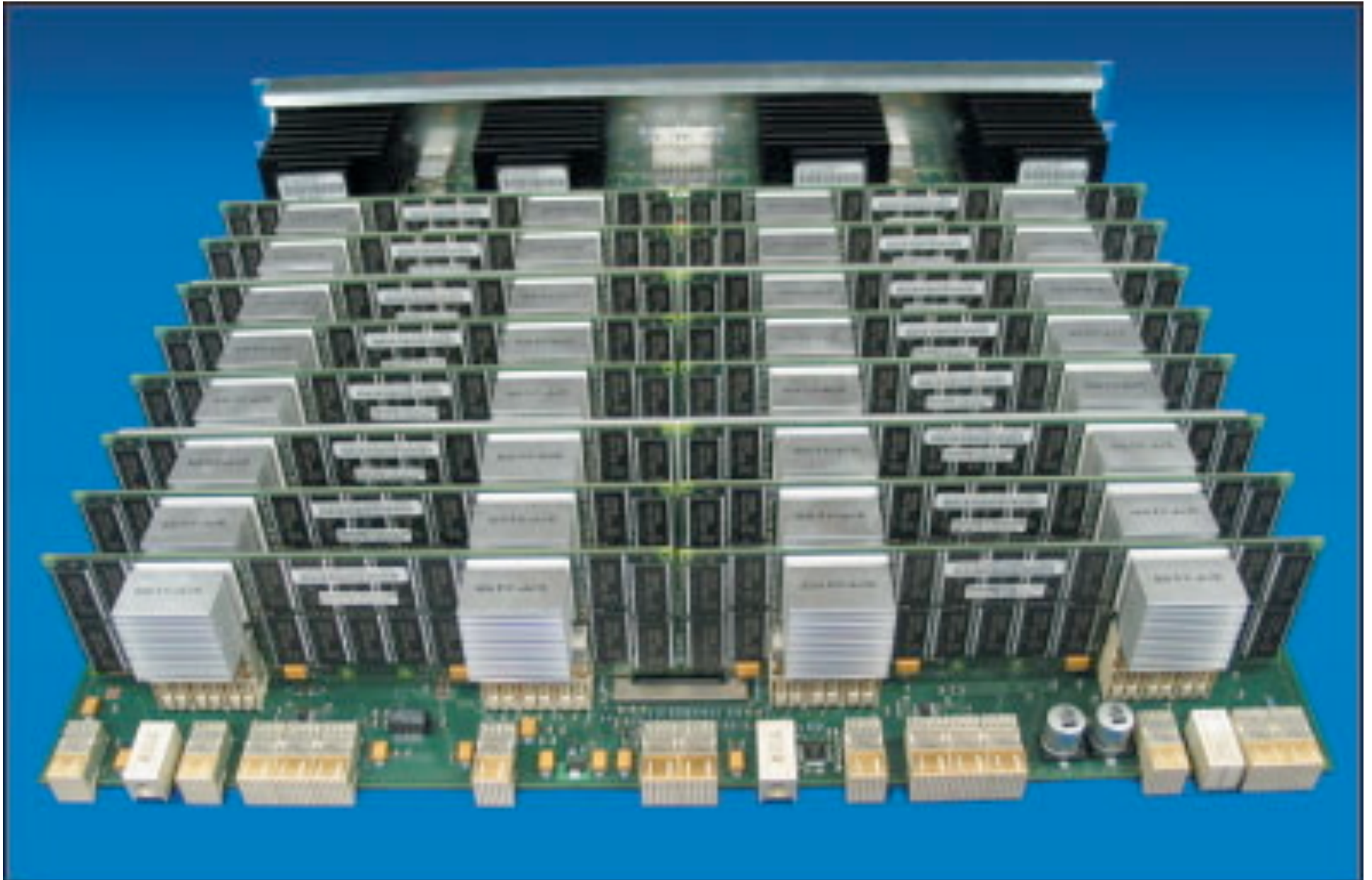
Simple Programming Model

Requires Large Thread Counts

Proprietary Software Environment



Blue Gene L Node Cards

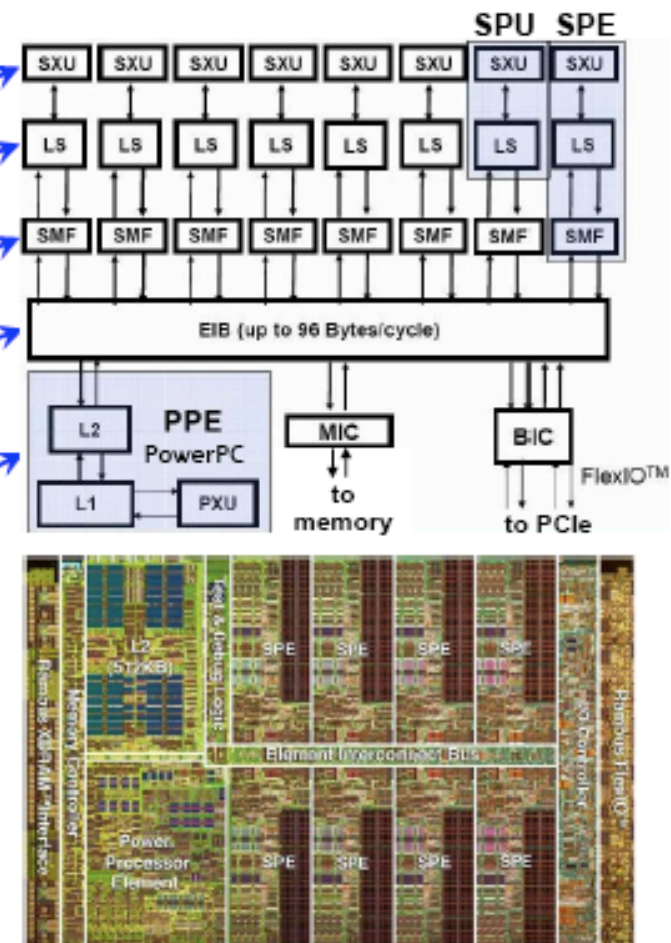


Blue Gene Node Cards



The Cell processor is an (8+1)-way heterogeneous parallel processor

- Cell Broadband Engine (CBE*) developed by Sony-Toshiba-IBM
 - used in Sony PlayStation 3
- 8 Synergistic Processing Elements (SPEs)
 - 128-bit vector engines
 - 256 kB local memory (LS = Local Store)
 - Direct Memory Access (DMA) engine (25.6 GB/s)
 - Chip interconnect (EIB)
 - Run SPE-code as POSIX threads (SPMD, MPMD, streaming)
- PowerPC PPE runs Linux OS
- Current performance:
 - 204.8 GF/s SP & 13.65 GF/s DP
 - 512 MB @ 25.6 GB/s XDR memory



* trademark of Sony Computer Entertainment, Inc.

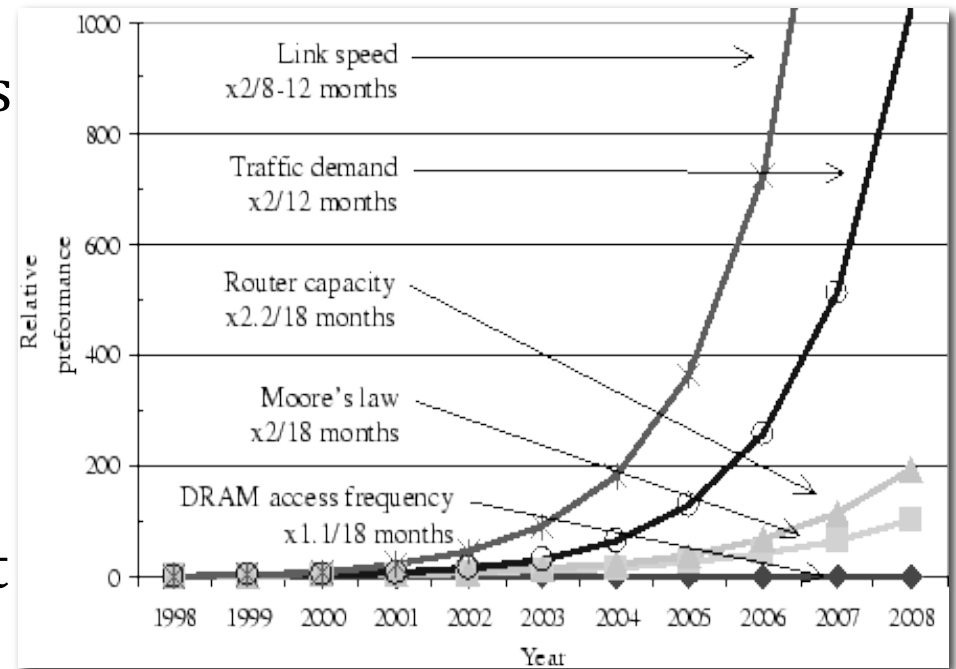
Roadrunner at a glance

- **Cluster of 18 Connected Units (CU)**
 - 6,912 AMD dual-core Opterons
 - 12,960 IBM Cell eDP accelerators
 - 49.8 Teraflops peak (Opteron)
 - 1.33 Petaflops peak (Cell eDP)
 - 1PF sustained Linpack
- **InfiniBand 4x DDR fabric**
 - 2-stage fat-tree; all-optical cables
 - Full bi-section BW within each CU
 - 384 GB/s (bi-directional)
 - Half bi-section BW among CUs
 - 3.45 TB/s (bi-directional)
 - Non-disruptive expansion to 24 CUs
- **80 TB aggregate memory**
 - 28 TB Opteron
 - 52 TB Cell
- **216 GB/s sustained File System I/O:**
 - 216x2 10G Ethernets to Panasas
- **RHEL & Fedora Linux**
- **SDK for Multicore Acceleration**
 - Cell compilers, libraries, tools
- **xCAT Cluster Management**
 - System-wide GigE network
- **3.9 MW Power:**
 - 0.35 GF/Watt
- **Area:**
 - 296 racks
 - 5500 ft²



Thinking about Trends

- Slow forces that build up over time and change things slowly but surely
 - Moore's Law
 - Global Warming
 - Wireless
 - Digital Imaging
- Unanticipated Rapid Impact
 - Peer-to-peer
 - Social Network Applications

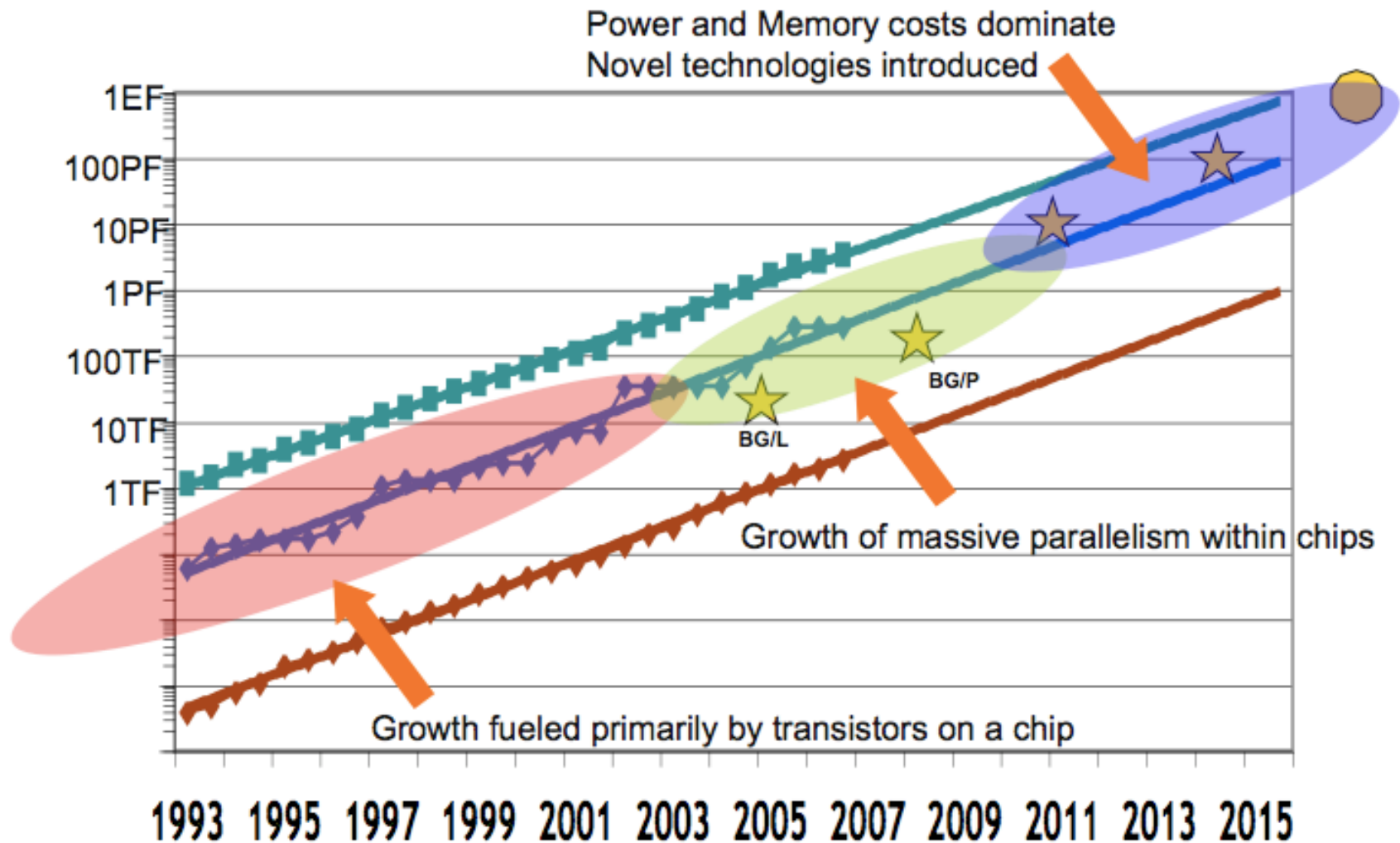


Thinking about Trends

- Slow forces that build up over time and change things slowly but surely
 - Moore's Law
 - Global Warming
 - Wireless
 - Digital Imaging
- Unanticipated Rapid Impact
 - Peer-to-peer
 - Social Network Applications



