Getting Ready for Exascale Science

Rick Stevens
Argonne National Laboratory
University of Chicago

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?

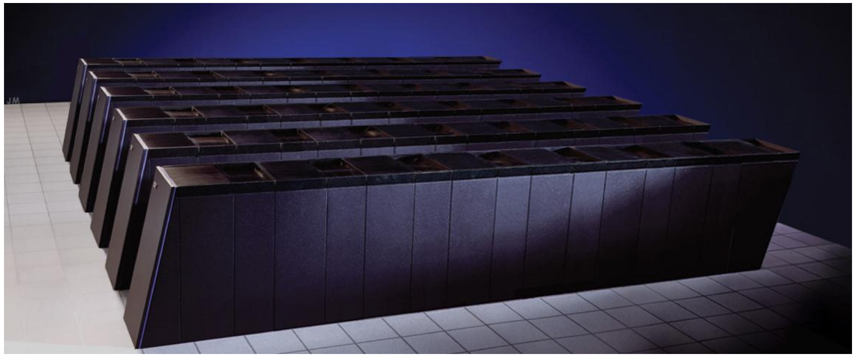
Argonne Leadership Computing Facility

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

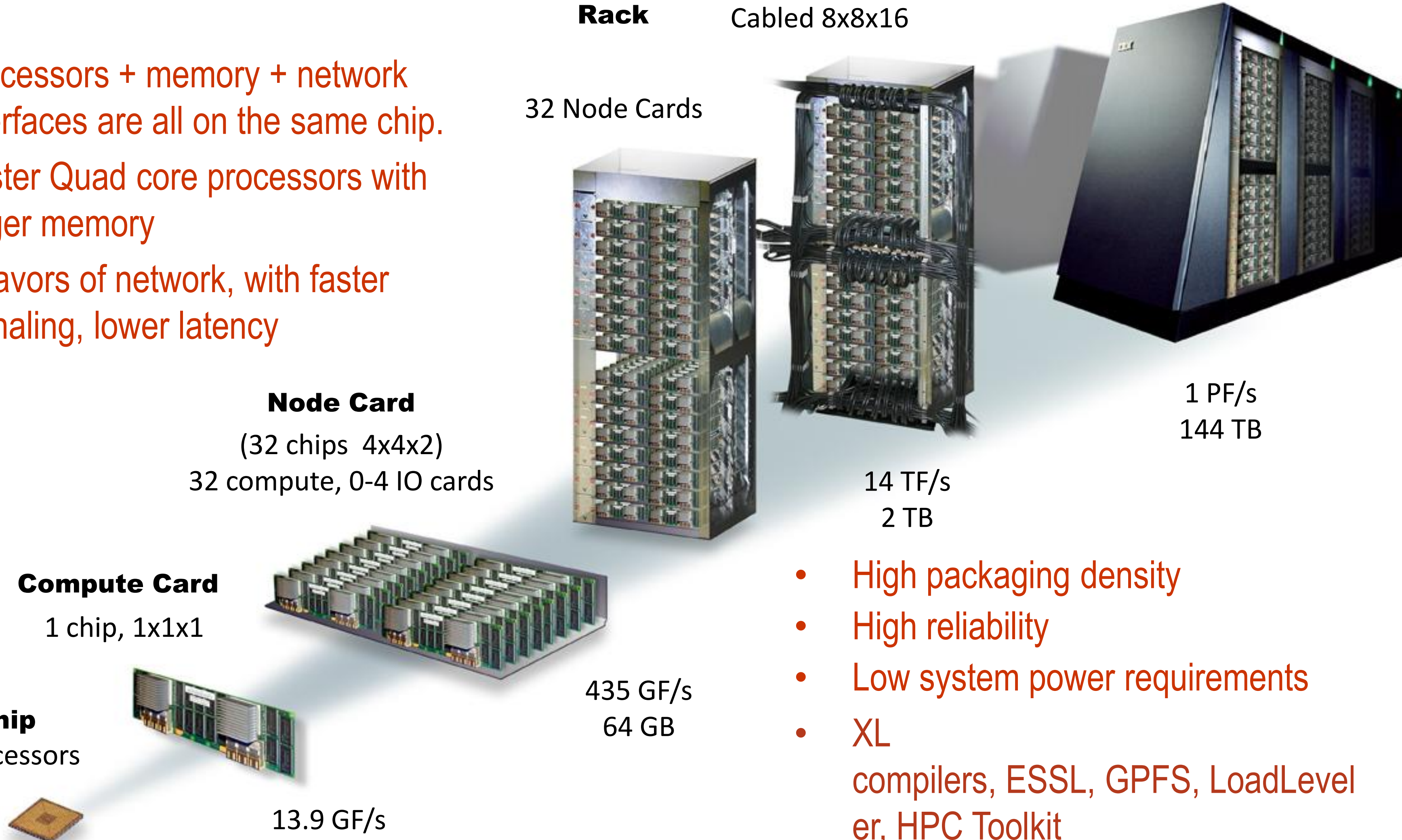
In 2004 DOE selected the ORNL, ANL and PNNL team based on a competitive peer review
- ORNL to deploy a series of Cray X-series systems
- ANL to deploy a series of IBM Blue Gene systems
- PNNL to contribute software technology



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

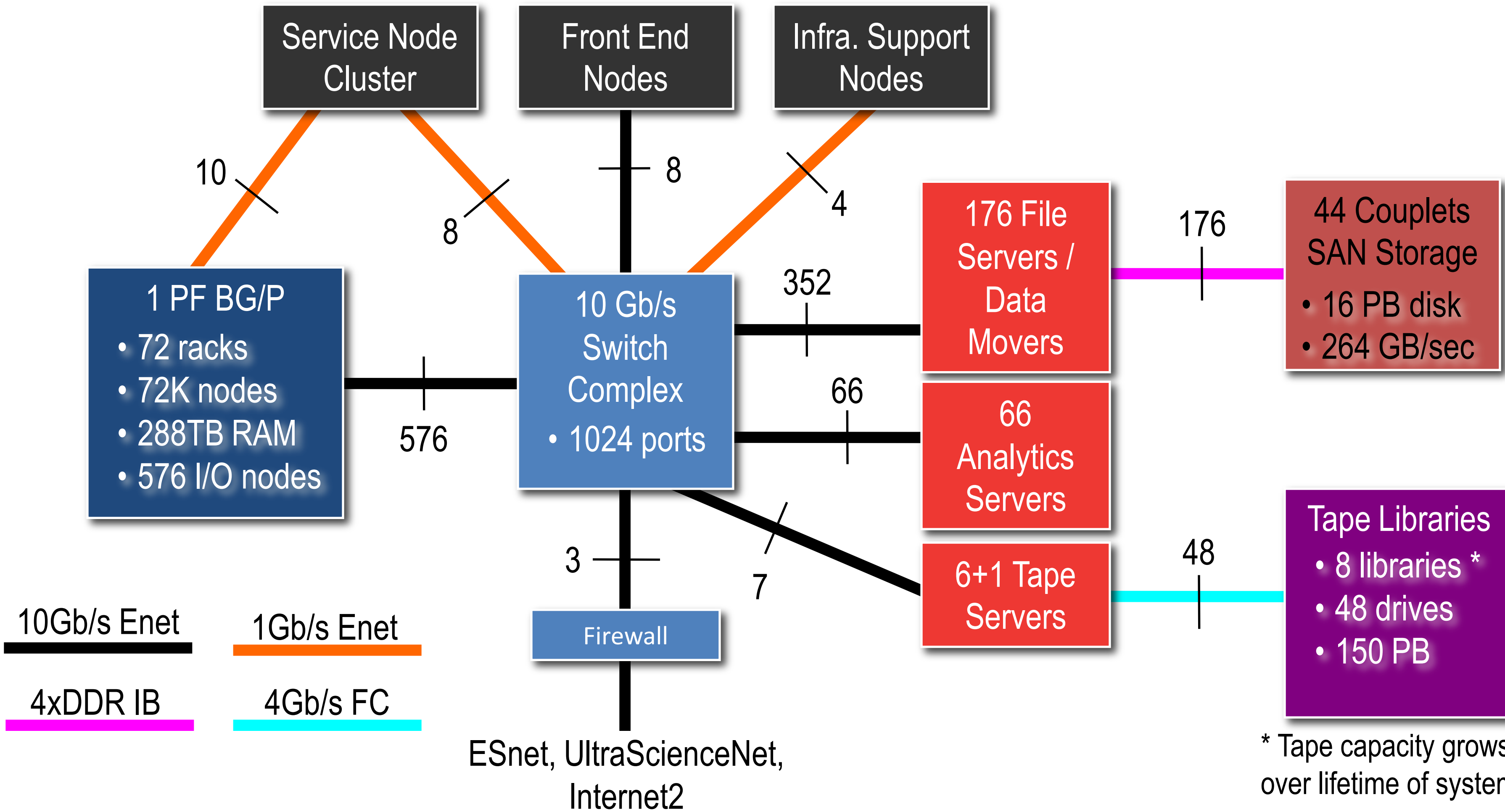
1. Molecular dynamics (mat)
2. Electronic structure
3. Reactor analysis/CFD
4. Fuel design (mat)
5. Reprocessing (chm)
6. Repository optimizations
7. Molecular dynamics (bio)
8. Genome analysis
9. QMC
10. QCD
11. Astrophysics

Blue Gene Advantage

	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**

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!