Looking to Exascale



A Three Step Path to Exascale

Begin Full System Delivery (Yr)	2004	2007	2012	2015	2019
Design Parameters	BG/L	BG/P	ONE	TWO	THREE
Cores / Node	2	4	8-24	32-64	96-128
Clock Speed (GHz)	0.7	0.85	1.6-4.1	2.3-4.8	2.8-6.0
Flops / Clock / Core	4	4	8-32	8-32	16-64
Nodes / Rack	1024	1024	100-512	256-1024	256-1024
Racks / Full System Config	64	72	128-350	128-400	256-400
MB RAM/core	256	512	1024-4096	1024-4096	1024-4096
Total Power	2.5MW	4.8MW	8MW-20MW	20MW-50MW	40MW-80MW
Flops / Node (GF)	5.6	14	128-640	640-2000	2000-6000
Flops / Rack (TF)	5.7	14	200-400	400-1200	1600-4800
LB Concurrency	5.E+05	1.E+06	1M-2M	10M-100M	400M-100M
Full System					
Total Cores (Millions)	0.13	0.3	.3M-1.2M	1M-10M	4M-30M
Total RAM (TB)	33.6	151	2,000-4,400	3,000-10,000	5,000-25,000
Total Racks	64	72	128-350	128-400	256-400
Peak Flops System (PF)	0.37	1	25	300	1200

E3 Advanced Architectures - Findings

- Exascale systems are likely feasible by 2017±2
- 10-100 Million processing elements (mini-cores) with chips as dense as 1,000 cores per socket, clock rates will grow slowly
- 3D chip packaging likely
- Large-scale optics based interconnects
- 10-100 PB of aggregate memory
- > 10,000's of I/O channels to 10-100 Exabytes of secondary storage, disk bandwidth to storage ratios not optimal for HPC use
- Hardware and software based fault management
- Simulation and multiple point designs will be required to advance our understanding of the design space
- Achievable performance per watt will likely be the primary metric of progress

E3 Advanced Architectures - Challenges

- Performance per watt -- goal 100 GF/watt of sustained performance \Rightarrow 10 MW Exascale system
 - Leakage current dominates power consumption
 - Active power switching will help manage standby power
- Large-scale integration -- need to package 10M-100M cores, memory and interconnect < 10,000 sq ft
 - 3D packaging likely, goal of small part classes/counts
- Heterogenous or Homogenous cores?
 - Mini cores or leverage from mass market systems
- Reliability -- needs to increase by 10^3 in faults per PF to achieve MTBF of 1 week
 - Integrated HW/SW management of faults
- Integrated programming models (PGAS?)
 - Provide a usable programming model for hosting existing and future codes

Top Pinch Points

- Power Consumption
 - Proc/mem, I/O, optical, memory, delivery
- Chip-to-Chip Interface Scaling (pin/wire count)
- Package-to-Package Interfaces (optics)
- Fault Tolerance (FIT rates and Fault Management)
 - Reliability of irregular logic, design practice
- Cost Pressure in Optics and Memory

Programming Models: Twenty Years and Counting

- In large-scale scientific computing today essentially all codes are message passing based (CSP and SPMD)
- Multicore is challenging the sequential part of CSP but there has not emerged a dominate model to augment message passing
- Need to identify new programming models that will be stable over long term

Quasi Mainstream Programming Models

- C, Fortran, C++ and MPI, CHARM++
- OpenMP, pthreads
- CUDA, RapidMind
- Clearspeeds Cn
- PGAS (UPC, CAF, Titanium)
- HPCS Languages (Chapel, Fortress, X10)
- HPC Research Languages and Runtime
- HLL (Parallel Matlab, Grid Mathematica, etc.)

Little's Law of High Performance Computing

Assume:

- Single processor-memory system.
- Computation deals with data in local main memory.
- Pipeline between main memory and processor is fully utilized.

Then by Little's Law, the number of words in transit between CPU and memory (i.e. length of vector pipe, size of cache lines, etc.)
= memory latency x bandwidth.

This observation generalizes to multiprocessor systems:

concurrency = latency x bandwidth,

where “concurrency” is aggregate system concurrency, and
“bandwidth” is aggregate system memory bandwidth.

This form of Little's Law was first noted by Burton Smith of Tera.

This slide stolen from David Bailey

Million Way Concurrency Today

- Little's law driven need for concurrency
 - To cover latency in memory path
 - Function of aggregate memory bandwidth and clock speed
 - Independent of technology and architecture to first order
- Mainstream CPUs (e.g. x86, PPC, SPARC)
 - 8-16 cores, 4-8 hardware threads per core,
 - Total system with $10^3 - 10^5$ nodes => 32K – 12M threads
 - BG/P example at 1 PF $72 \times 4K = 300,000$ (but each thread has to do 4 ops/clock) => 1.2M ops per clock
- GPU based cluster (e.g. 1000 Tesla 1 U nodes)
 - 3 x 128 cores x (32-96) threads per core x 1000 nodes = 12M – 36M threads

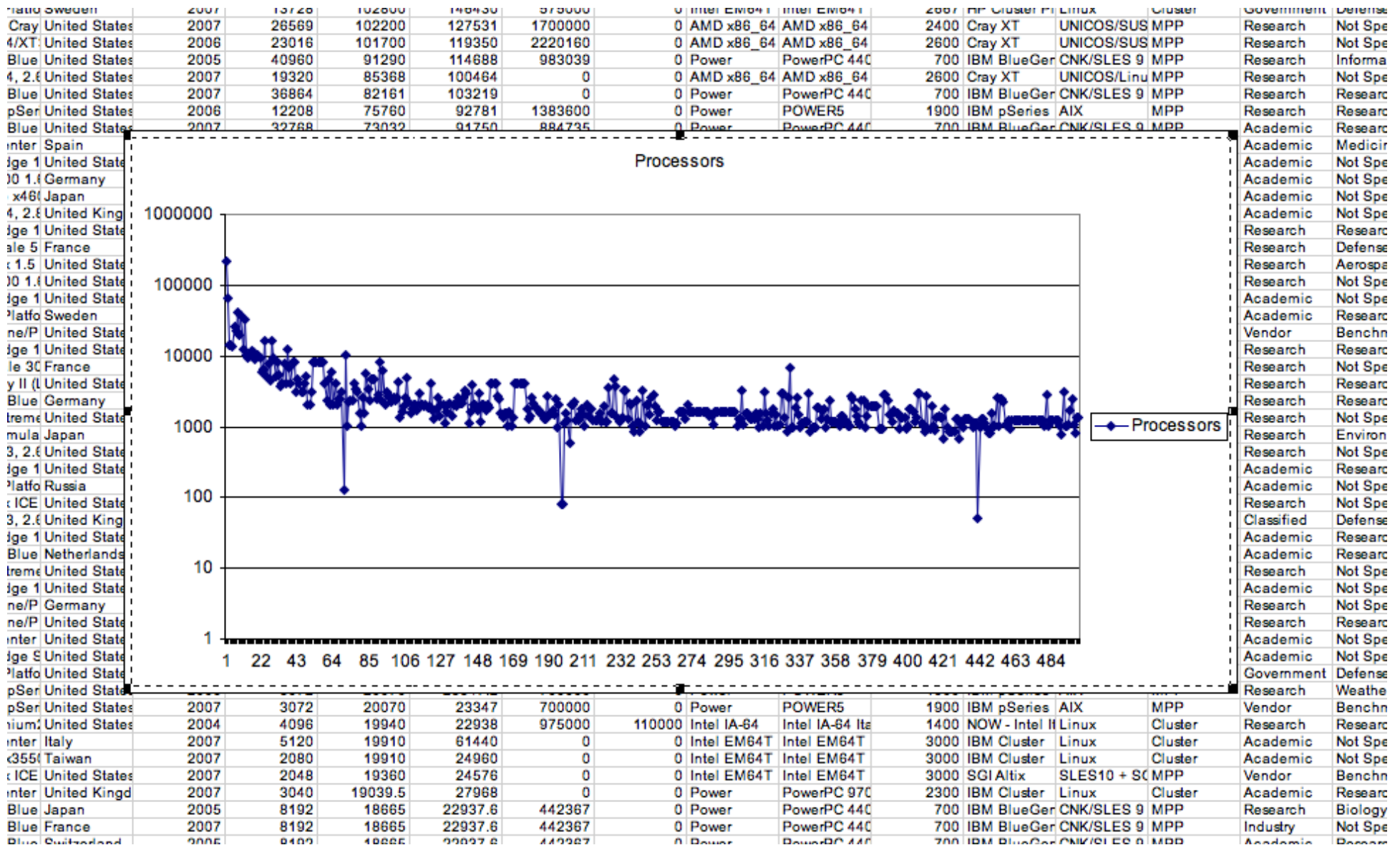
Lessons Learned from Terascale to Petascale

- The early adopters almost always self identify
- Approximately 1/3 of the petascale codes didn't exist 10 years ago
- Most of them did exist but required considerable investment, new implementation and tuning
- The simplest path forward (pure MPI) was the path of least resistance for most code groups
- The challenges moving forward are likely to be slightly different

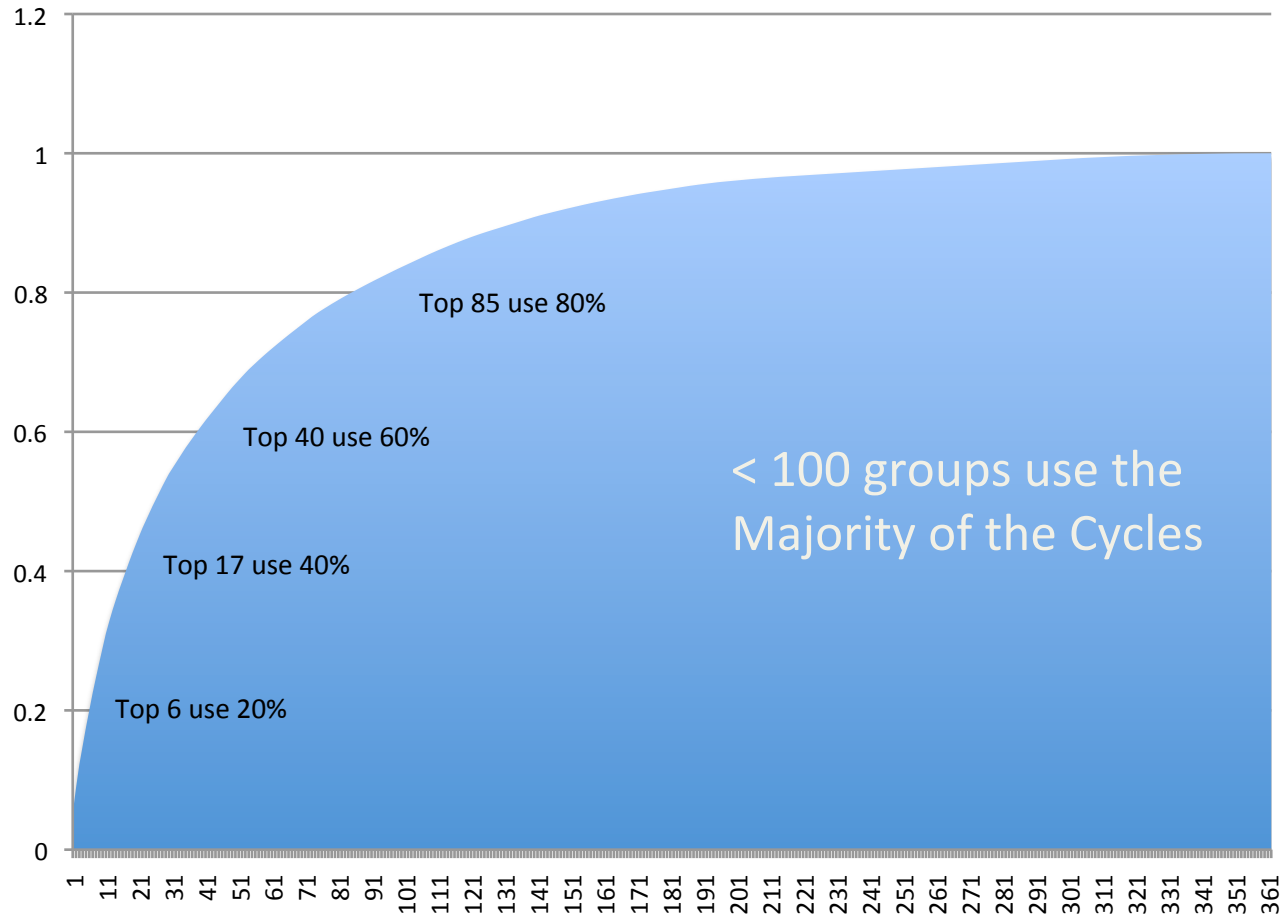
Existing Body of Parallel Software

- How many existing HPC science and engineering codes scale beyond 1000 processors?
 - My estimate is that it is less than 1000 world wide
 - Top users at NERSC, OLCF and ALCF < 200 groups
 - It appears likely that the bulk of cycles on Top500 are used in capacity mode with the exception of a sites with policies that enforce capability runs
- How quickly are new codes being generated?
 - Ab initio development
 - Migration and porting from previous generations
- There are different choices faced by large-established projects and personal explorations of new technologies

Number of Processors In the Top500



NERSC 2007 Rank Abundance

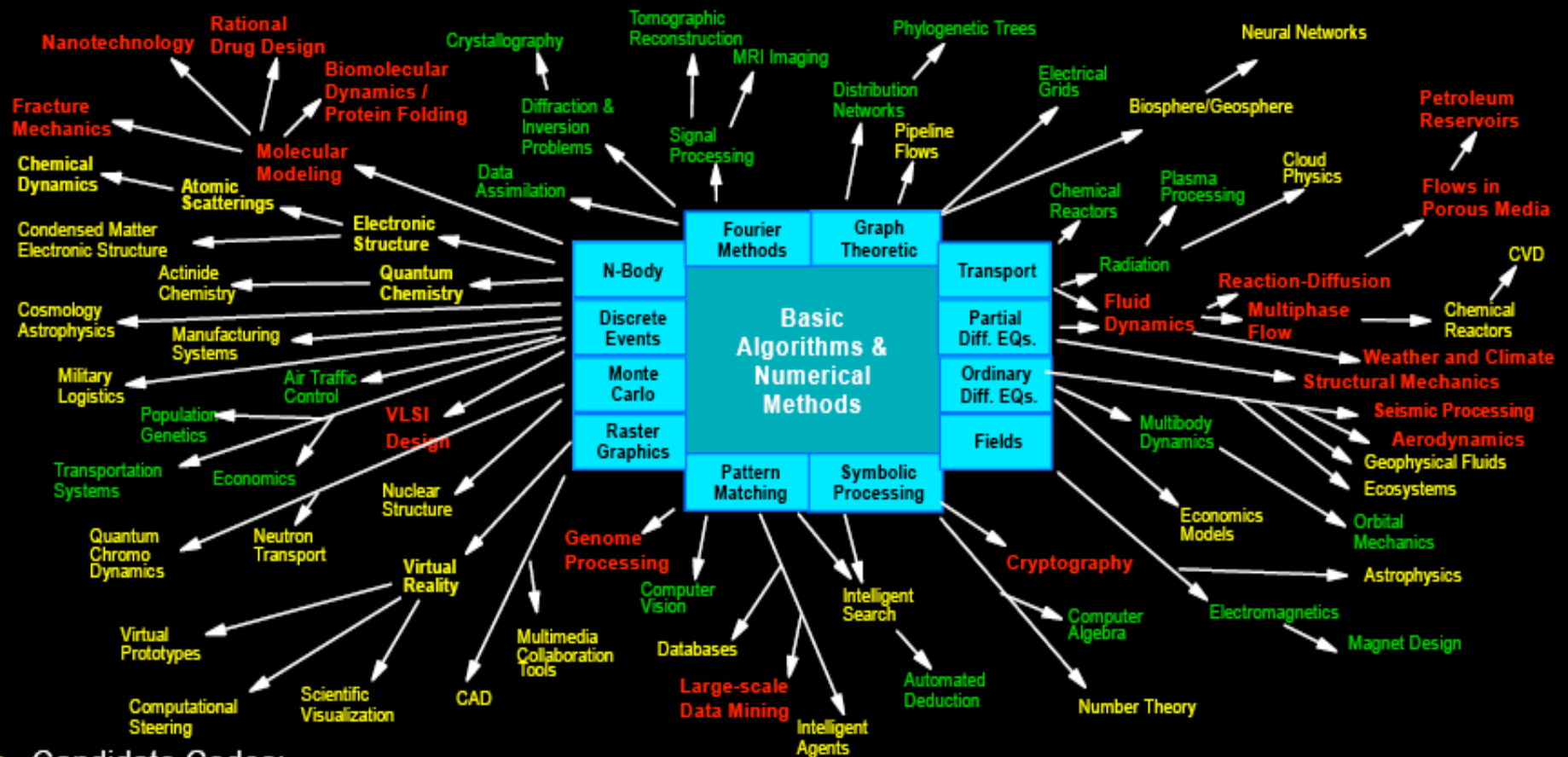


Existing Applications of Interest

- Climate and Weather (e.g. CCM3, POP, WRF)
- Plasma Physics (e.g. GTC, GYRO, M3D)
- Combustion (e.g. S3D, NCC)
- Multi-physics CFD (e.g. NEK, SHARP)
- Lattice QCD (e.g. MILC, CPS)
- Cosmology and Relativity (e.g. ENZO, Cactus)
- Astrophysics (e.g. FLASH, CHIMERA)
- Molecular Dynamics (e.g. NAMD, AMBER)
- Electronic Structure (e.g. QBOX, LSMS, QMC)
- Evolution (e.g. mrBayes, Clustalw-MPI)

Good Better Best

Many Classes of Applications are Massively Parallel



- Candidate Codes:
 - Inherently parallel; written using MPI
 - Memory required per MPI task is less than that available
 - Dominated by collective communication across all nodes
 - Locality of communications within 3D mapping
- Non-Candidate Codes:
 - Large memory footprints required on individual nodes
 - Client/server structures
 - Dominated by disk I/O

How Quickly Can A New Architecture Be Adopted?

Applied Mathematics and Computer Science are Essential to Advancing Science

- Programming models are needed for million way concurrency and beyond
- New classes of algorithms are needed that have better scaling properties
- Systems software is needed to make systems stable and usable
- New concepts are needed that enable whole new communities to access leadership class computing

Blue Gene Consortium

Ames National Laboratory/Iowa State U.
 Argonne National Laboratory
 Brookhaven National Laboratory
 Fermi National Laboratory
 Jefferson Laboratory
 Lawrence Berkeley National Laboratory
 Lawrence Livermore National Laboratory
 Oak Ridge National Laboratory
 Pacific Northwest National Laboratory
 Princeton Plasma Physics Laboratory

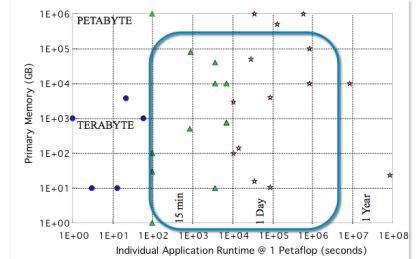
Boston University
 California Institute of Technology
 Columbia University
 Cornell
 DePaul
 Harvard University
 Illinois Institute of Technology
 Indiana University
 Iowa State
 Louisiana State University
 Massachusetts Institute of Technology
 National Center for Atmospheric Research
 New York University/Courant Institute
 Northern Illinois University
 Northwestern University
 Ohio State University
 Pennsylvania State University
 Pittsburgh Supercomputing Center
 Princeton University
 Purdue
 Rutgers
 Stony Brook University
 Texas A&M University
 University of California - Irvine

University of California - San Francisco
 University of CA - San Diego/SDSC
 University of Chicago
 University of Colorado
 University of Delaware
 University of Hawaii
 University of Illinois Urbana Champaign
 University of Minnesota
 University of North Carolina
 University of Southern California/ISI
 University of Texas at Austin/TACC
 University of Utah
 University of Wisconsin

Engineered Intelligence Corporation
 IBM
 Gene Network Science
 Allied Engineering (Japan)

Center of Excellence for Applied Research (CERT)
 Ecole Polytechnique Federale de Lausanne (EPFL)
 Trinity College, Ireland
 National University of Ireland
 Astron
 AIST, Japan
 John von Neumann Institute, Germany
 NIWS Co., Ltd, Japan
 University of Edinburgh, EPCC Scotland
 Institut de Physique du Globe de Paris
 University of Tokyo

Petaflops Applications Coverage

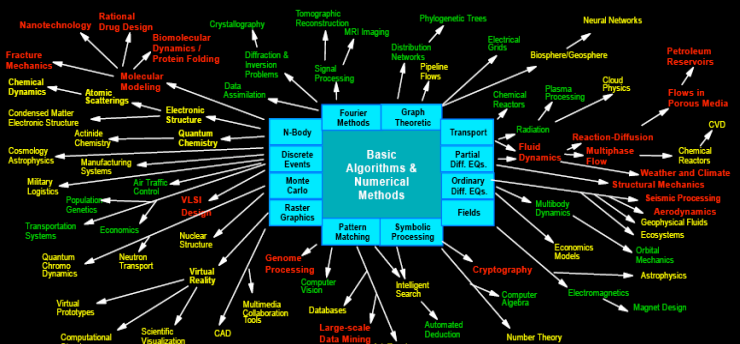


Example Applications Ported to BG/L and BG/P

- How fast can a community adopt a new machine architecture ?

General Domain	Code	Institution	General Domain	Code	Institution
Astro Physics	Enzo	UCSD/SDSC	Material Sciences	ALE3D	LLNL
Astro Physics	Flash	UC/Argonne	Material Sciences	LSMS	LLNL
Basic Physics	CPS	Columbia	Molecular Biology	mpiBLAST	Argonne
Basic Physics	QCD kernel	IBM	Molecular dynamics	MDCASK	LLNL
Basic Physics	QCD	Argonne	Molecular Dynamics	Amber	UCSF
Basic Physics	QMC	CalTech	Molecular dynamics	APBS	UCSD
Basic Physics	QMC	Argonne	Molecular dynamics	Blue Matter	IBM
BioChemistry	BGC.5.0	NCAR	Molecular Dynamics	Charmm	Harvard
BioChemistry	BOB	NCAR	Molecular dynamics	LJMD	CalTech
CAE/FEM Structure	PAM-CRASH	ESI	Molecular Dynamics	NAMD	UIUC/NCSA
CFD	Miranda	LLNL	Molecular Dynamics	Qbox	LLNL
CFD	Raptor	LLNL	Molecular Dynamics	Shake & Bake	Buffalo
CFD	SAGE	LLNL	Molecular Dynamics	MDCASK	LLNL
CFD	TBME	LLNL	Molecular dynamics	Paradis	LLNL
CFD	sPPM	LLNL	Nano-Chemistry	DL_POLY	Argonne
CFD	mpcuglies	LLNL	Neuroscience	pNEO	Argonne
CFD	Nek5	Argonne	neutron transport	SWEEP3D	L'Argonne
CFD	Enzo	Argonne	Nuclear Physics	QMC	Argonne
CFD	TLBE	LLNL	Quantum Chemistry	CPMD	IBM
Financial	KOJAK	NIC, Juelich	Quantum Chemistry	GAMESS	Ames/Iowa State
Financial	Nissei	NIWS	Seismic wave propagation	SPECFEM3D	GEOFRAMEWORK.org
Finite Element Solvers	HPCMW	RIST	Transport	SPHOT	LLNL
Fusion	GTC	PPPL	Transport	UMT2K	LLNL
Fusion	Nimrod	Argonne	Weather & Climate	MM5	NCAR
Fusion	Gyro	GA	Weather & Climate	POP	Argonne

Many Classes of Applications are Massively Parallel



- Candidate Codes:
 - Inherently parallel; written using MPI
 - Memory required per MPI task is less than that available
 - Dominated by collective communication across all nodes
 - Locality of communications within 3D mapping
- Non-Candidate Codes:
 - Large memory footprints required on individual nodes
 - Client/server structures
 - Dominated by disk I/O

Humanity's Top Ten Problems for next 50 years

1. **ENERGY**
2. WATER
3. FOOD
4. ENVIRONMENT
5. POVERTY
6. TERRORISM & WAR
7. DISEASE
8. EDUCATION
9. DEMOCRACY
10. POPULATION