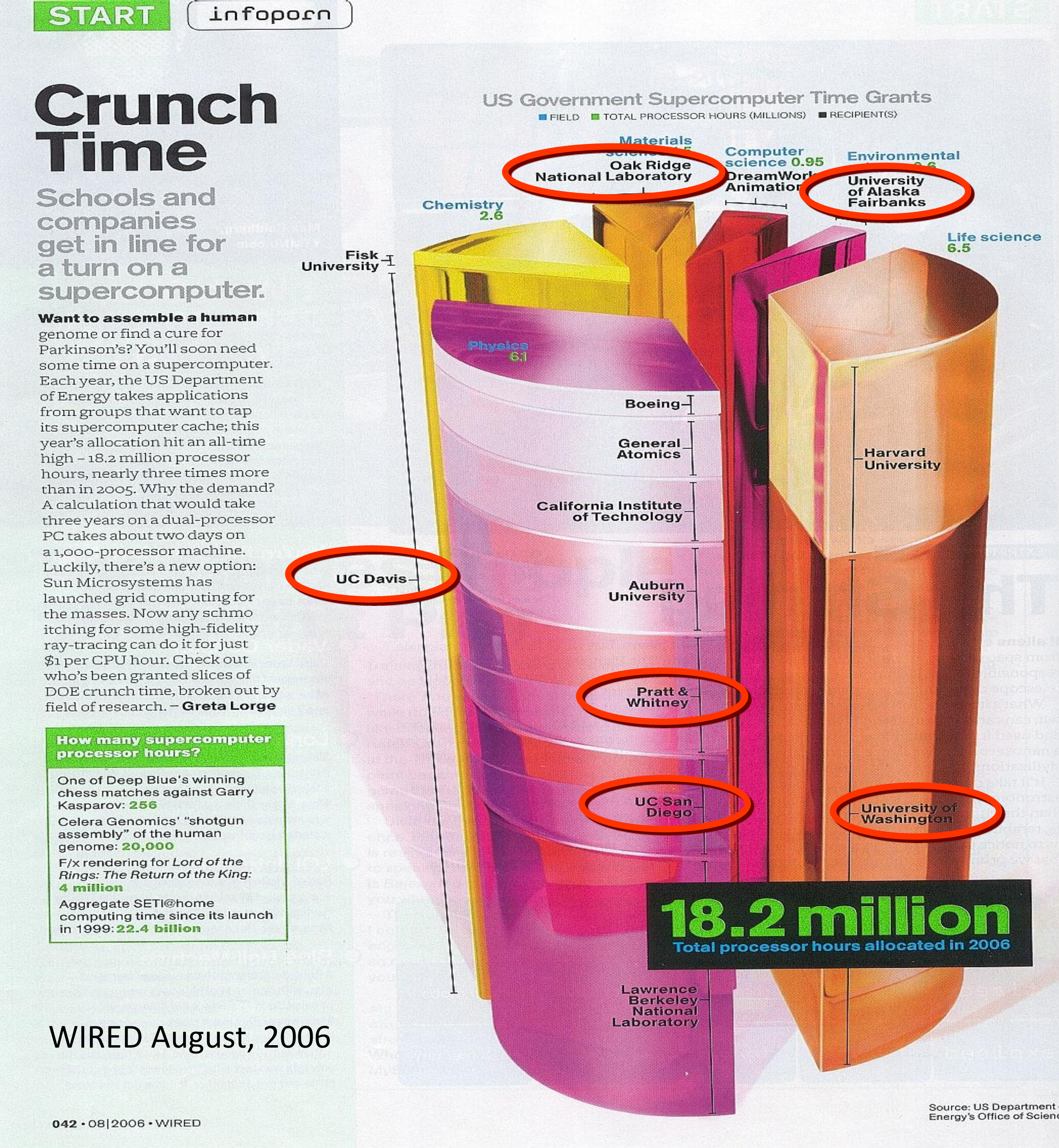
U.S. Department of Energy
Office of Science

Since 2004

DOE INCITE Program
Innovative and Novel Computational Impact on Theory and Experiment

- **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**

INCITE Awards in 2006



START | infoporn

Crunch Time

Schools and companies get in line for a turn on a supercomputer.

Want to assemble a human genome or find a cure for Parkinson's? You'll soon need some time on a supercomputer. Each year, the US Department of Energy takes applications from groups that want to tap its supercomputer cache; this year's allocation hit an all-time high – 18.2 million processor hours, nearly three times more than in 2005. Why the demand? A calculation that would take three years on a dual-processor PC takes about two days on a 1,000-processor machine. Luckily, there's a new option: Sun Microsystems has launched grid computing for the masses. Now any schmo itching for some high-fidelity ray-tracing can do it for just $1 per CPU hour. Check out who's been granted slices of DOE crunch time, broken out by field of research. – **Greta Lorge**

How many supercomputer processor hours?

- One of Deep Blue's winning chess matches against Garry Kasparov: **256**
- Celera Genomics' "shotgun assembly" of the human genome: **20,000**
- F/x rendering for *Lord of the Rings: The Return of the King*: **4 million**
- Aggregate SETI@home computing time since its launch in 1999: **22.4 billion**

WIRED August, 2006

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.

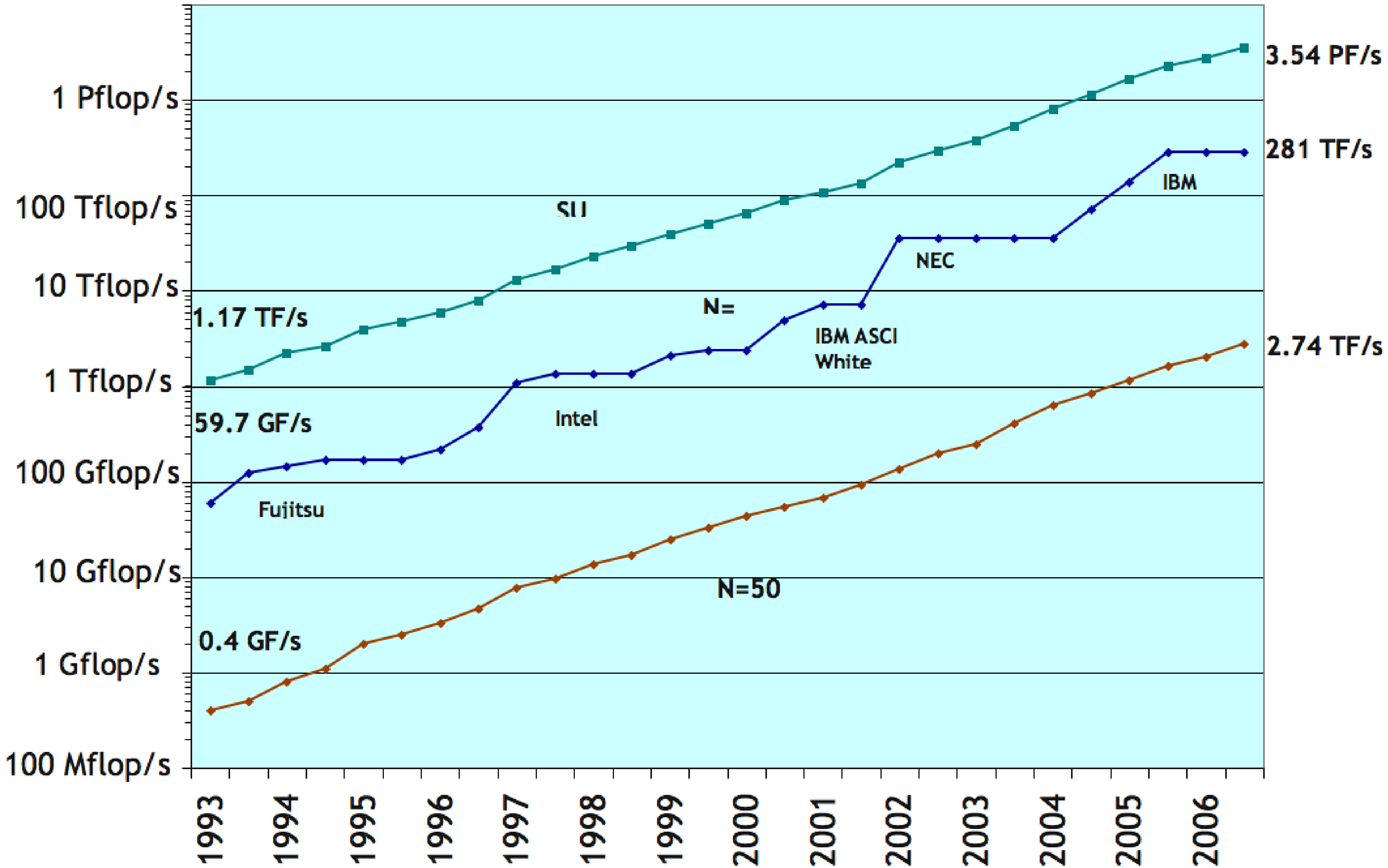
Argonne Theory and Computing Sciences Building

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

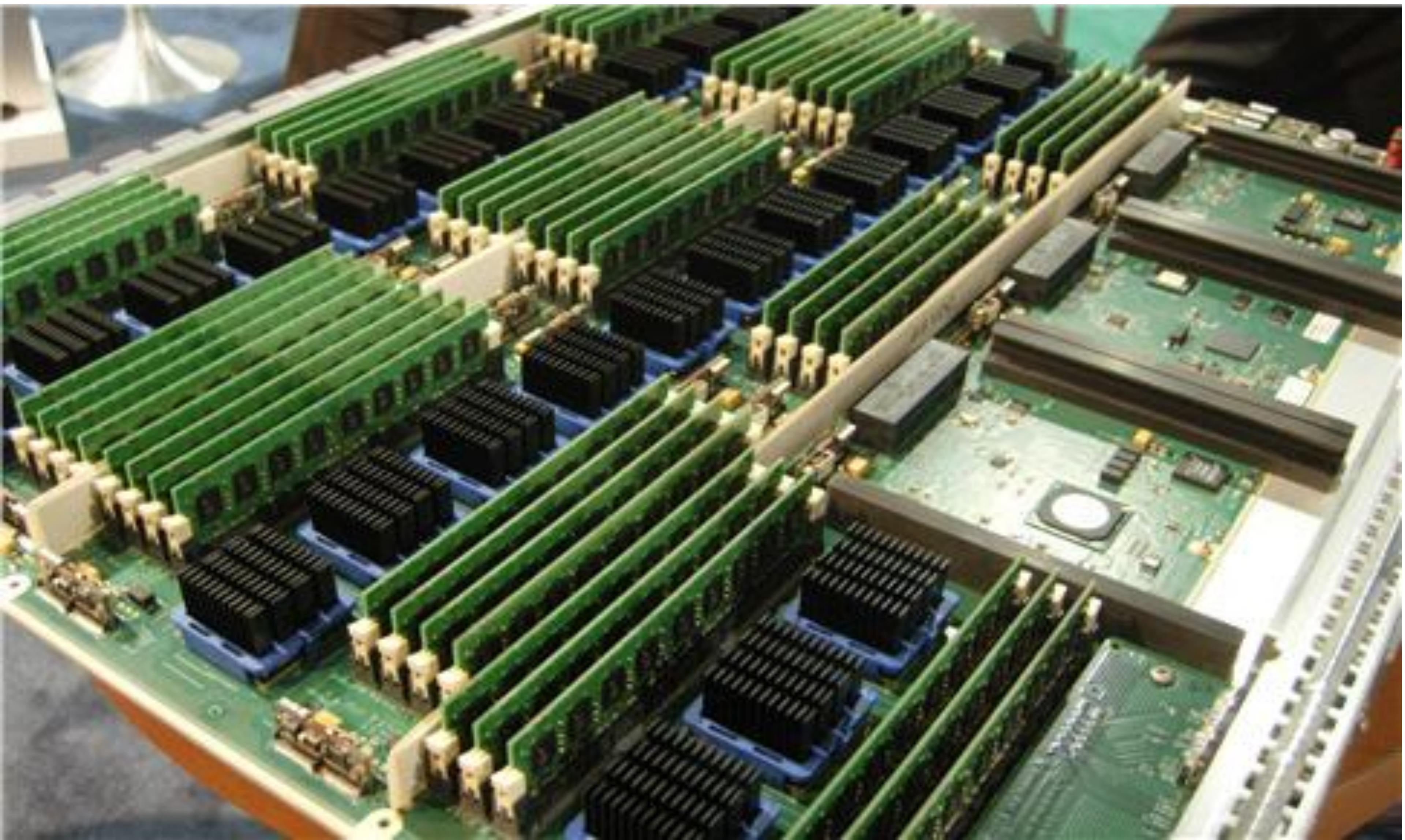
Supercomputing& Cloud Computing

- Two macro architectures dominate large-scale (intentional) computing infrastructures (vs embedded & ad hoc)
- 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