2007	7	Billion People
2050	8-10	Billion People

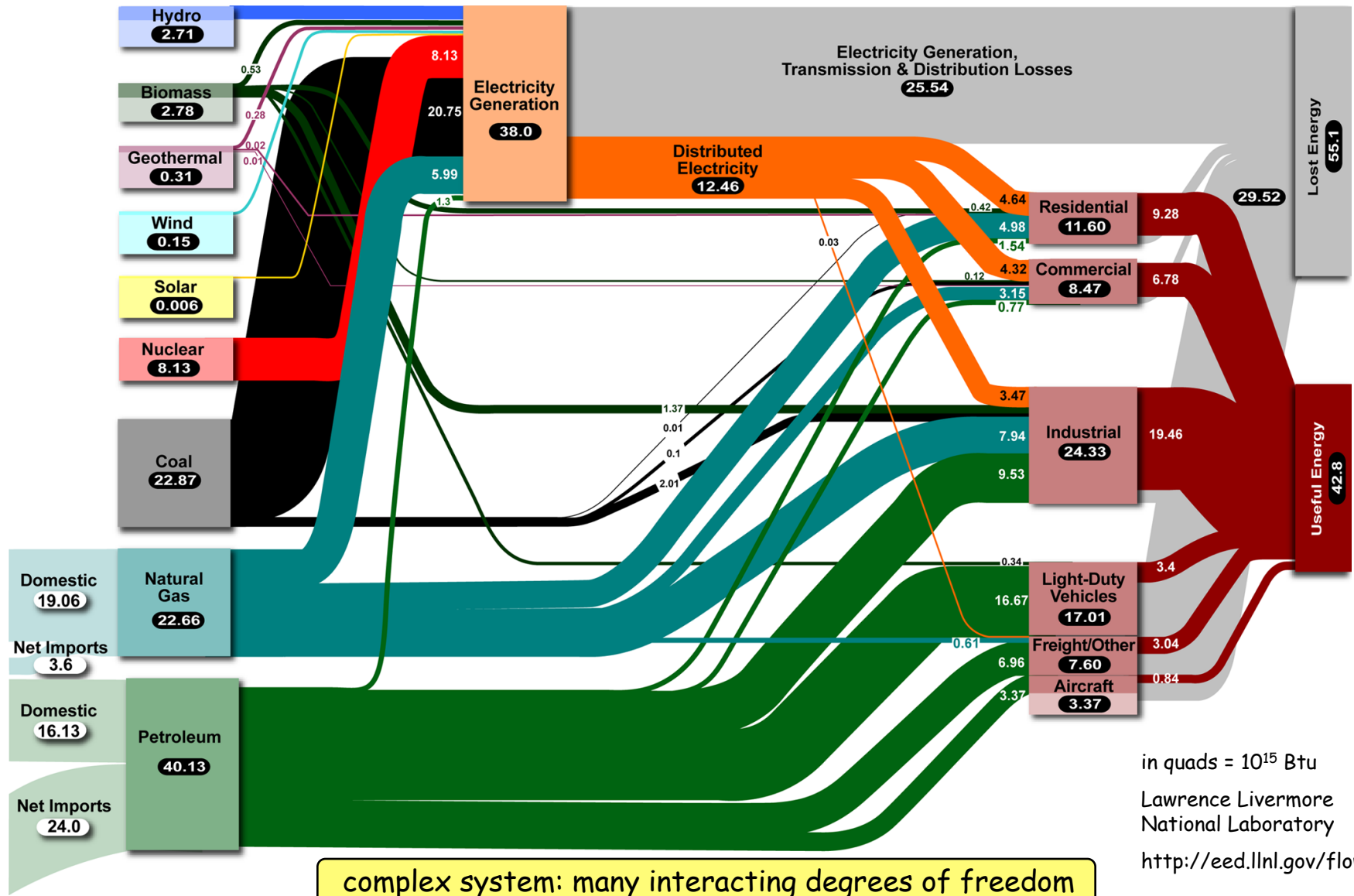
Richard Smalley's Top Ten List

The Grid - the Triumph of 20th Century Engineering



clean versatile power everywhere, at the flick of a switch

Energy Flows in 2005



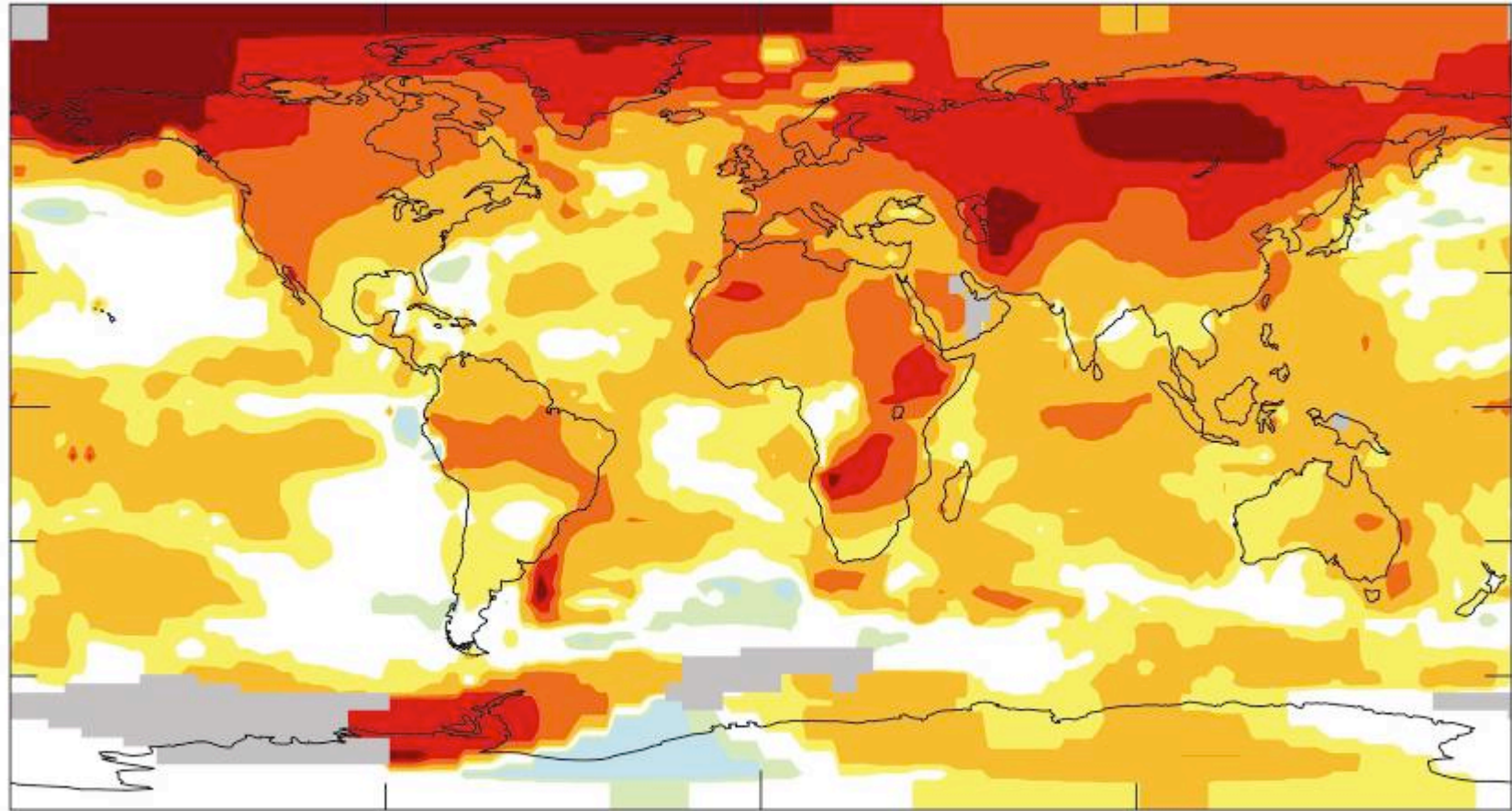
complex system: many interacting degrees of freedom

in quads = 10^{15} Btu
 Lawrence Livermore
 National Laboratory
<http://eed.llnl.gov/flow/>

2001-2005 mean ΔT_{avg} above 1951-80 base, °C

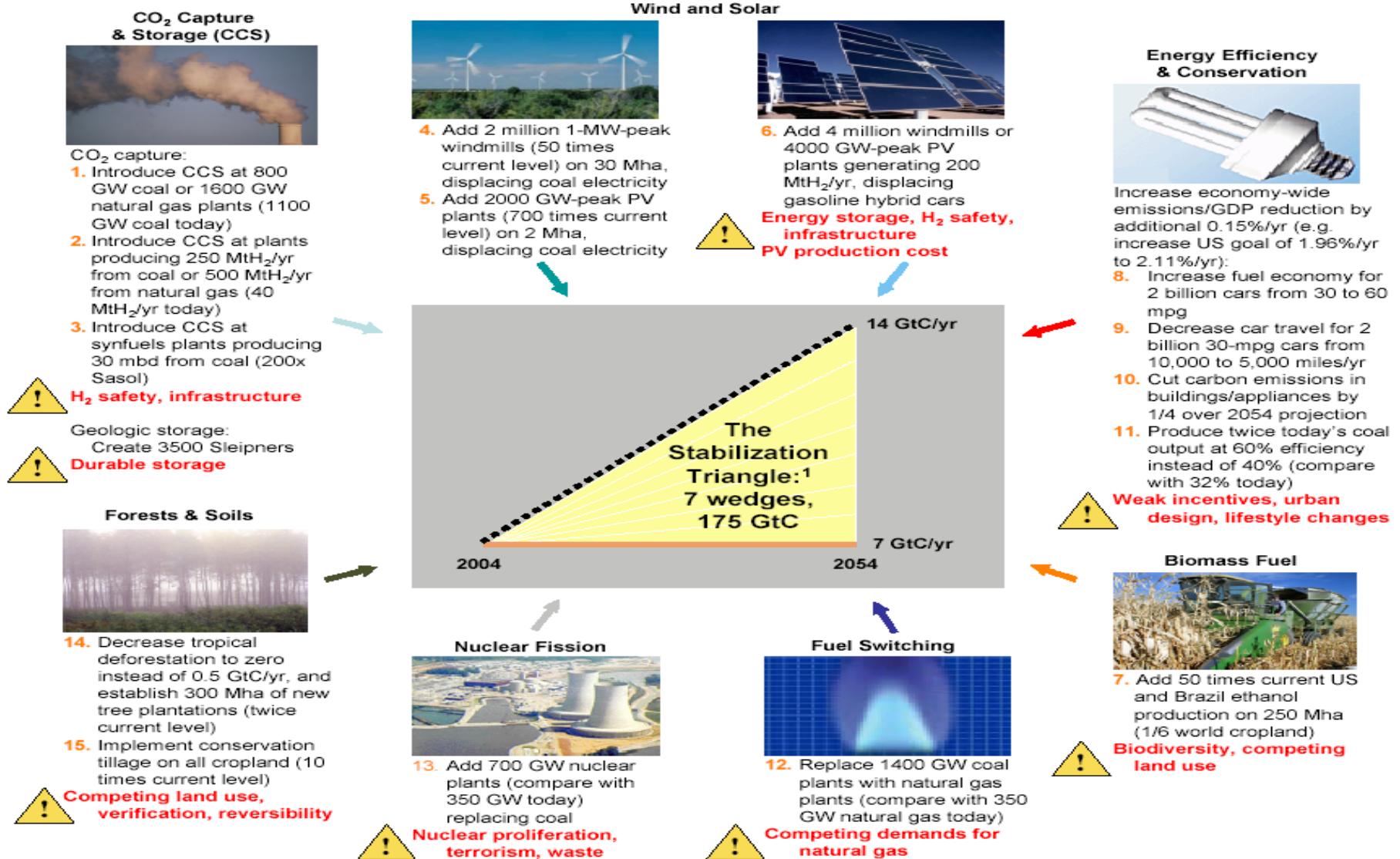
Base Period = 1951-1980

Global Mean = 0.53



Temperature increases are nonuniform: higher mid-continent, highest of all in far North.
(These are observations, not modeling results.)

There are more than 7 wedges to choose from: Here are 15 candidates.




Beyond 2054

More wedges will be needed to maintain the trajectory established by the stabilization triangle, and scaling up the above technologies are unlikely to be enough to satisfy growing energy demand. Therefore, it is imperative that advanced technologies, including **artificial photosynthesis, satellite solar power, nuclear fusion, and geoeengineering strategies** be developed now,³ so that the second and subsequent "runners" have the necessary tools to do their jobs.

References

1. Pacala, S. and R. Socolow, "Stabilization wedges: Solving the climate problem for the next 50 years with current technologies," *Science*, 305, 966 (2004), 13 August.
2. O'Neill, B. C. and M. Oppenheimer, "Dangerous climate impacts and the Kyoto Protocol," *Science*, 296, 1971 (2002).
3. Hoffert, M. I. et al., "Advanced technology paths to global climate stability: Energy for a greenhouse planet," *Science*, 295, 981 (2002).
4. Appenzeller, T., "The end of cheap oil," *National Geographic*, 205, 80 (2004), June.
5. UN Population Division, *World Population in 2300: Proceedings of the United Nations Expert Meeting on World Population in 2300*, United Nations, New York (2004).
6. Siegenthaler, U. and F. Joos, "Use of a simple model for studying oceanic tracer distributions and the global carbon cycle," *Tellus*, 44B, 186 (1992); Joos, F. et al., "An efficient and accurate representation of complex oceanic and biospheric models of anthropogenic carbon uptake," *Tellus*, 48B, 397 (1996).

Modeling and Simulation at the Exascale for Energy and the Environment

 U.S. Department of Energy
Office of Science

Simulation and Modeling at the Exascale
for Energy, Ecological Sustainability and Global Security
An Initiative

The objective of this ten-year vision, which is in line with the Department of Energy's Strategic Goals for Scientific Discovery and Innovation, is to focus the computational science experiences gained over the past ten years on the opportunities introduced with exascale computing to revolutionize our approaches to energy, environmental sustainability and security global challenges.

Executive Summary

The past two decades of national investments in computer science and high-performance computing have placed the DOE at the forefront of many areas of science and engineering. This initiative capitalizes on the significant gains in computational science and boldly positions the DOE to attack global challenges through modeling and simulation. The planned petascale computer systems and the potential for exascale systems shortly provide an unprecedented opportunity for science, one that will make it possible to use computation not only as a critical tool along with theory and experiment in understanding the behavior of the fundamental components of nature but also for fundamental discovery and exploration of the behavior of complex systems with billions of components including those involving humans.

Through modeling and simulation, the DOE is well-positioned to build on its demonstrated and widely-recognized leadership in understanding the fundamental components of nature to be a world-leader in understanding how to assemble these components to address the scientific, technical and societal issues associated with energy, ecology and security on a global scale.

For these types of problems, the time-honored, or subsystems, approach in which the forces and the physical environments of a phenomenon are analyzed, is approaching a state of diminishing returns. The approach for the future must be systems based and simulation programs are developed in the context of encoding all known relevant physical laws with engineering practices, production, utilization, distribution and environmental factors.