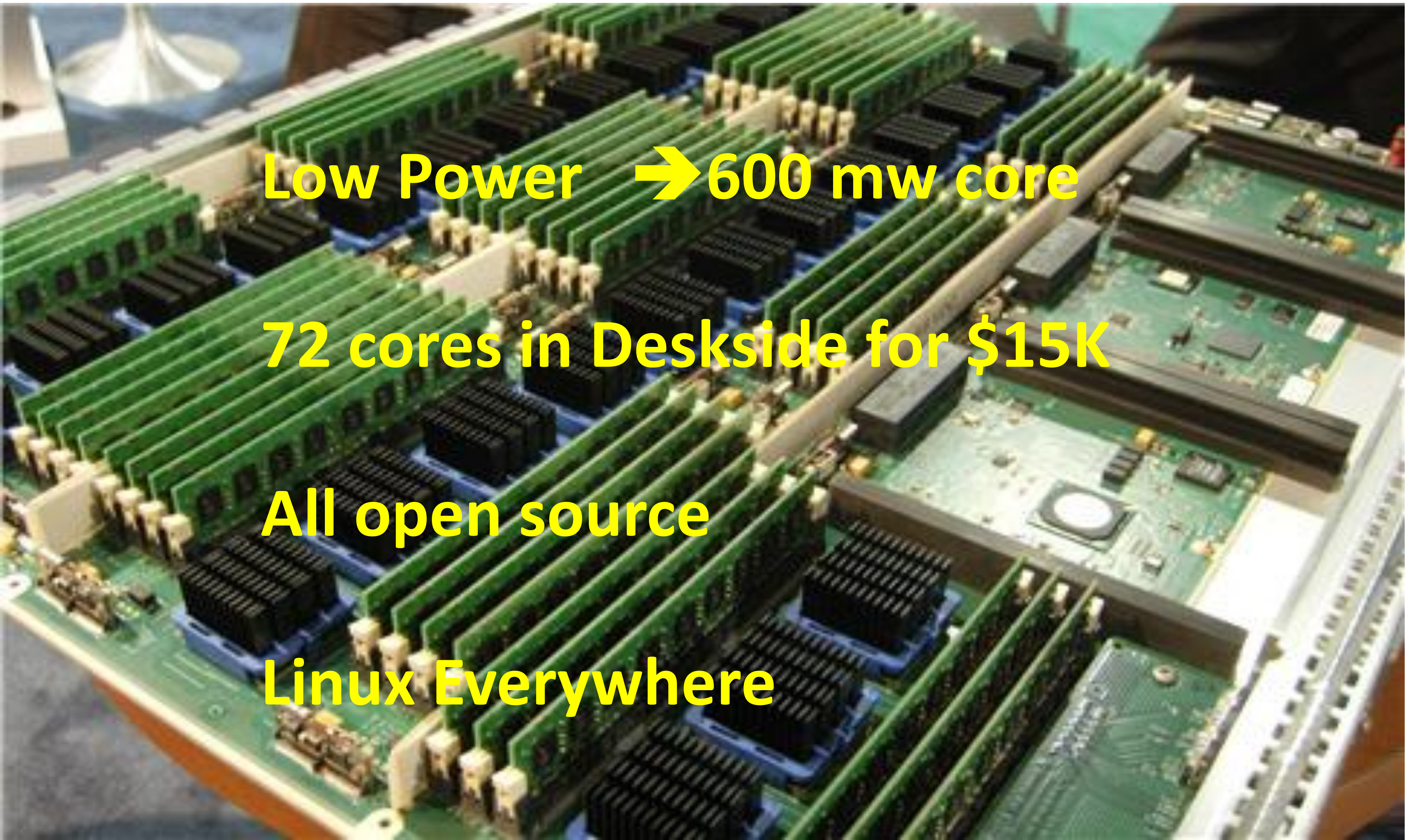
Top 500 Trends



SiCortex Node Board



SiCortex Node Board



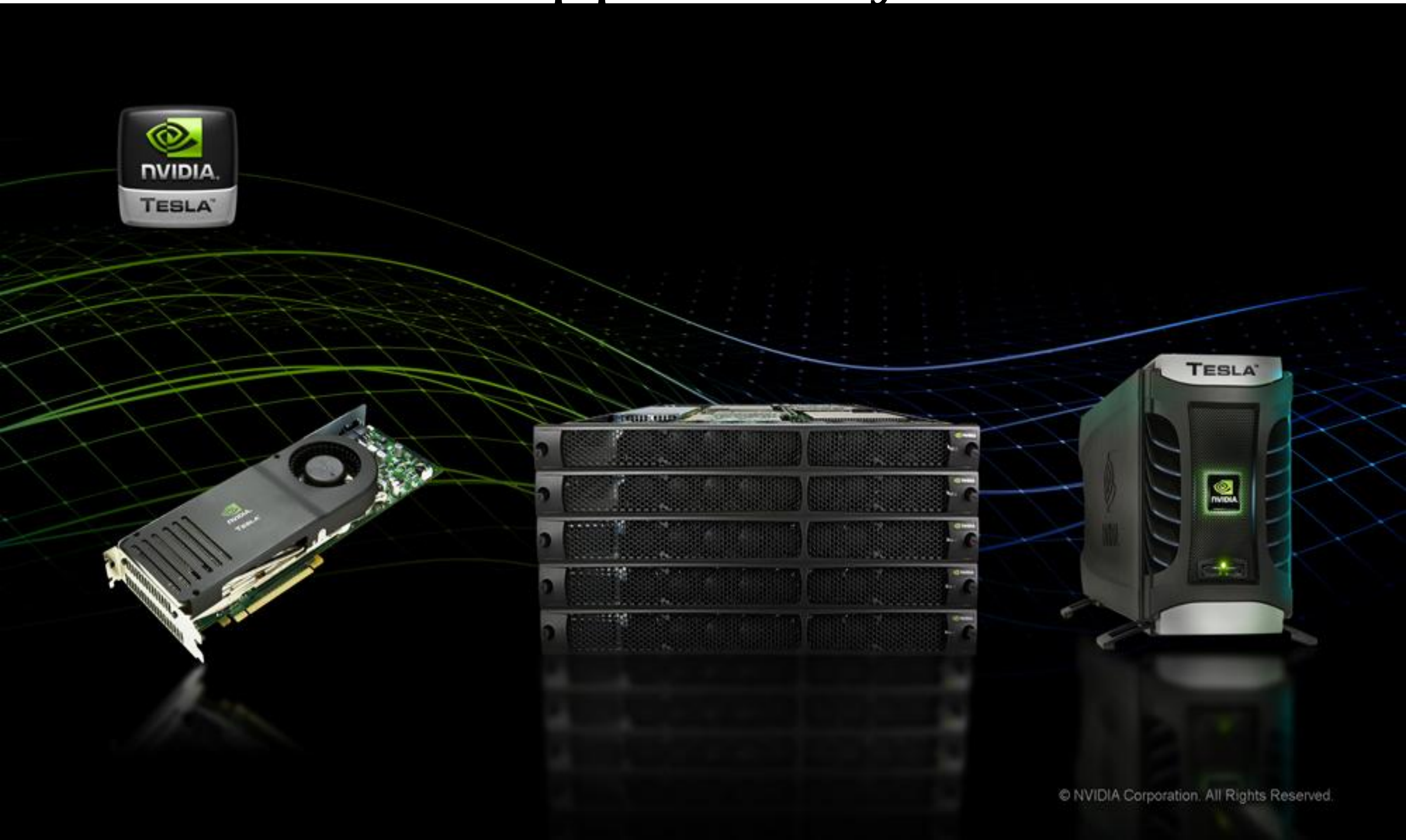
Low Power → 600 mw core

72 cores in Deskside for $15K

All open source

Linux Everywhere

The NVIDIA Challenge and Opportunity



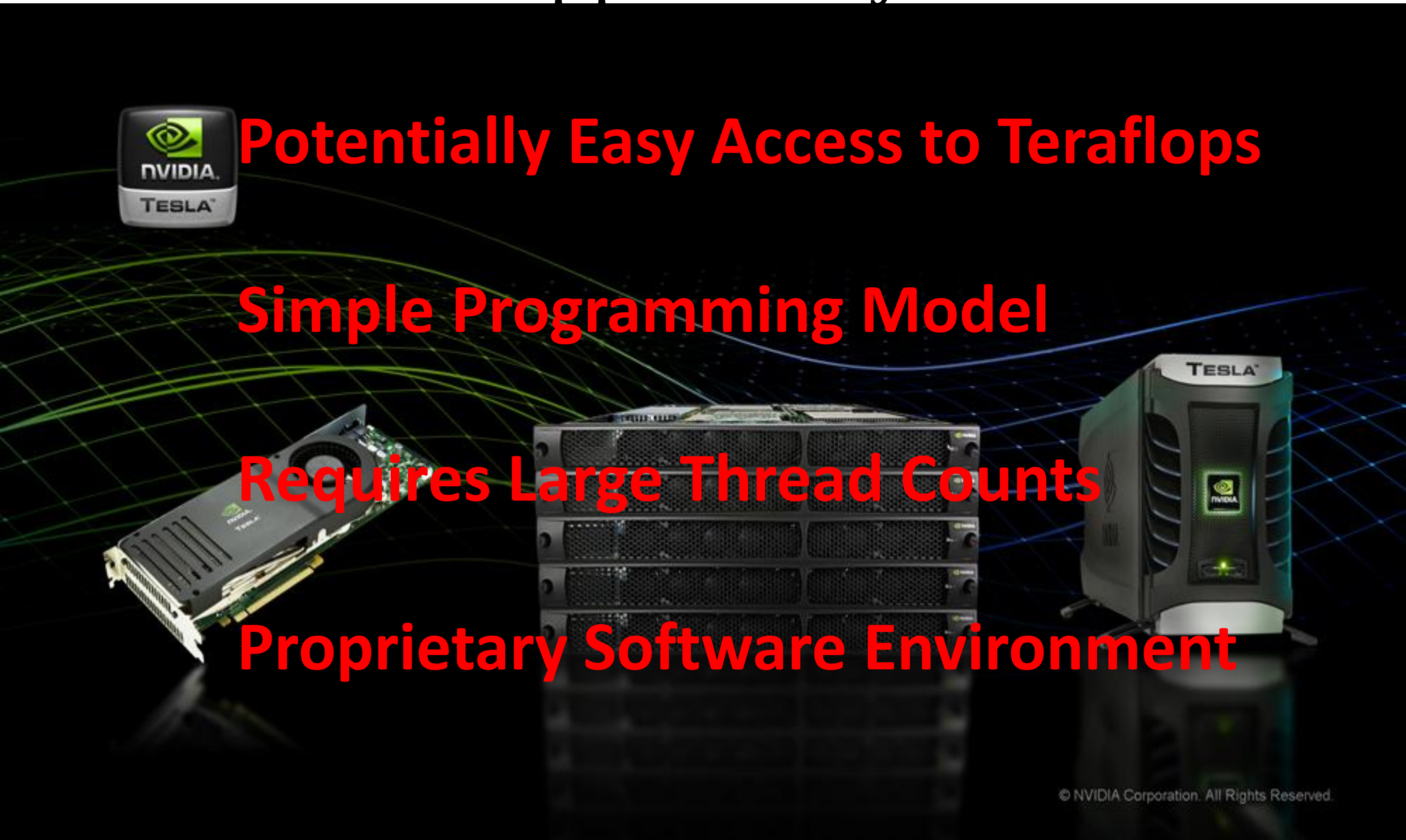
The NVIDIA Challenge and Opportunity

Potentially Easy Access to Teraflops

Simple Programming Model

Requires Large Thread Counts

Proprietary Software Environment



Blue Gene L Node Cards

Blue Gene Node Cards

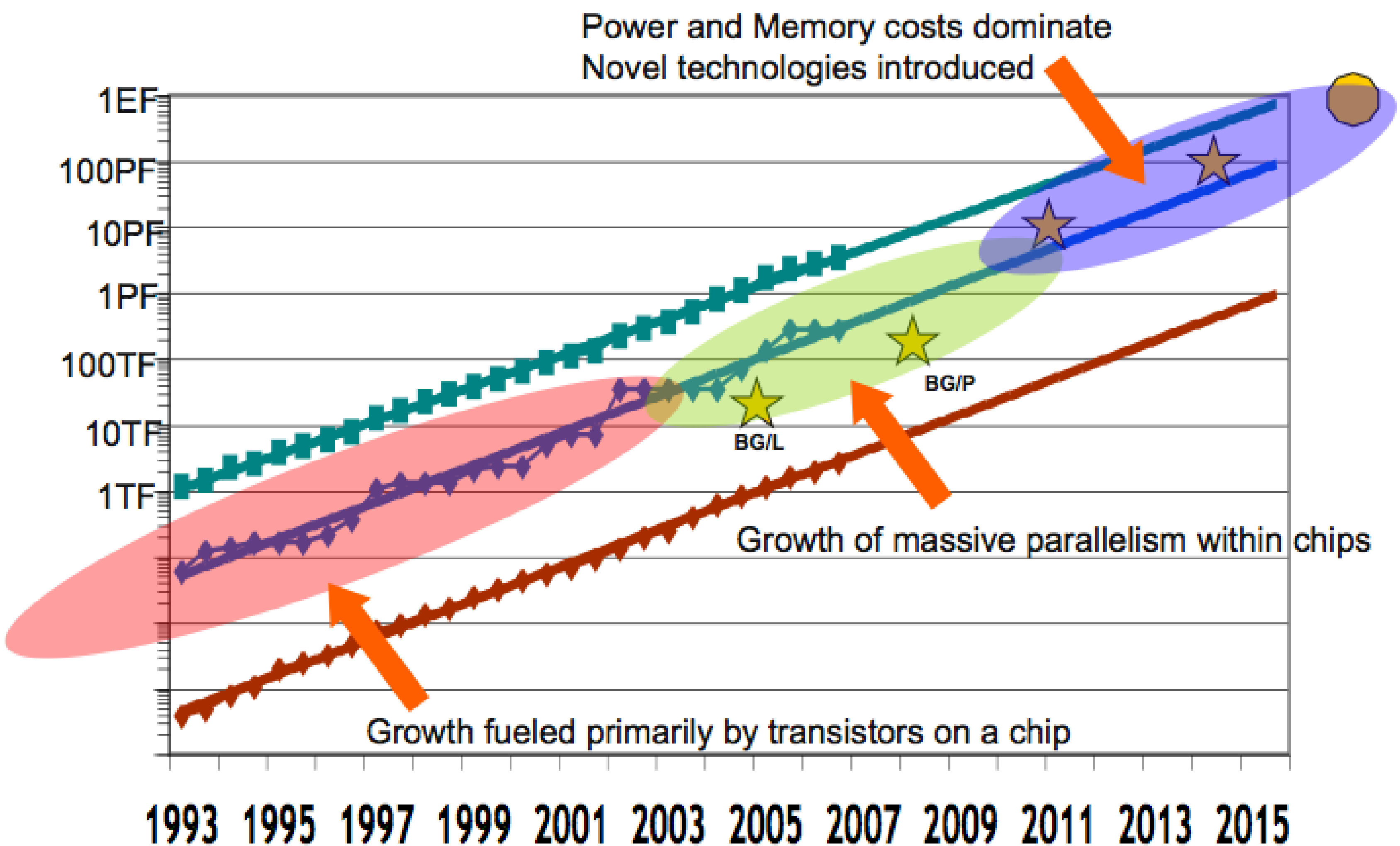
Fine Grain and Low Power

Existing Programming Model

Extremely Scalable

Mostly Open Software Environment

Looking to Exascale



A Three Step Path to Exascale

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

Failure Rates and Reliability of Large Systems

Table 2 Uncorrectable hard failure rates of the Blue Gene/L by component.

Component	*FIT per component*†	*Components per 64Ki compute node partition*	*FITs per system (K)*	*Failure rate per week*