This new approach will

- **Integrate, not reduce.** The full suite of physical, chemical, biological, chemical and engineering processes in the context of existing infrastructures and human behavior will be dynamically and realistically linked, rather than focusing on more detailed understanding of smaller and smaller components.
- **Leverage the interdisciplinary approach to computational sciences.** Current algorithms, approaches and levels of understanding may not be adequate. A key challenge in development of these models will be the creation of a framework and semantics for model

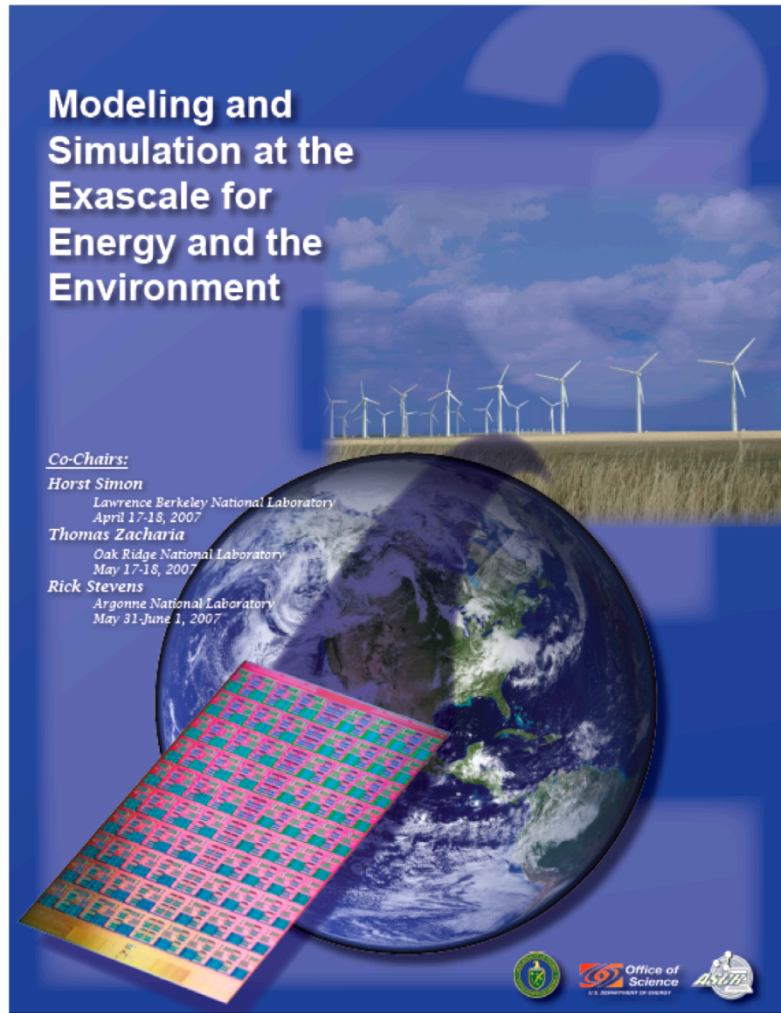
The objective of this ten-year vision, which is in line with the Department of Energy's Strategic Goals for Scientific Discovery and Innovation, is to focus the computational science experiences gained over the past ten years on the opportunities introduced with exascale computing to revolutionize our approaches to energy, environmental sustainability and security global challenges.

Based on this initial white paper, ANL, LBNL, and ORNL organized the community input process in the form of three town hall meetings.

The Opportunity

- Attack global challenges through modeling and simulation
- Planned petascale and the potential exascale systems provide an unprecedented opportunity
- Beyond computation as an critical tool along with theory and experiment
- Understanding the behavior of the fundamental components of nature
- Fundamental discovery and exploration of complex systems with billions of components including those involving humans

Planning for the Exascale Future!

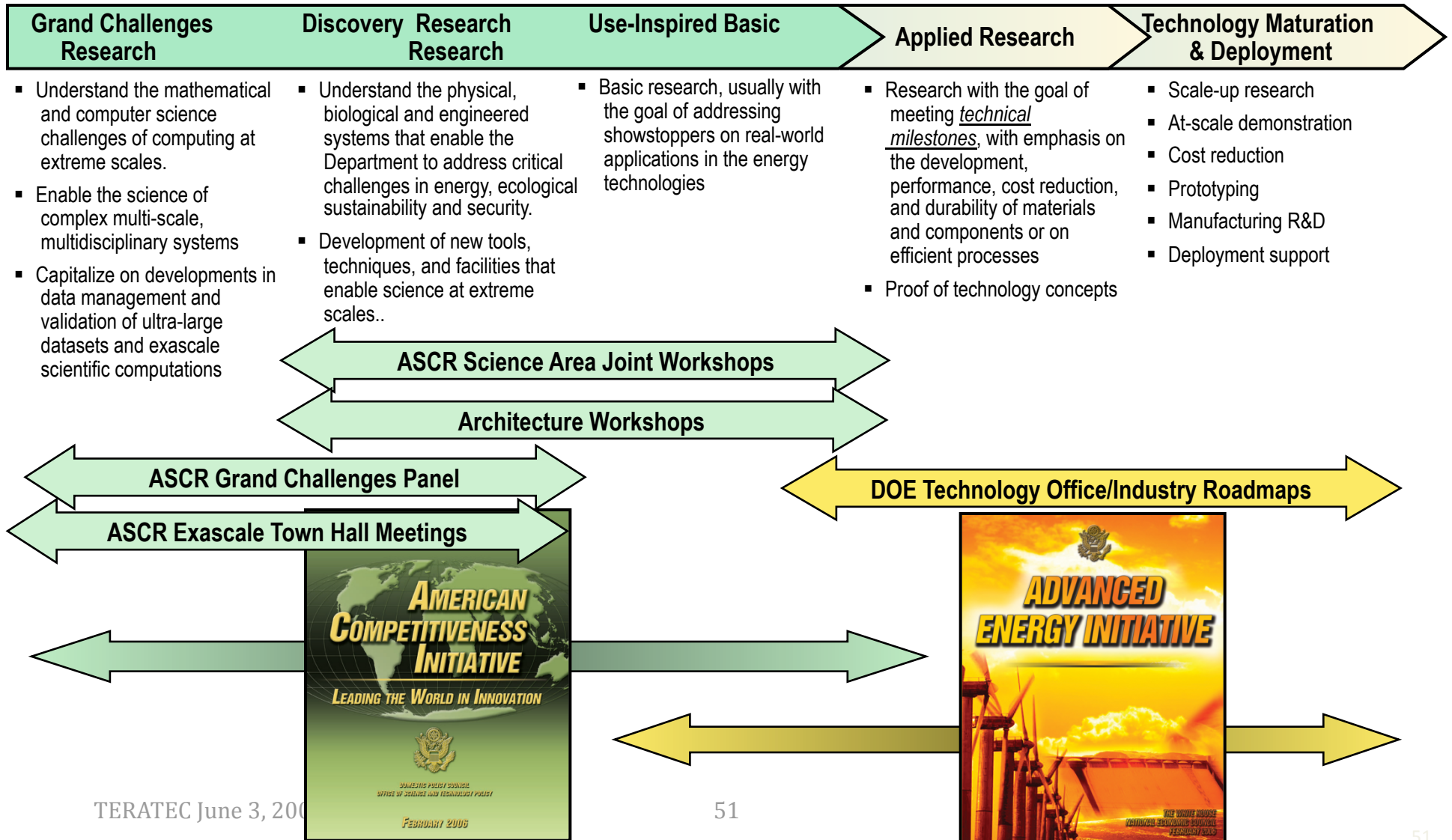


During the spring of 2007
Argonne, Berkeley and Oak Ridge held
three Townhall meetings to chart
future directions

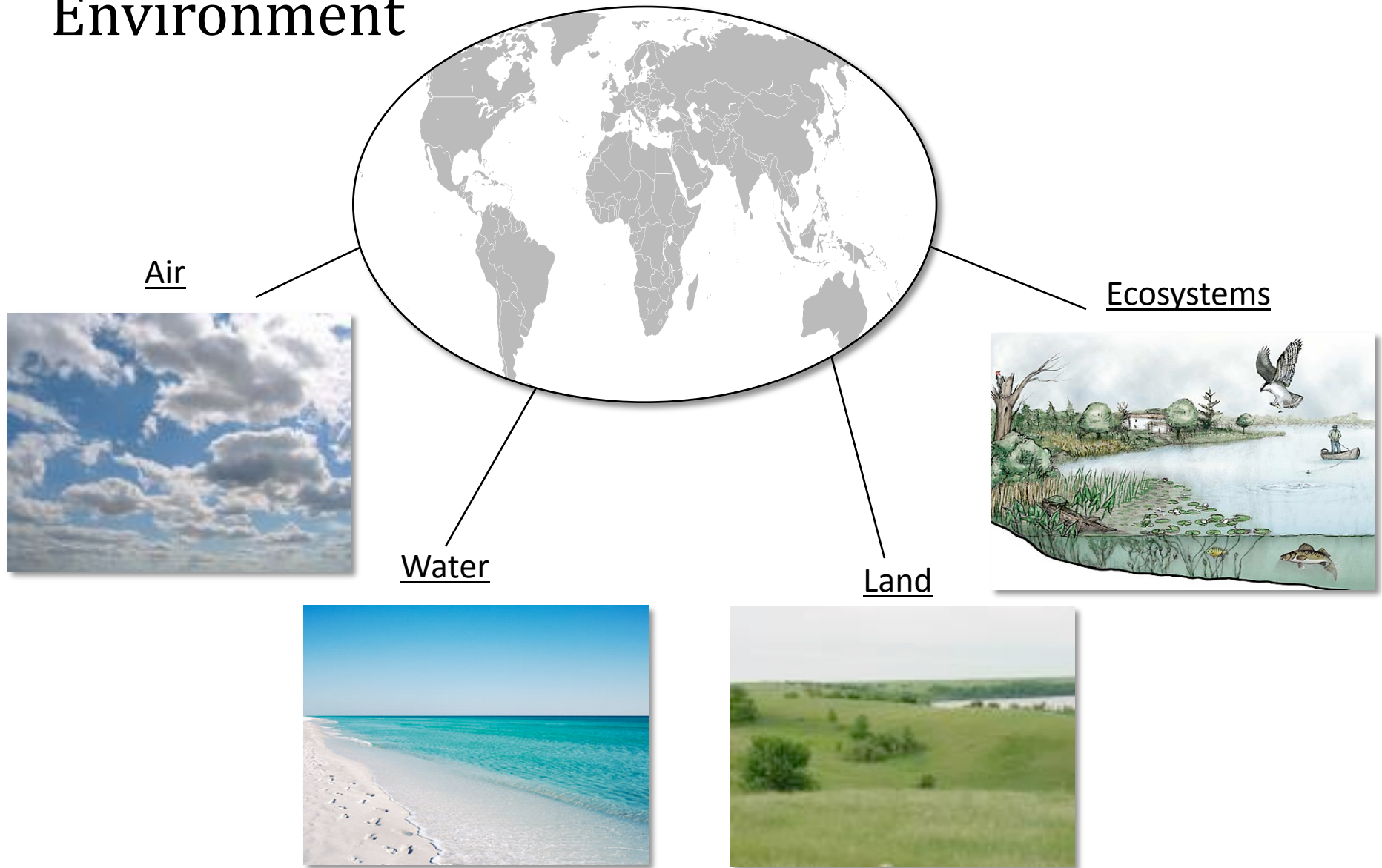
- Exascale Computing Systems
 - Hardware Technology
 - Software and Algorithms
 - Scientific Applications
- Energy
 - Combustion
 - Fission and Fusion
 - Solar and Biomass
 - Nanoscience and Materials
- Environment
 - Climate Modeling
 - Socio-economics
 - Carbon Cycle



Planning for Science at Extreme Scales of Computing



The Economic Systems Sit Within the Physical Environment



Petascale Geoscience

Geoscience Applications

Discipline	Requirement	Current Capability	PCG Capability
Climate Modelling	5 simulated yrs/day	Resolve atmosphere and ocean at 110 km and larger, parameterize mesoscale processes	Directly resolve mesoscale structure of ocean (10km) and atmosphere (20km)
Oceanography	40 simulated yrs/month	10-20 km eddy-permitting global circulation models	5-10 km eddy-resolving global circulation models coupled to ecosystem models with 10-20 biological constituents
Weather Research	2 simulated hrs/day	3 km thunderstorm simulation	10m tornado simulation
Seismology, Earthquake simulation	10 global earthquake simulations/month	O(10 billion) grid points: global seismic wave analysis limit 0.3Hz	O(500 billion) grid points: global seismic wave resolution at or better than 1 Hz
Seismology: Imaging Earth's Interior	1 global assimilation of thousands of earthquakes per month	1000 km resolution of Earth's interior	Imaging at 100 km resolution of core boundary in Earth's interior
Hydrology	1 decadal basin-scale simulation per week	1 year simulation of 1 km Rio Grande river basin	Decadal 1 km Columbia River Basin (100 times larger than Rio Grande River Basin)
Space Weather	Coronal mass ejection faster than real time	Resolve magnetic configuration associated with large sunspots 1/40 solar radius	Resolve fine structure of corona magnetic field inside active regions: 1/320 solar radius

Src: Petascale Collaboratory for the Geosciences, 2005

WRF

- Modern code, candidate for extensive work
 - Single source code tree w/layered sw architecture
 - Multilevel parallel decomposition
- High res simulations or ensembles
- Performance model for estimates
 - Tornado: 2 hr simulation with 10m resolution and 2-category microphysics
 - 100km x 100km x 20 km domain should be effective on 62,500 processors using 40x40 2D horizontal subdomains
 - 150 TF sustained
 - Using more realistic microphysical parameterization with 5 categories will double the computation time
 - Exploratory work with 100s of microphysical variables

Geosciences Applications Requirements

Application Name/Discipline	Problem	Max Required Sustained TFLOPS	System Memory (Tbytes)	Mass Storage Archive Rate (Pbytes/year)	Disk Bandwidth for 5% overhead (Gbytes/sec)
flow_solve/oceanography	3-D turbulence	2.5	6.5	0.14	1.1
POP/oceanography	10 km global mesoscale eddy	6	0.15	0.32 to 3.2	0.2 to 2.0
POP/oceanography	5 km global mesoscale	120	1.5	3.2 to 32	2 to 20
MITgcm/ocean data assimilation	15 km global ocean	7.3	0.82	0.66	0.4
WRF/meteorology	10m tornado simulation	150	20	2 to 24	25 to 300
	5 years of 3 km global nonhydrostatic simulation	66	1.75	1	8
CAM/climate modeling	5 instances of T341L52	13	0.5	4.6	1.1
CRCP/climate modeling	2 km global sub-grid scale model	22	-	-	-
ABINIT/minerology	DFT calculation	1.6	-	-	-
inverse problem/regional seismology	100M point inverse problem	17	0.01	0.12	0.07
forward problem/global seismology	36.6 billion degrees of freedom	10.4	7.3	0.01	0.0002
LADHS/regional hydrology	100m Columbia river basin	10	0.3	20.8	0.66