Control–FPGA complex	160	3,024	484	0.08
DRAM	5	608,256	3,041	0.51
Compute + I/O ASIC	20	66,560	1,331	0.22
Link ASIC	25	3,072	77	0.012
Clock chip	6.5	~1,200	8	0.0013
Nonredundant power supply	500	384	384	0.064
Total (65,536 compute nodes)			5,315	0.89

†T = 60°C, V = Nominal, 40K POH. *FIT* = Failures in ppm/KPOH. One FIT = 0.168×16^{-6} fails per week if the machine runs 24 hours a day.

Theory

Experiment

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
- ClearspeedsCn
- 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 x 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

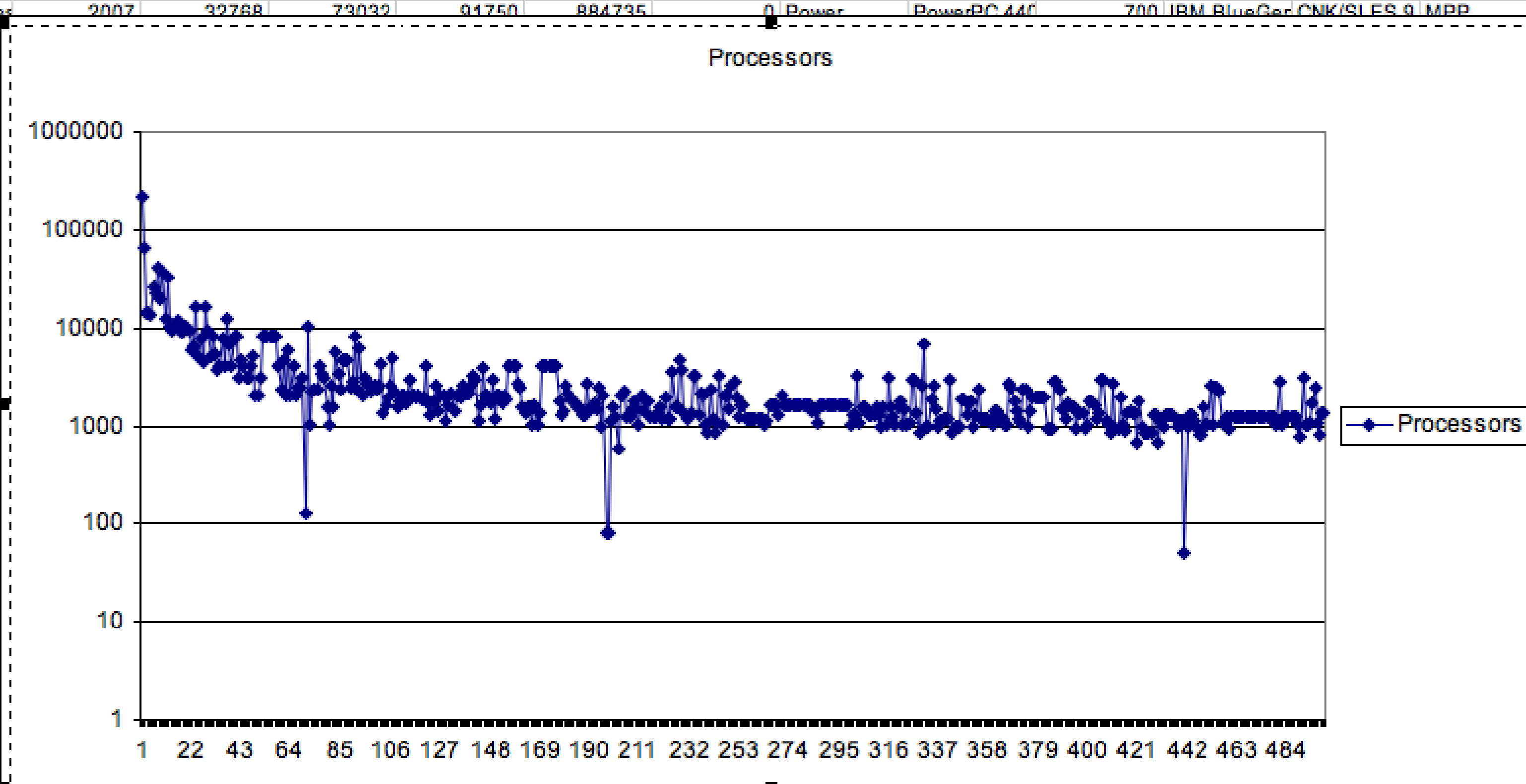
- The early adopters almost always self identify
- Approximately 1/3 of the petascale codes didn’t exist 10 years ago
- Most of them did exist but required considerable investment, new implementation and tuning
- The simplest path forward (pure MPI) was the path of least resistance for most code groups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

Existing Body of Parallel Software

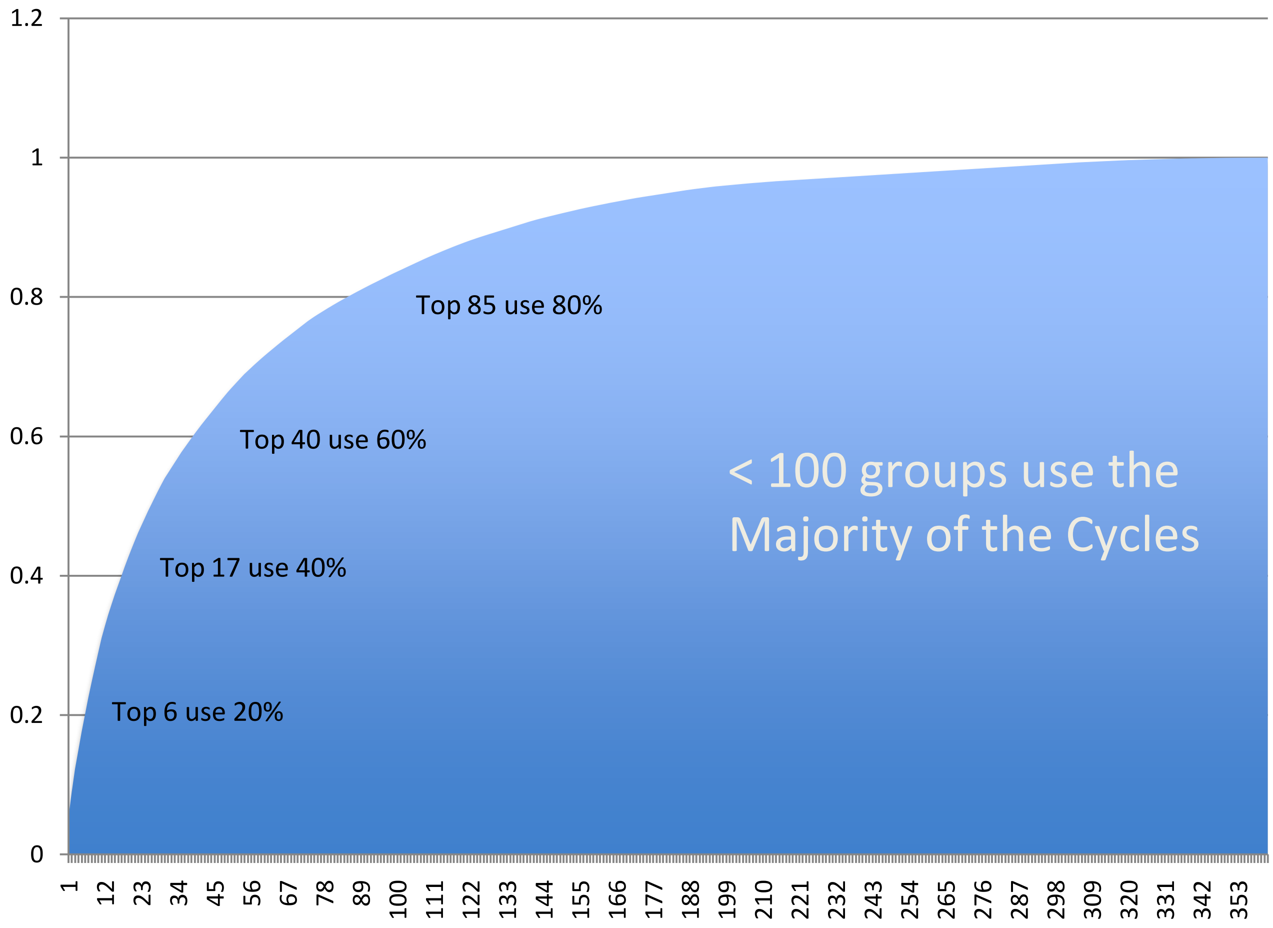
- How many existing HPC science and engineering codes scale beyond 1000 processors?
  - My estimate is that it is less than 1000 world wide
  - Top users at NERSC, OLCF and ALCF < 200 groups
  - It appears likely that the bulk of cycles on Top500 are used in capacity mode with the exception of a sites 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

Platfo	Sweden	2007	13728	102800	146430	375000	0	Intel EM64T	Intel EM64T	2667	HP Cluster Pl	Linux	Cluster	Government	Defens
Cray	United States	2007	26569	102200	127531	1700000	0	AMD x86_64	AMD x86_64	2400	Cray XT	UNICOS/SUS	MPP	Research	Not Sp
4/XT	United States	2006	23016	101700	119350	2220160	0	AMD x86_64	AMD x86_64	2600	Cray XT	UNICOS/SUS	MPP	Research	Not Sp
Blue	United States	2005	40960	91290	114688	983039	0	Power	PowerPC 440	700	IBM BlueGen	CNK/SLES 9	MPP	Research	Informa
4, 2.6	United States	2007	19320	85368	100464	0	0	AMD x86_64	AMD x86_64	2600	Cray XT	UNICOS/Linu	MPP	Research	Not Sp
Blue	United States	2007	36864	82161	103219	0	0	Power	PowerPC 440	700	IBM BlueGen	CNK/SLES 9	MPP	Research	Resear
pSer	United States	2006	12208	75760	92781	1383600	0	Power	POWER5	1900	IBM pSeries	AIX	MPP	Research	Resear
Blue	United States	2007	32768	73032	91750	884735	0	Power	PowerPC 440	700	IBM BlueGen	CNK/SLES 9	MPP	Academic	Resear
nter	Spain													Academic	Medici
dge 1	United State													Academic	Not Sp
00 1.6	Germany													Academic	Not Sp
x460	Japan													Academic	Not Sp
4, 2.6	United King													Academic	Not Sp
dge 1	United State													Research	Resear
ale 5	France													Research	Defens
x 1.5	United State													Research	Aerosp
00 1.6	United State													Research	Not Sp
dge 1	United State													Academic	Not Sp
Platfo	Sweden													Academic	Resear
ne/P	United State													Vendor	Bench
dge 1	United State													Research	Resear
le 30	France													Research	Not Sp
y II (U	United State													Research	Resear
Blue	Germany													Research	Resear
treme	United State													Research	Not Sp
mula	Japan													Research	Enviro
3, 2.6	United State													Research	Not Sp
dge 1	United State													Academic	Resear
Platfo	Russia													Academic	Not Sp
x ICE	United State													Research	Not Sp
3, 2.6	United King													Classified	Defens
dge 1	United State													Academic	Resear
Blue	Netherlands													Academic	Resear
treme	United State													Research	Not Sp
dge 1	United State													Academic	Not Sp
ne/P	Germany													Research	Not Sp
ne/P	United State													Research	Resear
nter	United State													Academic	Not Sp
dge 9	United State													Academic	Not Sp
Platfo	United State													Government	Defens
pSer	United State													Research	Weathe
pSer	United States	2007	3072	20070	23347	700000	0	Power	POWER5	1900	IBM pSeries	AIX	MPP	Vendor	Bench
nium2	United States	2004	4096	19940	22938	975000	110000	Intel IA-64	Intel IA-64 Ita	1400	NOW - Intel It	Linux	Cluster	Research	Resear
nter	Italy	2007	5120	19910	61440	0	0	Intel EM64T	Intel EM64T	3000	IBM Cluster	Linux	Cluster	Academic	Not Sp
x3550	Taiwan	2007	2080	19910	24960	0	0	Intel EM64T	Intel EM64T	3000	IBM Cluster	Linux	Cluster	Academic	Not Sp
x ICE	United States	2007	2048	19360	24576	0	0	Intel EM64T	Intel EM64T	3000	SGI Altix	SLES10 + SG	MPP	Vendor	Bench
nter	United Kingd	2007	3040	19039.5	27968	0	0	Power	PowerPC 970	2300	IBM Cluster	Linux	Cluster	Academic	Resear
Blue	Japan	2005	8192	18665	22937.6	442367	0	Power	PowerPC 440	700	IBM BlueGen	CNK/SLES 9	MPP	Research	Biology
Blue	France	2007	8192	18665	22937.6	442367	0	Power	PowerPC 440	700	IBM BlueGen	CNK/SLES 9	MPP	Industry	Not Sp
Blue	Switzerland	2005	8192	18665	22937.6	442367	0	Power	PowerPC 440	700	IBM BlueGen	CNK/SLES 9	MPP	Academic	Resear