Src: Petascale Collaboratory for the Geosciences, 2005

WRF

- High resolution global models
 - Below 5 km, scales and physics change
- Global non-hydrostatic numerical weather model
 - 2 km resolution requires ~200 TF sustained
 - 1 km requires 1.6 PF sustained
- Major research problem just getting started

Resolution (km)	TFLOPS sustained to achieve 60 days/day	TFLOPS sustained to achieve 5 years/day	Global WRF Data volume TB/sim year
1	1609	48260	1892
2	212	6350	466
3	66	1975	206
4	29	875	116
5	15	467.5	74
8	4.3	129	29
10	2.4	71	18.5

Src: Petascale Collaboratory for the Geosciences, 2005

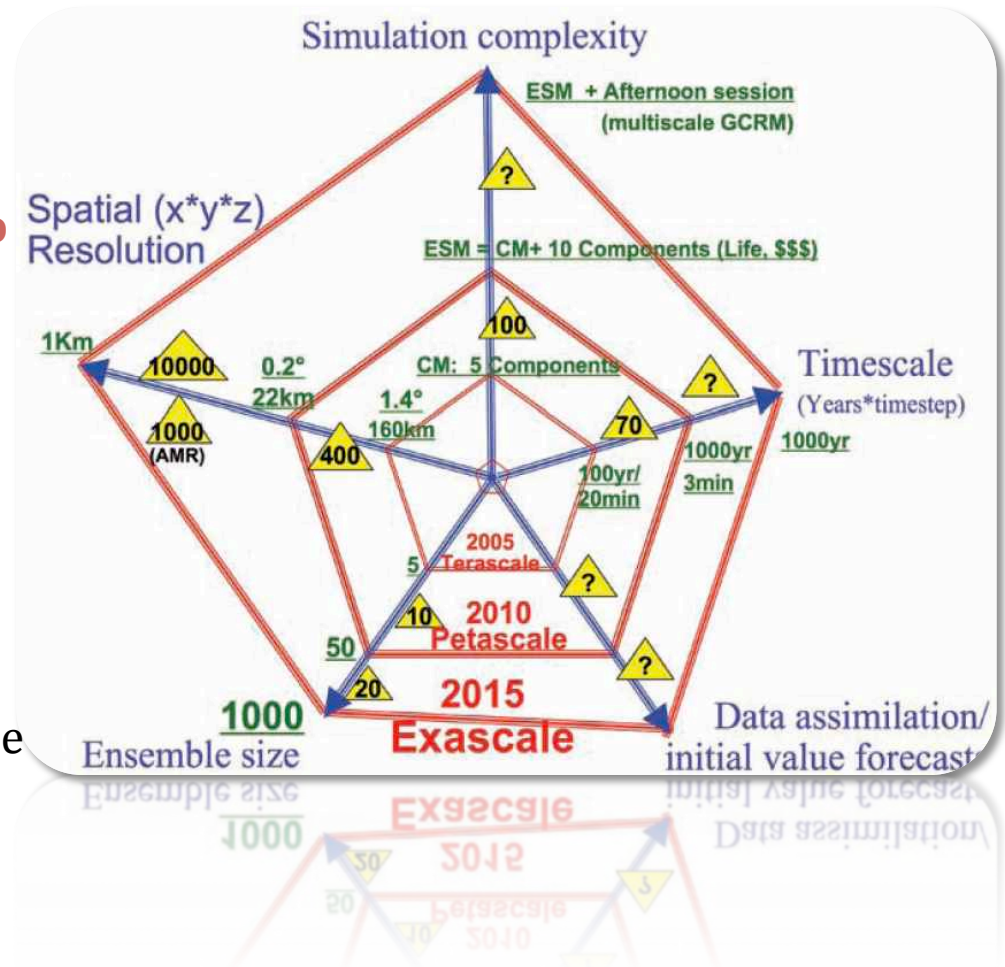
Reliable Climate Forecasts from Next Generation Earth System Models

- **Key Challenges**

- High certainty forecasts for the next few decades
- Long term forecasts relevant to regional/community scales

- **Urgent Questions for Petascale to Exascale Simulations**

- Carbon sequestration option models
- Systems understanding of carbon-climate coupling
- Triggering mechanisms for extreme weather shifts
- Stability/sustainability of tropical rainforests and polar ice caps
- Sustainability of sea and land/agricultural ecosystems

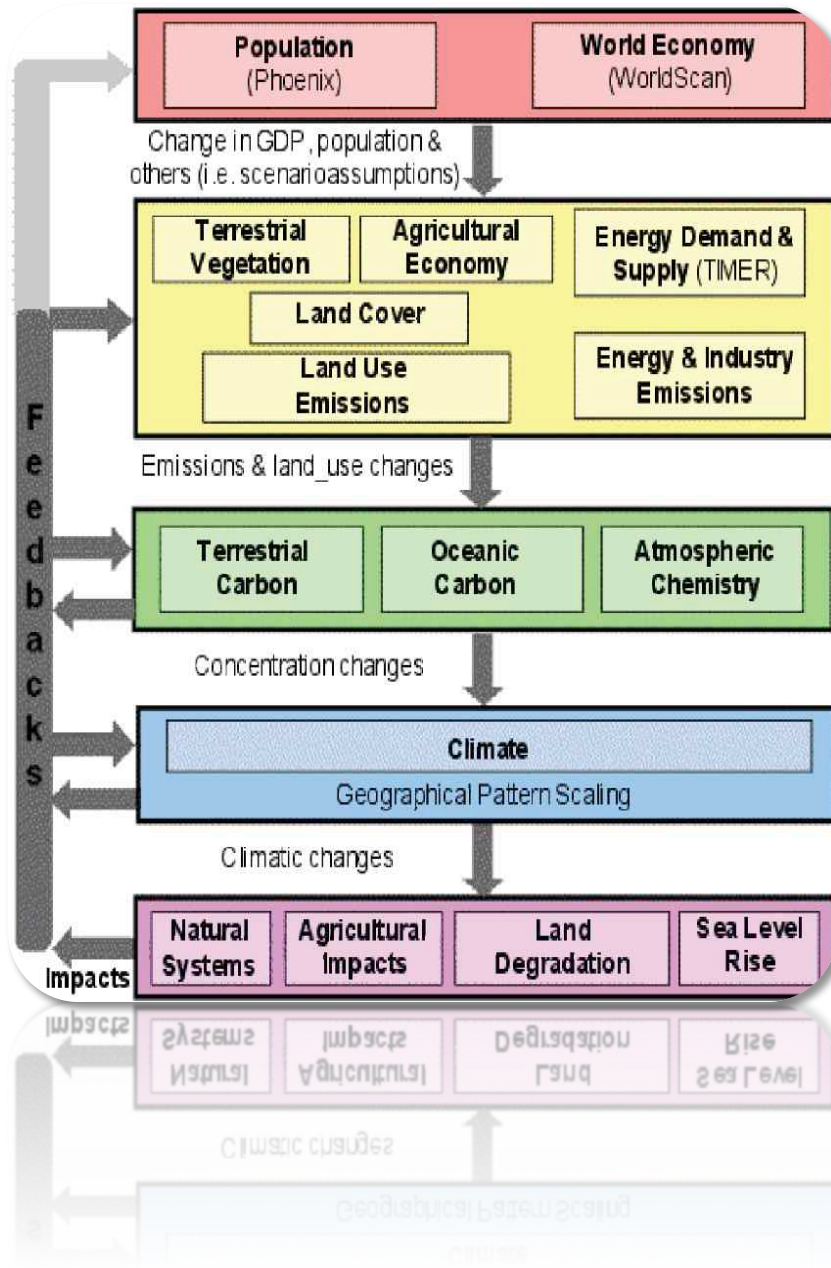


Trajectory of Climate Model Developments



Figure 11: Development of Climate Models over the Last 30 Years. The development of climate models over the last 30 years showing how the different components are first developed separately and later coupled into comprehensive climate models. *Credit: CCSP Strategic Plan, Chapter 10 (2003).*

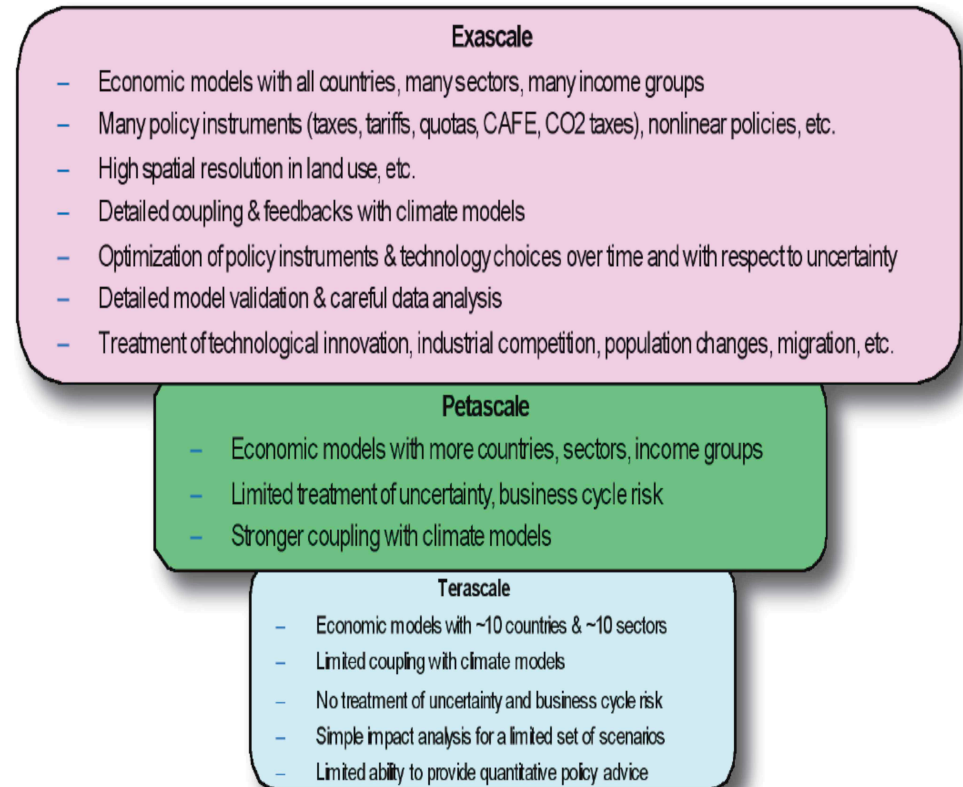
From Earth System Modeling to Computational Socio-Economics



- Earth system modeling has progressed to a point where there is considerable confidence in predictions of continental- and global-scale climate changes over the next 100 years [IPCC 2007]
- Integrated modeling of the social, economic, and environmental system with an extensive treatment of couplings among these different elements and consequent nonlinearities and uncertainties would have great impact.
- Computational limitations have prevented existing models from including substantial regional and sectoral disaggregation, dynamic treatment of world economic development and industrialization, and detailed accounting for technological innovation, industrial competition, population changes and migration.

Impact of Socio-Economic Modeling

- Emergence of petascale and prospect of exascale computers enable a fully integrated treatment of diverse factors.
- Models have potential to transform understanding of socio-economic-environmental interactions.
- How will climate change impact energy demand and prices?
- How will nonlinearities, thresholds, and feedbacks impact both climate and energy supply?
- How will different adaptation and mitigation strategies effect energy supply and demand, the economy, the environment, etc.?
- How can computational approaches help identify good strategies for R&D, policy, and technology adoption under conditions of future uncertainty?



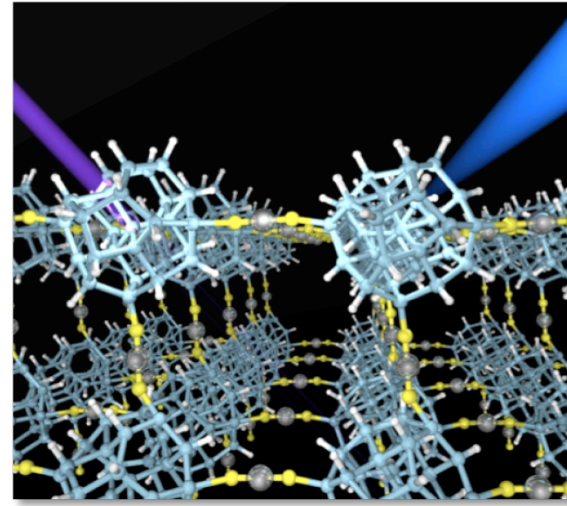
Nanoscale Materials by Design

Major challenges in nano/materials science

1. Numerical approximations and models for accurate physics and properties
2. Integrated diverse models to simulate the whole system or process
3. Large-scale systems (>100K atoms) and long duration dynamics (nanoseconds or microseconds)

Requires both computers larger than petascale and algorithms with better scaling with problem size

Today's $O(N^3)$ DFT methods will be limited to $\sim 50K$ atom single point electronic structures on petaflops



Addressing these issues opens many valuable design avenues

- Optimal materials for dense hydrogen storage
- Inexpensive, efficient and environmentally benign solar cells
- Nanostructured data storage
- Bio-nano electronics

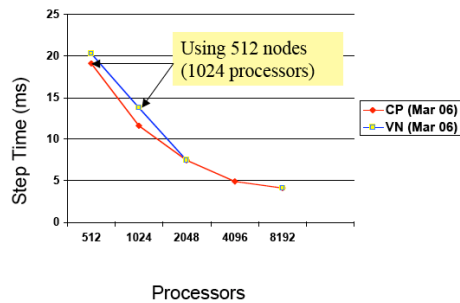
These problems each have very large parameter spaces, so design optimizations take many runs

Petascale Molecular Modeling

NAMD on BlueGene/L

Procs	Time (ms)
1	9000
32	347
128	97.2
512	23.7
1024	13.8
2048	8.6
4096	6.2
8192	5.2

Oct'05

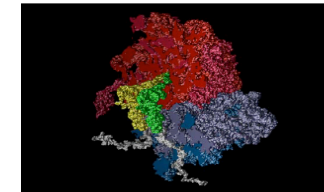
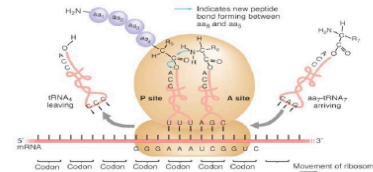


9/6/06

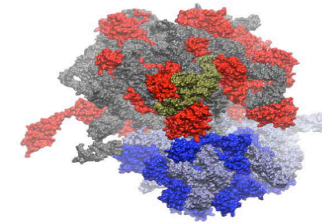
Parallel Programming Lab, SIAM 06

10

Petascale Project 2: Ribosome

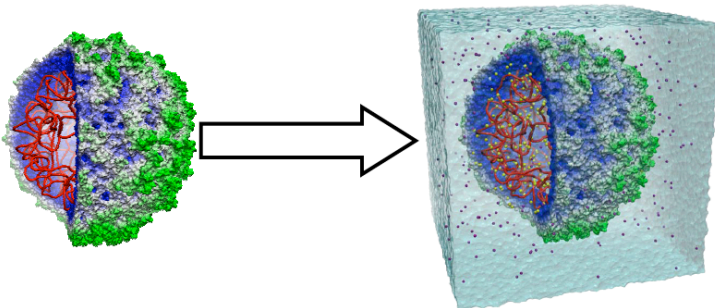


- Very large system size: ~3,000,000 atoms
- Great biological and biomedical relevance
- Simulations with close collaboration with leading experimentalist
- All-atom, coarse graining, and multiscale simulations



Petascale Project 1: Virus Capsid

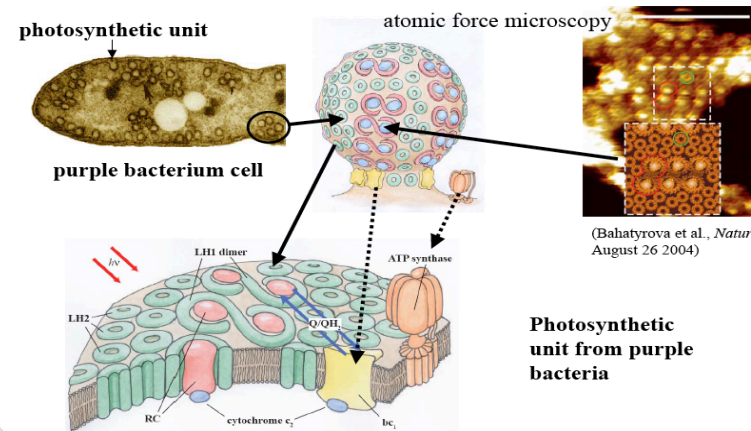
Solvate the virus in a 220Åx220Åx220Å water box, add Mg²⁺ ions to neutralize RNA and Cl⁻ ions to neutralize the protein



132,000 atoms of protein, 30,000 atoms of RNA,
~1,000,000 atoms in total

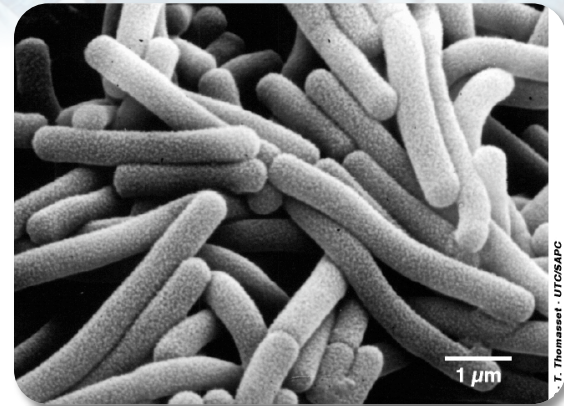
Peter Freddolino, Anton Arkhipov, Steven Larson, Alexander McPherson, and Klaus Schulten, *Structure*, 14:437 (2006)

Petascale Project 3: Chromatophore

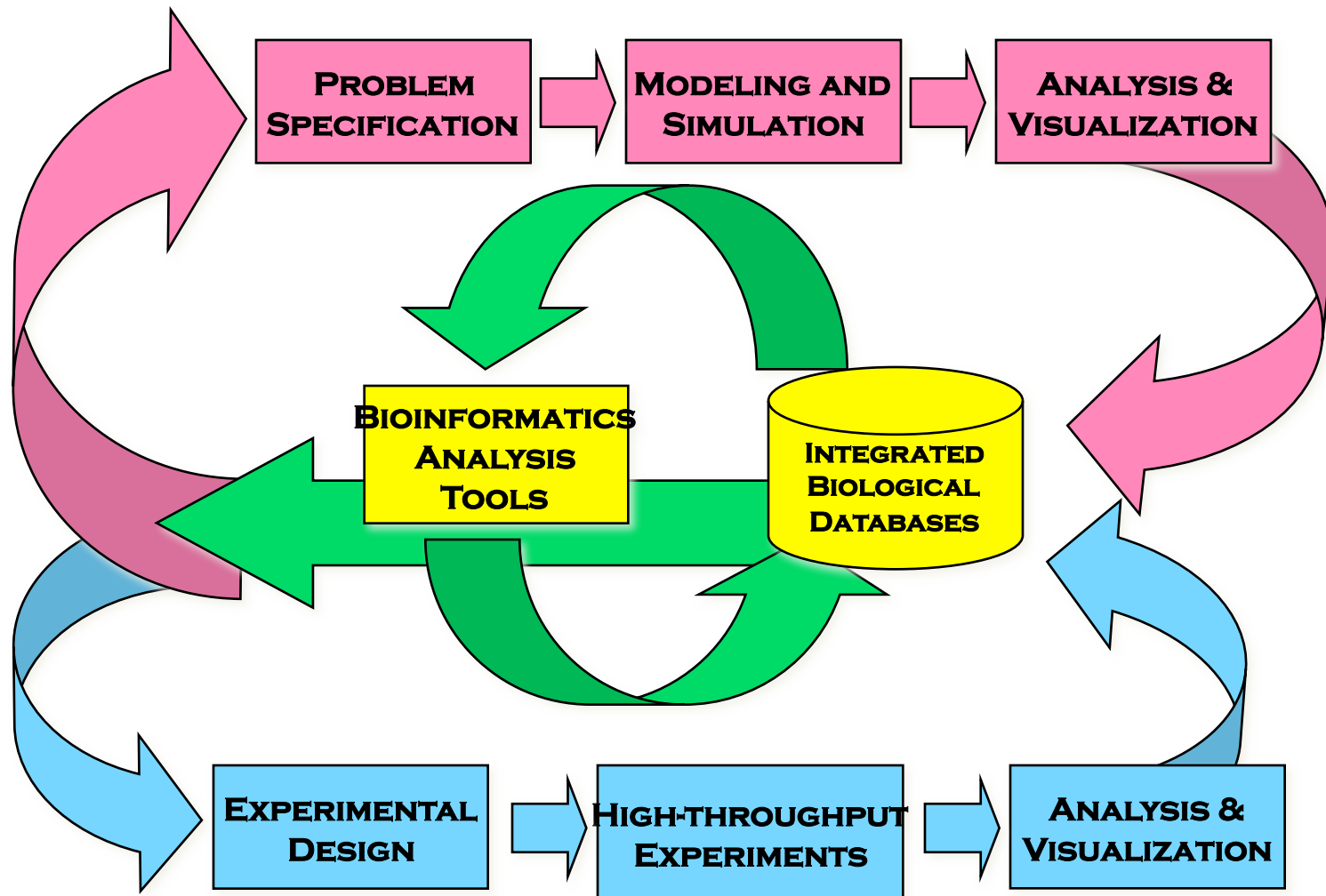


Petascale Impact on Biological Theory

- **Potential high impact on theory development**
 - The ability to run large-scale simulations that can capture non-trivial variation in an evolutionary process could have a dramatic impact on our ability to move from qualitative to quantitative theory in biology
- **Software readiness for petascale systems**
 - While physical process oriented software is on a trajectory to achieve scalable performance on petascale systems, agent based evolution and ecosystem modeling environments are lagging far behind
 - Data analysis and bioinformatics environments are in the middle, hindered in part by the lack of data intensive infrastructure
- **Capability and capacity computing estimates**
 - First principles MD and QM simulations have enormous computing requirements, but perhaps limited impact on large-scale theory
 - Agent based simulations have not been effectively scoped
- **Related experimental support is needed**
 - Validation experiments driven by the simulation and modeling will be required



An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics

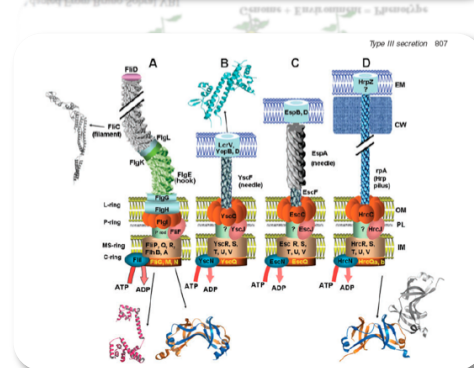
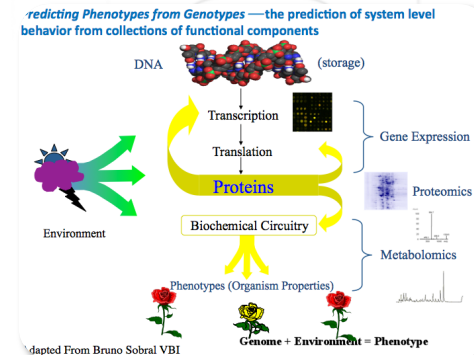
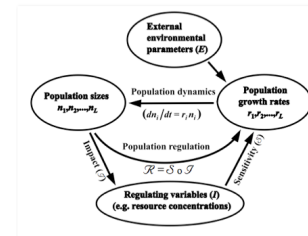


Six Open Problems in Basic Biology

Where Computing Can Have an Impact

1. **Applicability of the Competitive Exclusion Principle** — the nature and scale of ecological niches and relationships between competition and diversity
2. **Predicting Phenotypes from Genotypes** — the prediction of system level behavior from collections of functional components
3. **Understanding the Evolution of Biological Networks** — structure, complexity and mechanisms
4. **Reconstruction of Horizontal Gene Transfer Events** — rapid evolution of complexity and non-inherited adaptation mechanisms
5. **Understanding the Range of Permitted Biologies** — possible origins and the fundamental limits to life and life processes
6. **Understanding Convergent Evolution** — the repertoire of form and function, independent evolution of similar structures or functions in similar or different environments

Framework for Modeling Diversity and Niche Exclusion



Emergent Biogeography of Microbial Communities in a Model Ocean

Michael J. Follows,^{1*} Stephanie Dutkiewicz,¹ Scott Grant,^{1,2} Sallie W. Chisholm³

Fig. 1. Annual mean biomass and biogeography from single integration. (A) Total phytoplankton biomass ($\mu\text{M P}$, 0 to 50 m average). (B) Emergent biogeography: Modeled photo-autotrophs were categorized into four functional groups; color coding is according to group locally dominating annual mean biomass. Green, analogs of *Prochlorococcus*; orange, other small photo-autotrophs; red, diatoms; and yellow, other large phytoplankton. (C) Total biomass of *Prochlorococcus* analogs ($\mu\text{M P}$, 0 to 50 m average). Black line indicates the track of AMT13.