NERSC 2007 Rank Abundance

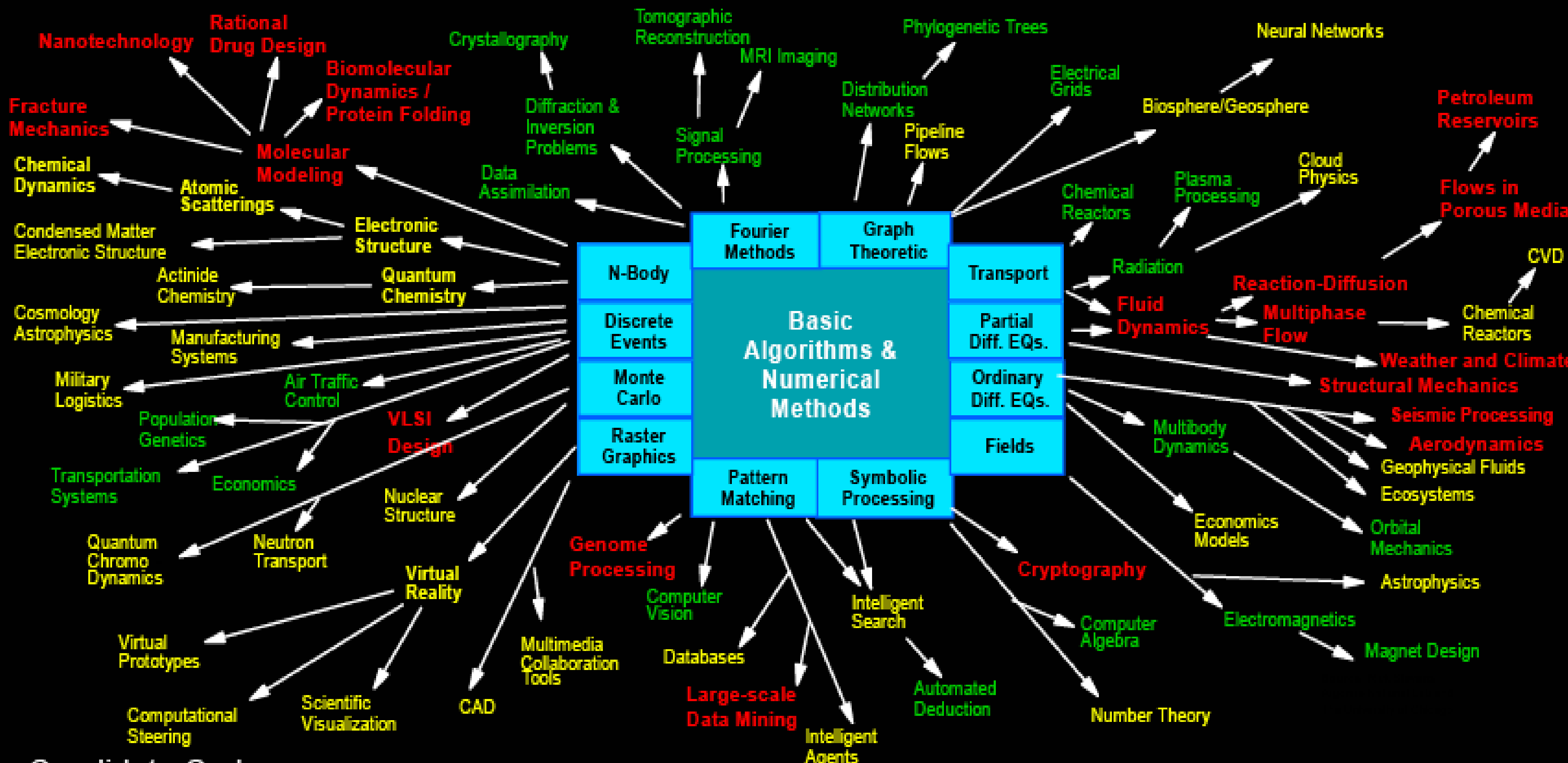


Existing Applications of Interest

- 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

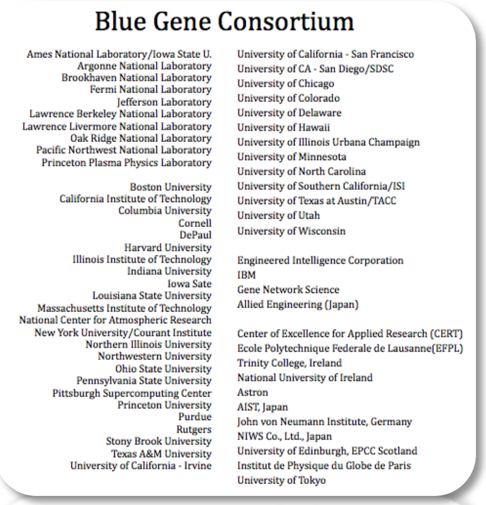


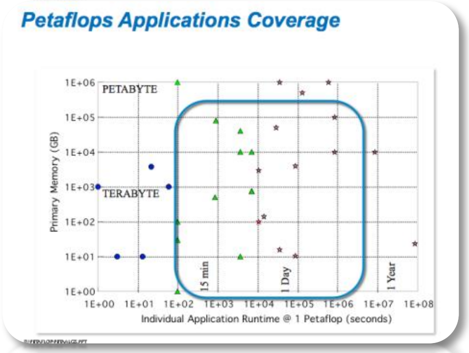
- Candidate Codes:
  - Inherently parallel; written using MPI
  - Memory required per MPI task is less than that available
  - Dominated by collective communication across all nodes
  - Locality of communications within 3D mapping
- Non-Candidate Codes:
  - Large memory footprints required on individual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

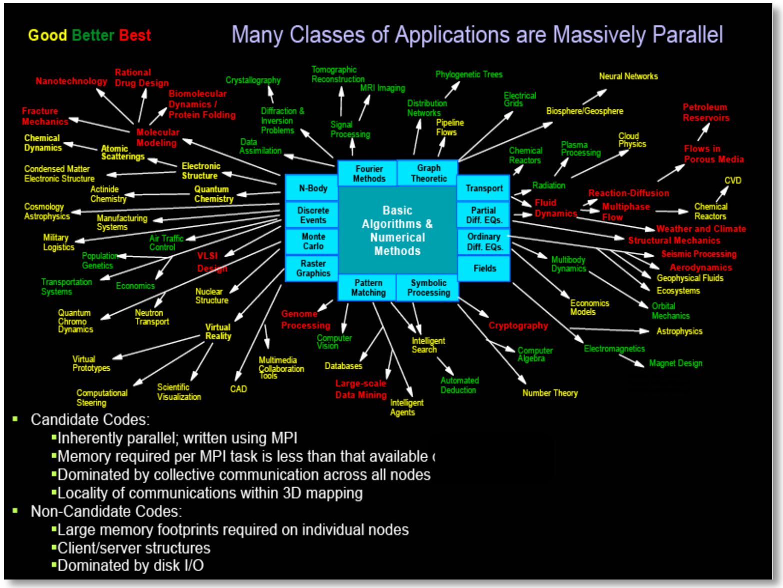
How Quickly Can A New Architecture Be Adopted?

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







Example Applications Ported to BG/L and BG/P

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

Humanity's Top Ten Problems for next 50 years

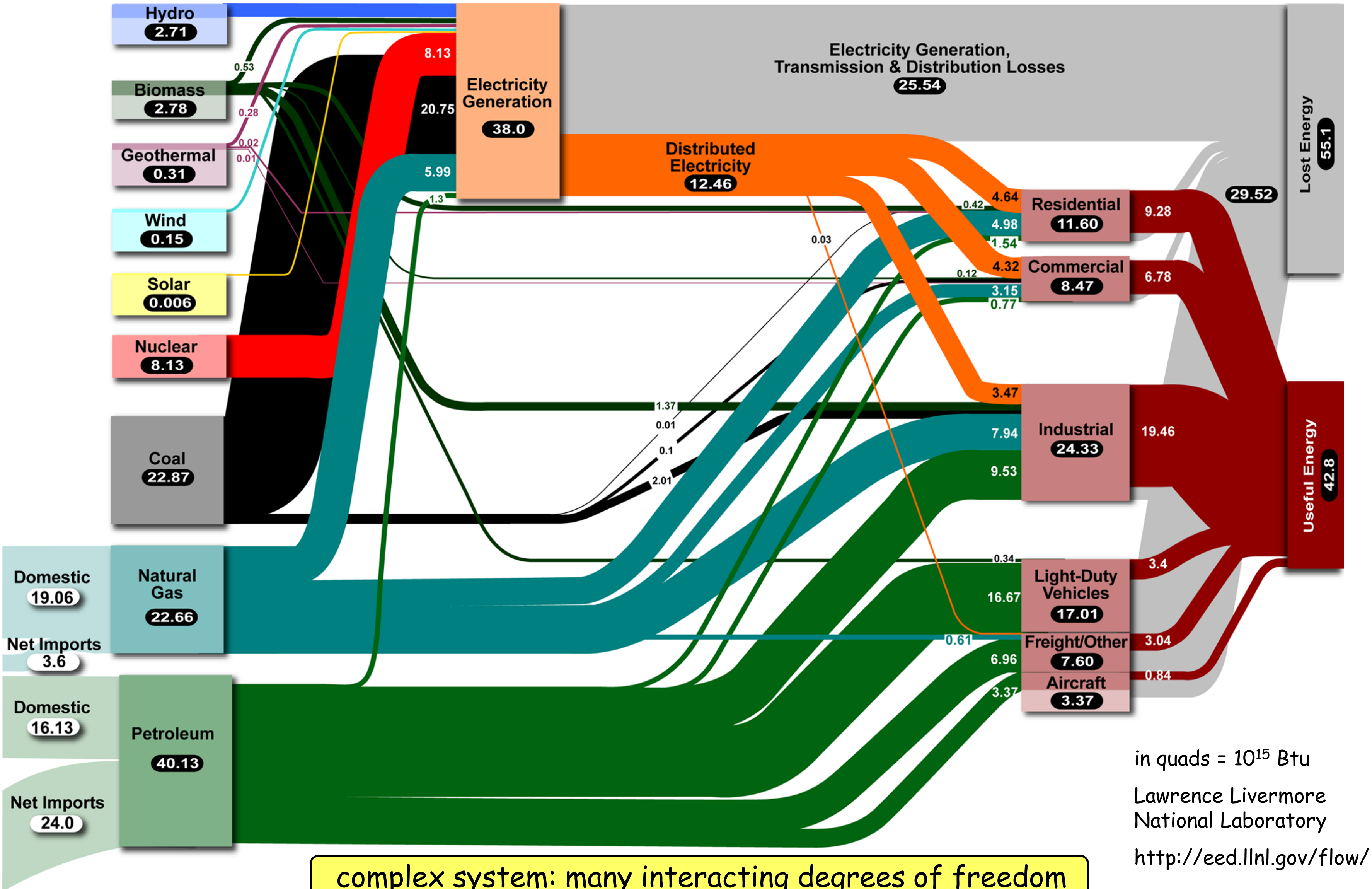
1. **ENERGY**
2. WATER
3. FOOD
4. ENVIRONMENT
5. POVERTY
6. TERRORISM & WAR
7. DISEASE
8. EDUCATION
9. DEMOCRACY
10. POPULATION

2007 7 Billion People
2050 8-10 Billion People

Richard Smalley's Top Ten List

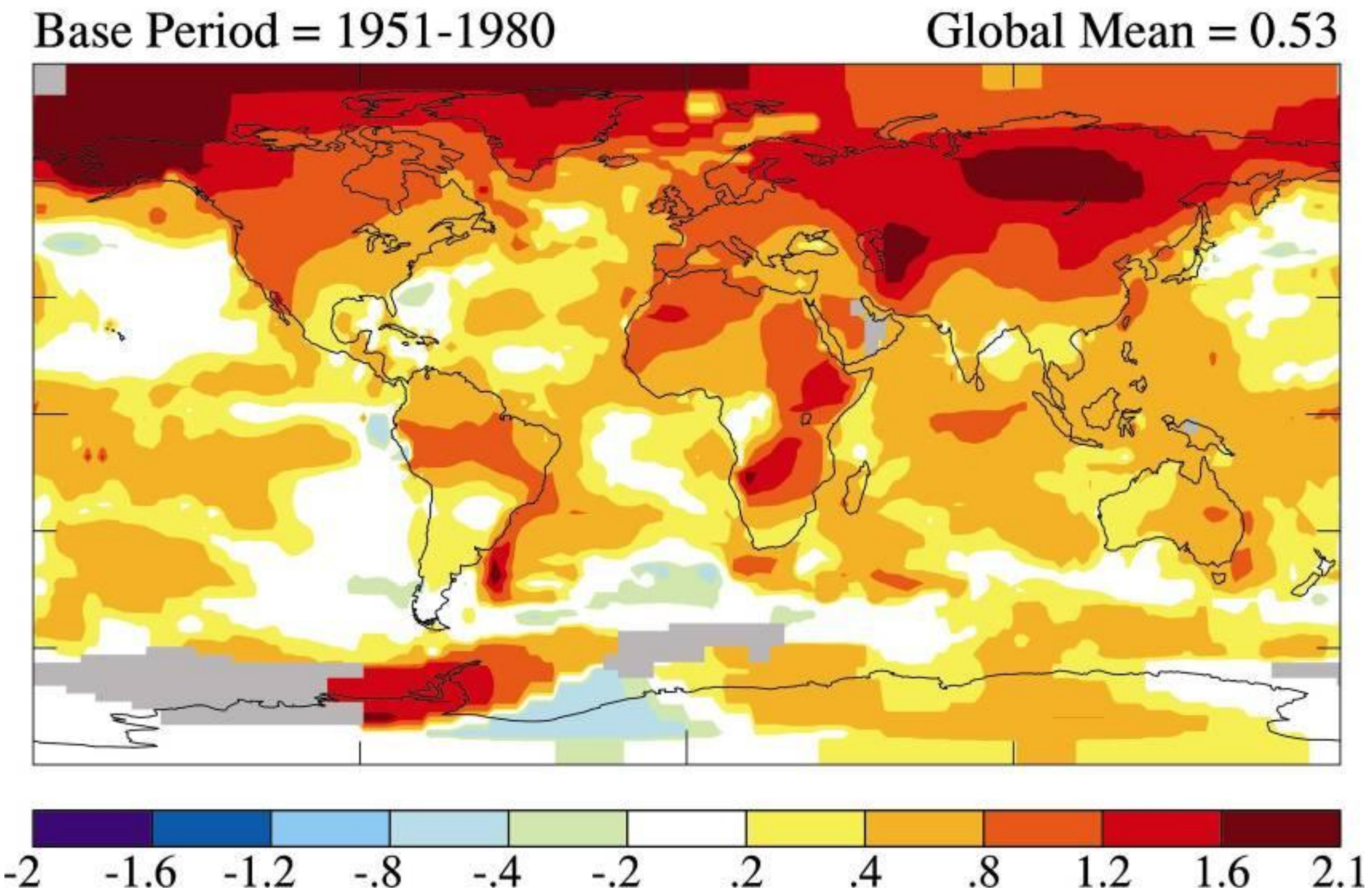
The Grid - the Triumph of 20th Century Engineering

Energy Flows in 2005



complex system: many interacting degrees of freedom

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



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

J. Hansen et al., *PNAS 103*: 14288-293 (26 Sept 2006)

The 21st Century: A Different Set of Challenges

capacity

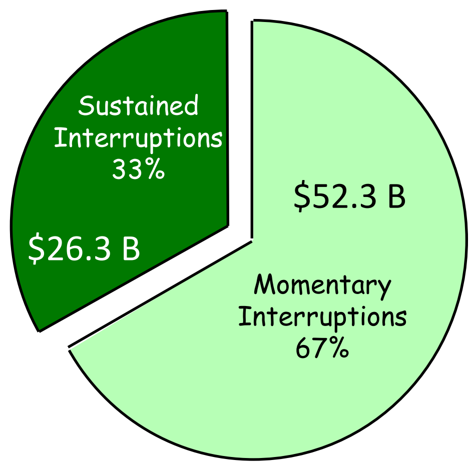
growing electricity uses
growing cities and suburbs
high people / power density
urban power bottleneck

2030
50% demand growth (US)
100% demand growth (world)

reliability
power quality

average
power loss/customer
(min/yr)

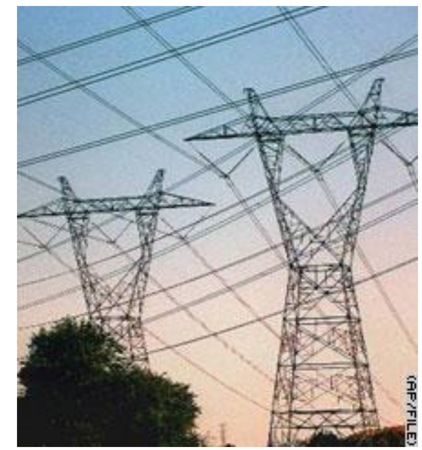
US	214
France	53
Japan	6



$79 B economic loss (US)

LaCommare & Eto, Energy 31, 1845 (2006)

efficiency
lost energy



62% energy lost in production / delivery

8-10% lost in grid

40 GW lost (US)
~ 40 power plants

2030: 60 GW lost (US)
340 Mtons CO_2

The Energy Alternatives

Fossil	Nuclear	Renewable	Fusion
		solar, wind, hydroelectric ocean tides and currents biomass, geothermal	

energy gap
~ 14 TW by 2050
~ 33 TW by 2100

→

10 TW = 10,000 1 GW power plants

1 new power plant/day for 27 years

China: 1 GW / week

no single solution

diversity of energy sources required

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

CO_2 Capture & Storage (CCS)

CO_2 capture:

1. Introduce CCS at 800 GW coal or 1600 GW natural gas plants (1100 GW coal today)
2. Introduce CCS at plants producing 250 MtH_2/yr from coal or 500 MtH_2/yr from natural gas (40 MtH_2/yr today)
3. Introduce CCS at synfuels plants producing 30 mbd from coal (200x Sasol)

H_2 safety, infrastructure

Geologic storage:
Create 3500 Sleipners

Durable storage

Wind and Solar

4. Add 2 million 1-MW-peak windmills (50 times current level) on 30 Mha, displacing coal electricity
5. Add 2000 GW-peak PV plants (700 times current level) on 2 Mha, displacing coal electricity

6. Add 4 million windmills or 4000 GW-peak PV plants generating 200 MtH_2/yr, displacing gasoline hybrid cars

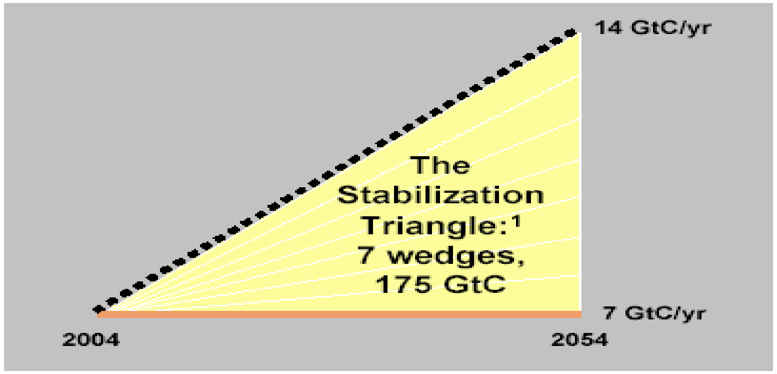
Energy storage, H_2 safety, infrastructure
PV production cost

Energy Efficiency & Conservation

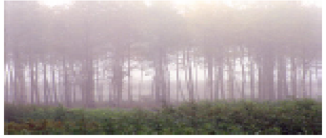
Increase economy-wide emissions/GDP reduction by additional 0.15%/yr (e.g. increase US goal of 1.96%/yr to 2.11%/yr):

8. Increase fuel economy for 2 billion cars from 30 to 60 mpg
9. Decrease car travel for 2 billion 30-mpg cars from 10,000 to 5,000 miles/yr
10. Cut carbon emissions in buildings/appliances by 1/4 over 2054 projection
11. Produce twice today's coal output at 60% efficiency instead of 40% (compare with 32% today)

Weak incentives, urban design, lifestyle changes



Forests & Soils



14. Decrease tropical deforestation to zero instead of 0.5 GtC/yr, and establish 300 Mha of new tree plantations (twice current level)
15. Implement conservation tillage on all cropland (10 times current level)

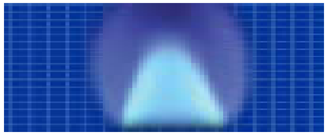
Competing land use, verification, reversibility

Nuclear Fission

13. Add 700 GW nuclear plants (compare with 350 GW today) replacing coal

Nuclear proliferation, terrorism, waste

Fuel Switching



12. Replace 1400 GW coal plants with natural gas plants (compare with 350 GW natural gas today)

Competing demands for natural gas

Biomass Fuel

7. Add 50 times current US and Brazil ethanol production on 250 Mha (1/6 world cropland)

Biodiversity, competing land use

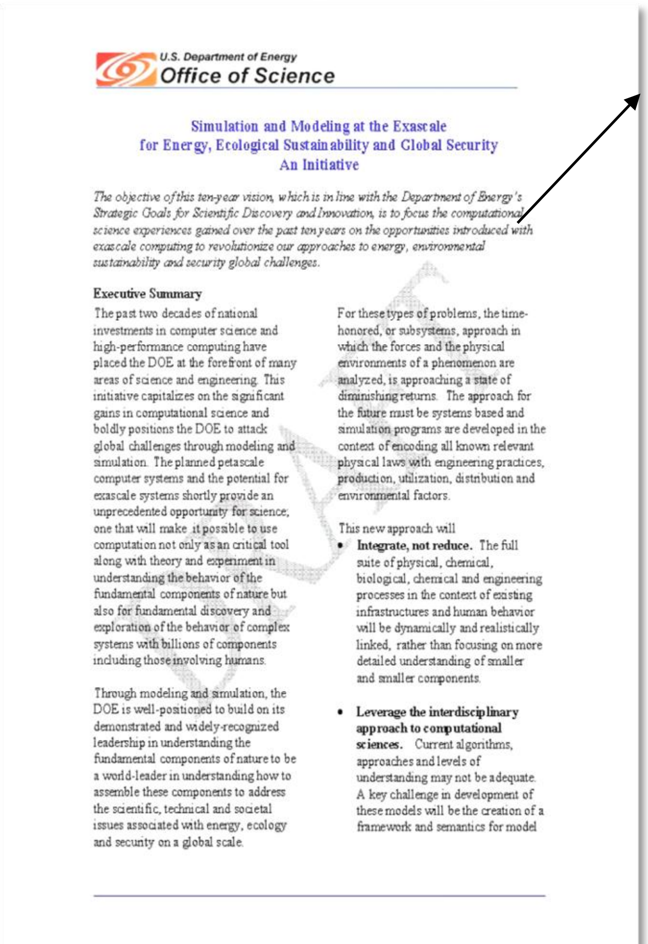
Beyond 2054

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 **geoengineering 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, 968 (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*, 298, 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 an 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.

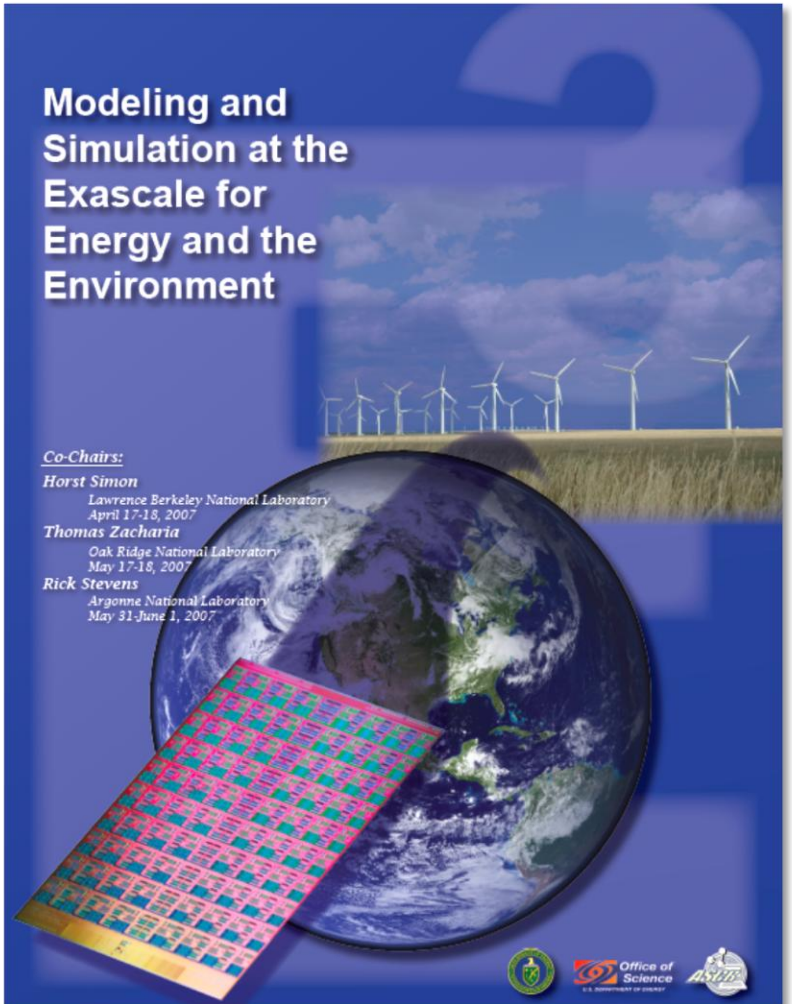
This new approach will

- **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.