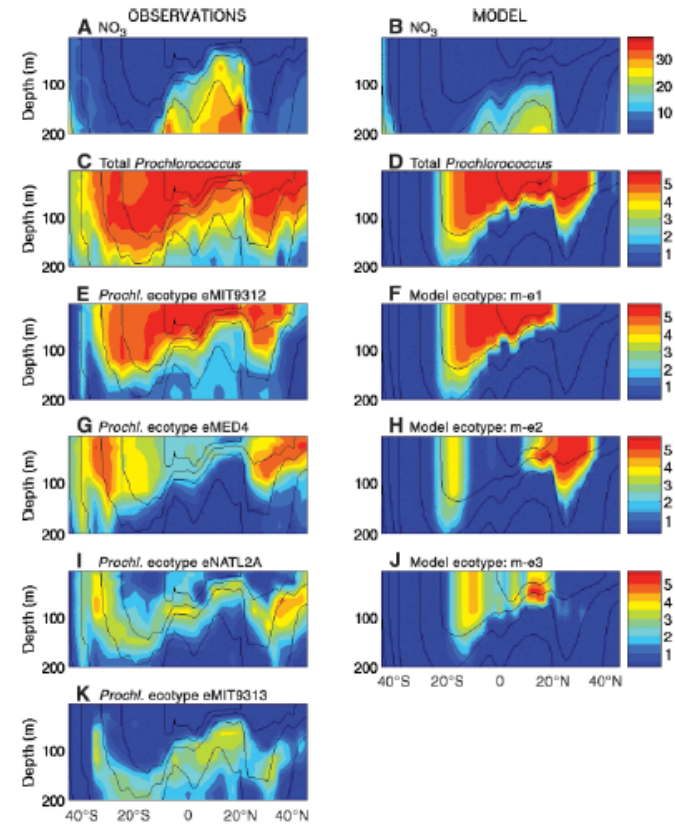
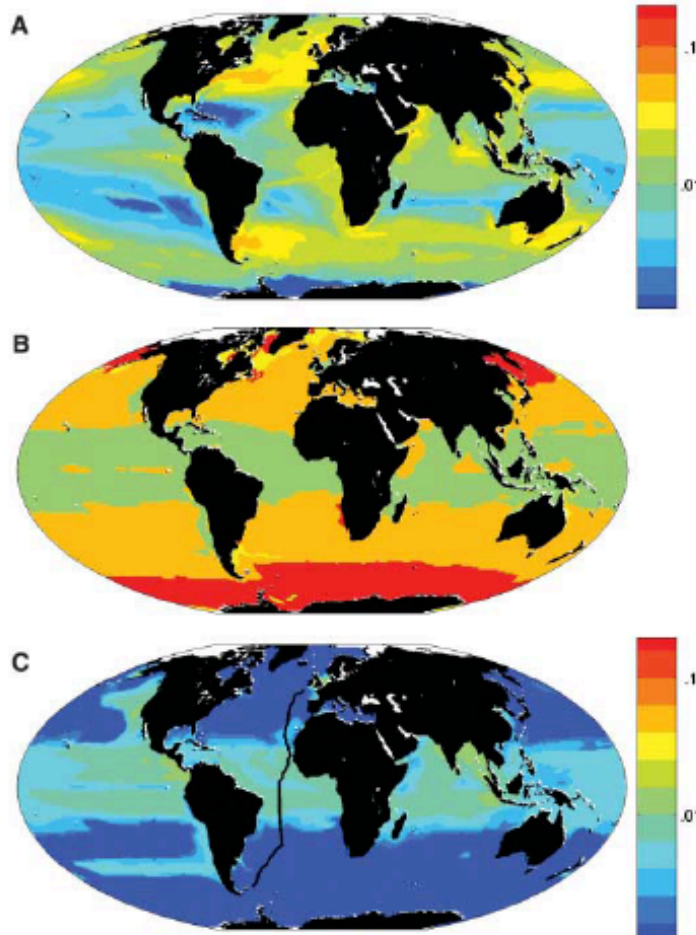


Fig. 2. Observed and modeled properties along the AMT13 cruise track. Left column shows observations (17), right column shows results from a single model integration. (A and B) Nitrate ($\mu\text{mol kg}^{-1}$); (C and D) total *Prochlorococcus* abundance [$\log(\text{cells ml}^{-1})$]. (E, G, I, and K) Distributions of the four most abundant *Prochlorococcus* ecotypes [$\log(\text{cells ml}^{-1})$] ranked vertically. (F, H, and J) The three emergent model ecotypes ranked vertically by abundance. Model *Prochlorococcus* biomass was converted to cell density assuming a quota of 1 fg P cell^{-1} (27). Black lines indicate isotherms.

Challenges for Cell and Ecosystem Simulation

- Modeling cells rivals the complexity of climate and earth systems models
 - Multiple space and time scales
 - Millions of interacting parts
 - Populations of cells to understand emergent behavior
 - Integrated modeling necessary to advance theory in systems biology
- Cell and ecosystems modeling will need Petascale computing and beyond
 - Dynamics of evolution
 - Genomics driven medicine

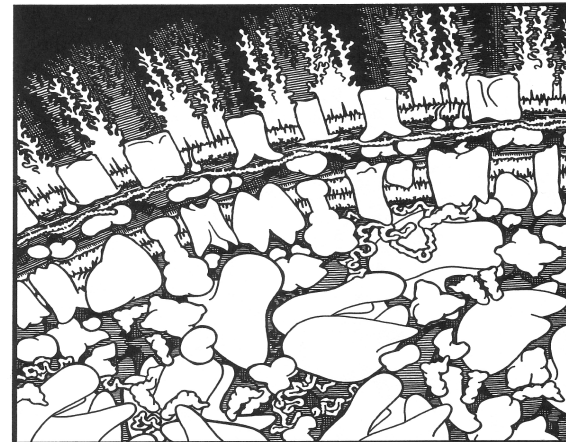
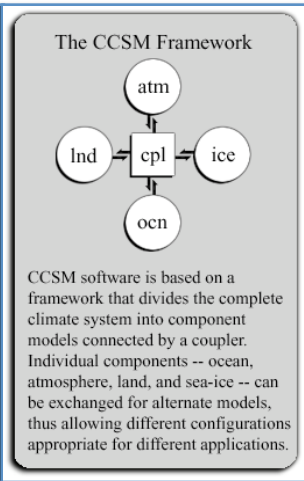
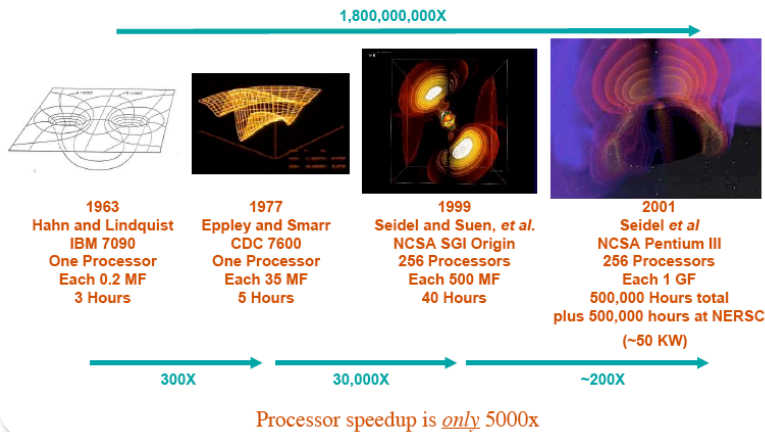


Figure 4.3 Cell Wall



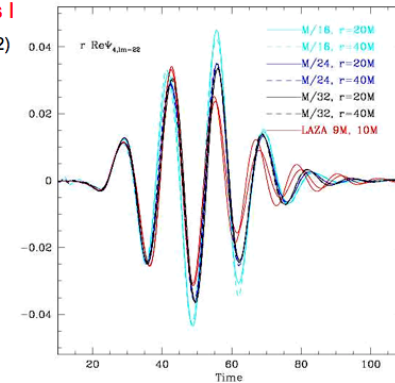
Colliding Black Holes

Black Hole Collision Problem



QCQ Waveforms

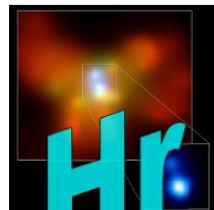
- Waveforms (Re $L=2, M=2$ mode) from three runs, $M/16, M/24, M/32$ extracted at $r_{\text{extract}}=20M$ (Solid), $40M$ (Dashed). Plotted are $(r \times \Psi_{lm})$.
- Good $O(1/r)$ propagation behavior; $M/24, M/32$ are very close.
- Comparison with **Lazarus I**



General Relativity 2005 Workshop, NASA/GSFC, NOV 2, 2005

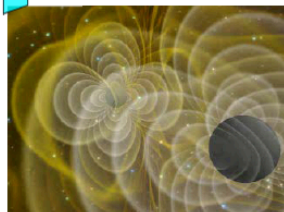
Colliding Black Holes

- Centrella, et al.
 - April 18, 2006
 - 4 orbits before infall
 - NASA Columbia system
 - 2032 Itanium2 Processors
 - 80 Hours total
 - "Combination of various crucial tools such as adaptive mesh refinement, Newman-Penrose scalars, waveform extraction, and novel gauge conditions in general relativity that critically contributed to results." Choi
 - "Code Performance (speedup): Scalability demonstrated up to ~864 processors now with highly complicated mesh structure: code scales with 90—95% efficiency." Choi



Chandra Image July, 2001

Credit: NASA/CXC/UM/ E.S. Komossa et al.



Credit: Henze, NASA

Centaurus A: X-ray Light View



NASA / SAO / R. Kraft, et al.

A telltale sign of a black hole is a high-energy jet blasting into space. This galaxy has a supermassive black hole in its center!

A telltale sign of a black hole is a high-energy jet blasting into space. This galaxy has a supermassive black hole in its center!

Quantum Chromodynamics

- Calculate weak interaction matrix elements of strongly interacting particles to the accuracy needed to make precise test of the standard model
- Determine the properties of strongly interacting matter at high temperatures and densities, such as those that existed immediately after the big bang
- With BG/Q (and beyond) data is cache resident, so memory access is not a factor
- However latency could be a big deal at exaflops, bounding scaling of present approaches [IBM Study]

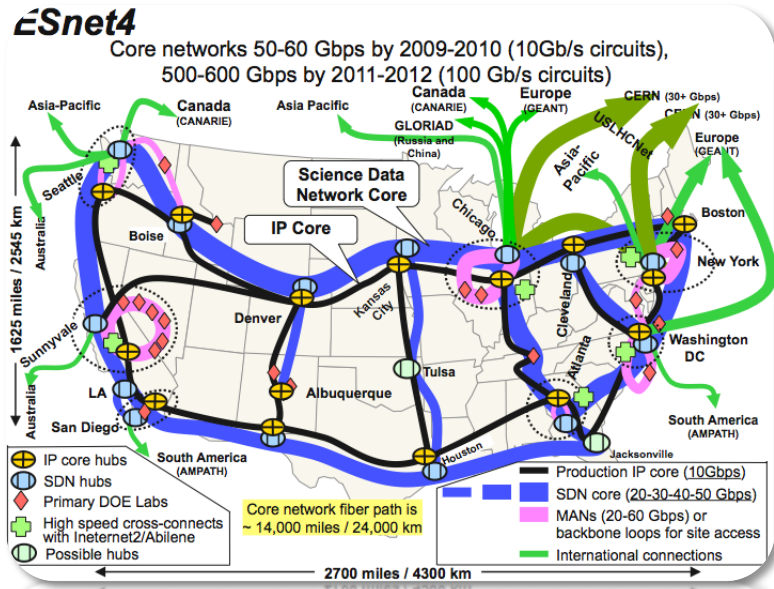
BG/P Configuration Generation Plans

QCD Action	Lattice Spacing (Fermi)	m_l/m_s	Lattice Dimensions	Lattice Size (GB)	TF-Years
ASQTAD	0.060	0.10	$60^3 \times 144$	9.0	2.0
ASQTAD	0.045	0.20	$56^3 \times 192$	9.7	1.9
ASQTAD	0.045	0.10	$80^3 \times 192$	28.3	13.7
ASQTAD	0.060	0.05	$84^3 \times 144$	24.6	23.2
DWF	0.094	0.27	$32^3 \times 64$	0.6	1.2
DWF	0.094	0.19	$48^3 \times 64$	2.0	7.8
DWF	0.094	0.11	$48^3 \times 64$	2.0	25.2
CLOVER	0.100	0.22	$32^3 \times 128$	1.2	0.8
CLOVER	0.100	0.15	$40^3 \times 128$	2.4	4.1
CLOVER	0.080	0.18	$40^3 \times 128$	2.4	4.5
CLOVER	0.080	0.15	$48^3 \times 128$	4.1	22

Lattice QCD calculations have 2 stages

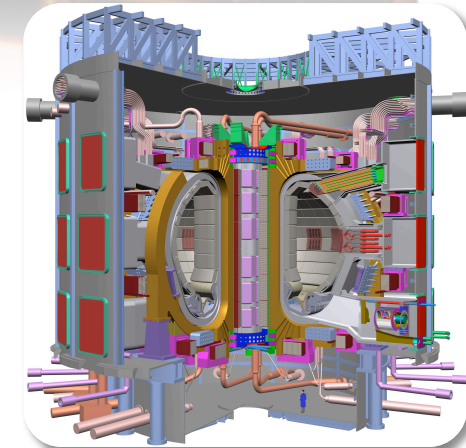
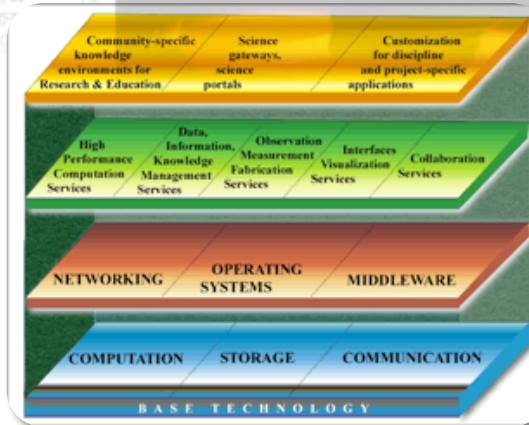
1. Monte Carlo methods generate representative configurations of the QCD ground state -- time intensive
2. Use configurations to calculate a wide variety of quantities of interest in high energy and nuclear physics.

Integrating Leadership Computing Into the International Research Infrastructure



Emerging Areas and Directions

- Economics
- Large-scale Optimization
- Large-scale Data Analysis
- Agent-Based Models (cyber security, evolution, social organizations)
- Parallel Symbolic Computation
- Interactive Exploratory Analysis
- Sensor Network Data Assimilation
- Comparative Genomics
- Cell Network Models



Some Final Words

- Scientific breakthroughs require flexibility and abundance of computing resources for serendipity and insight to work.
 - One must be able to make lots of mistakes.. therefore cost matters to make mistakes affordable
- High-capability platforms require considerable quantities of capacity platforms to make the capability effective.
 - We learn this from the distribution of computing allocations at major centers.. most scientific computing is warm-up exercises..
- The community needs a long term commitment not just to developing new high-end architectures, but also to deploying them as well supported infrastructure.
 - Scientists are very good at optimizing their time and generally will not respond to speculative availability of resources..