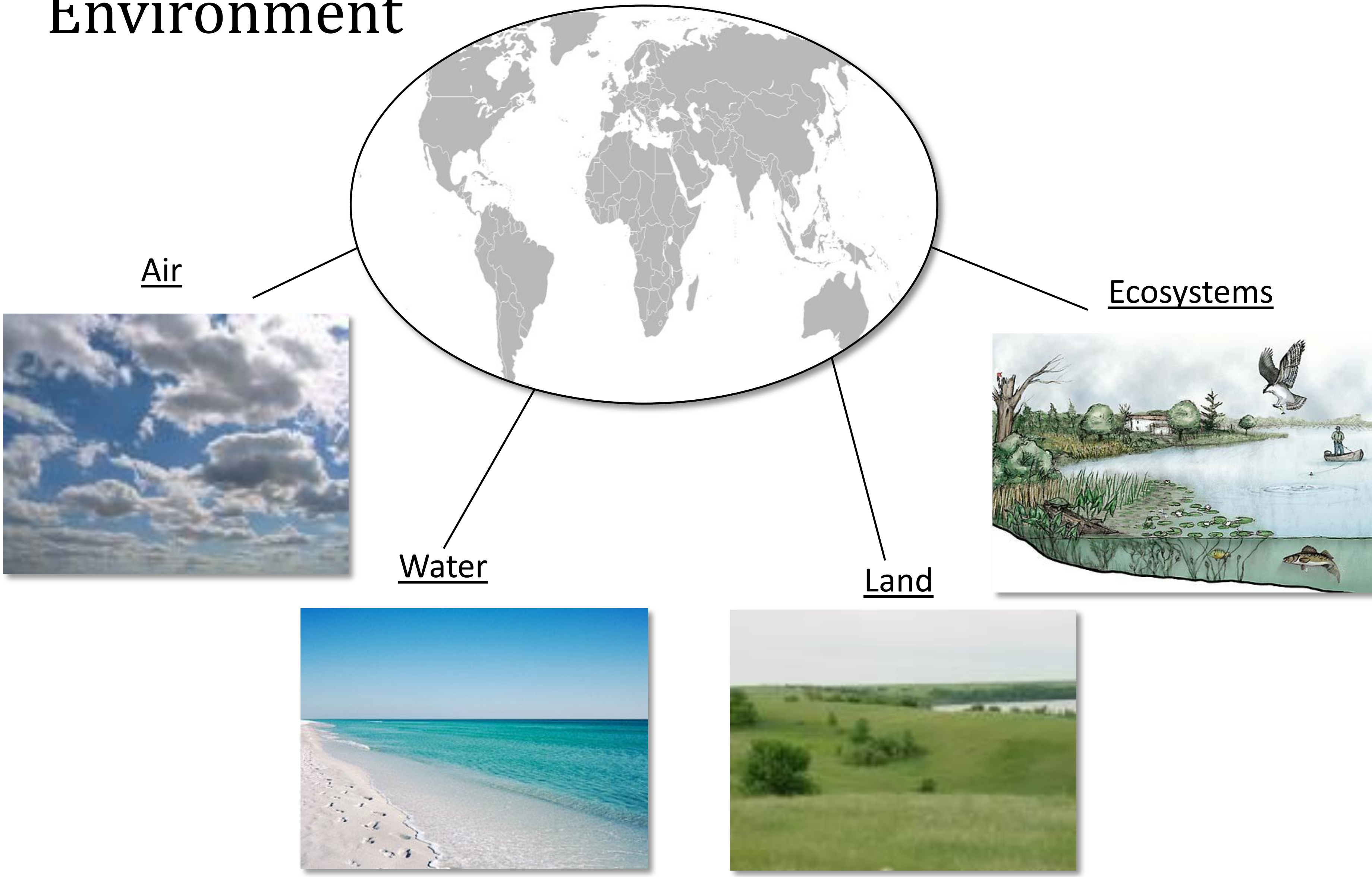
Planning for the Exascale Future!



During the spring of 2007 Argonne, Berkeley and Oak Ridge held three Townhallmeetings 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

The Economic Systems Sit Within the Physical Environment



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

Petascale Geoscience

Geoscience Applications

Discipine	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 curculation 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 microphyics
    - 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	0.00002
LADHS/regional hydrology	100m Columbia river basin	10	0.3	20.8	0.66

Src: Petascale Collaboratory for the Geosciences, 2005

WRF

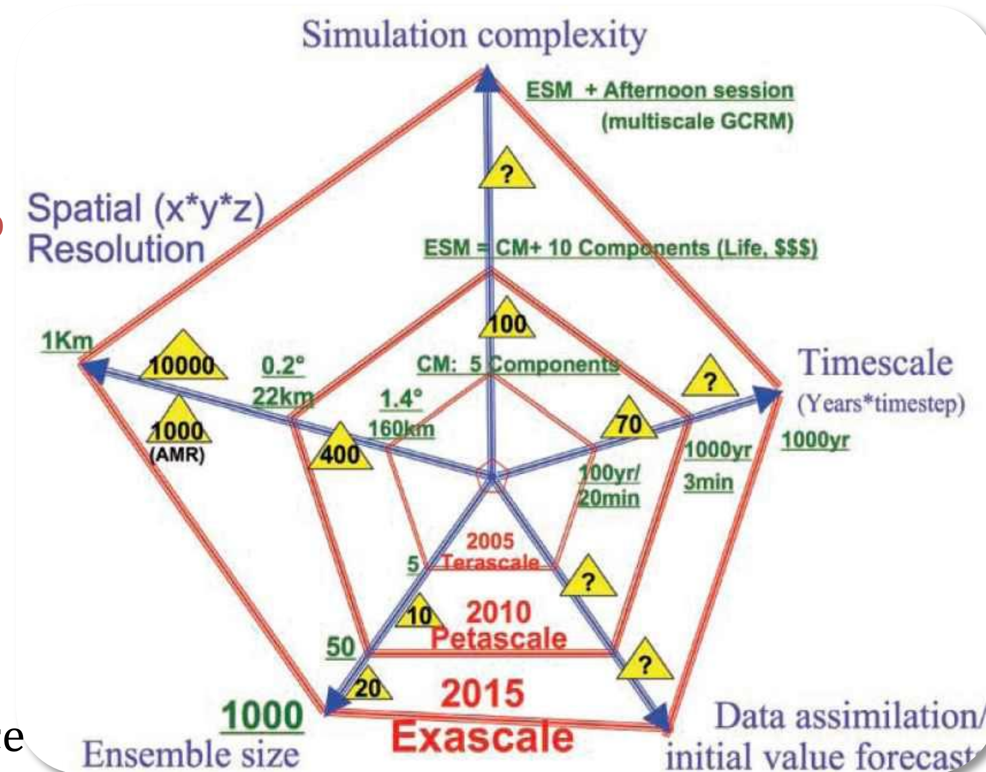
- High resolution global models
  - Below 5 km, scales and physics change
- Global non-hydrostatic numerical weather model
  - 2 km resolution requires ~200 TF sustained
  - 1 km requires 1.6 PF sustained
- Major research problem just getting started

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

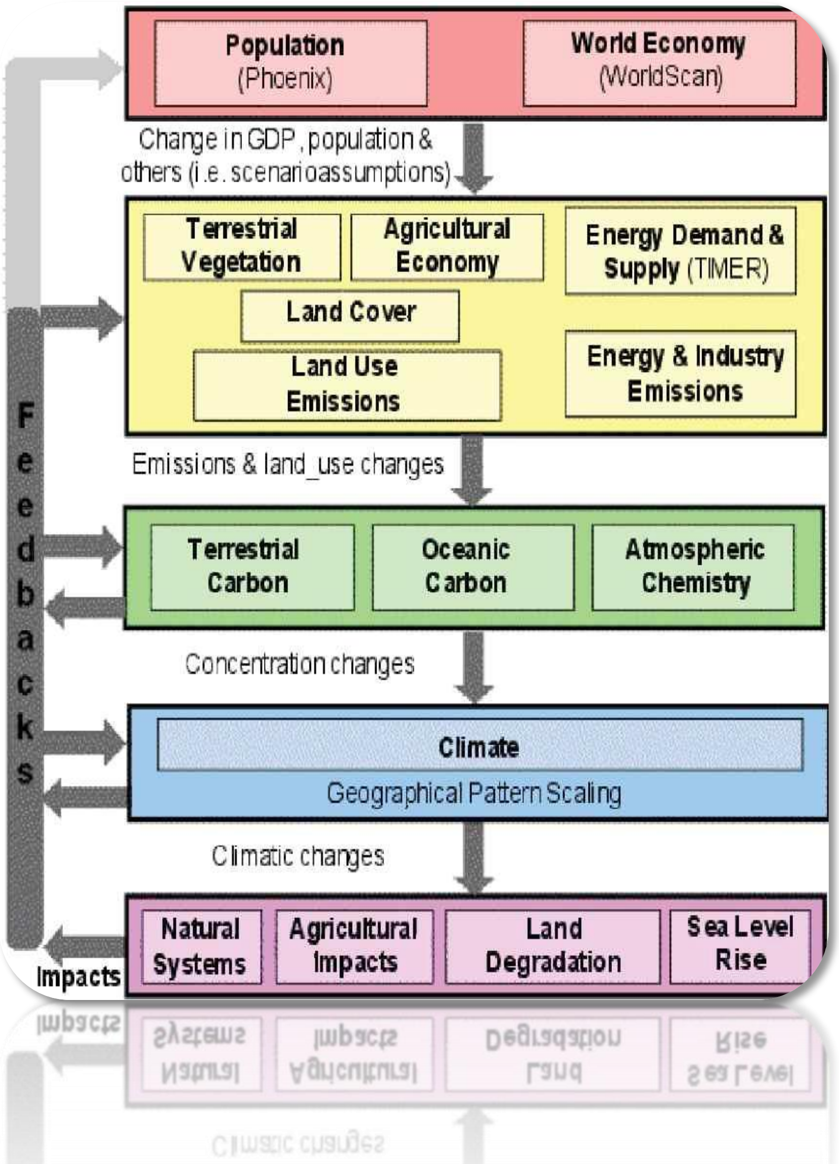
Reliable Climate Forecasts from Next Generation Earth System Models

- **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

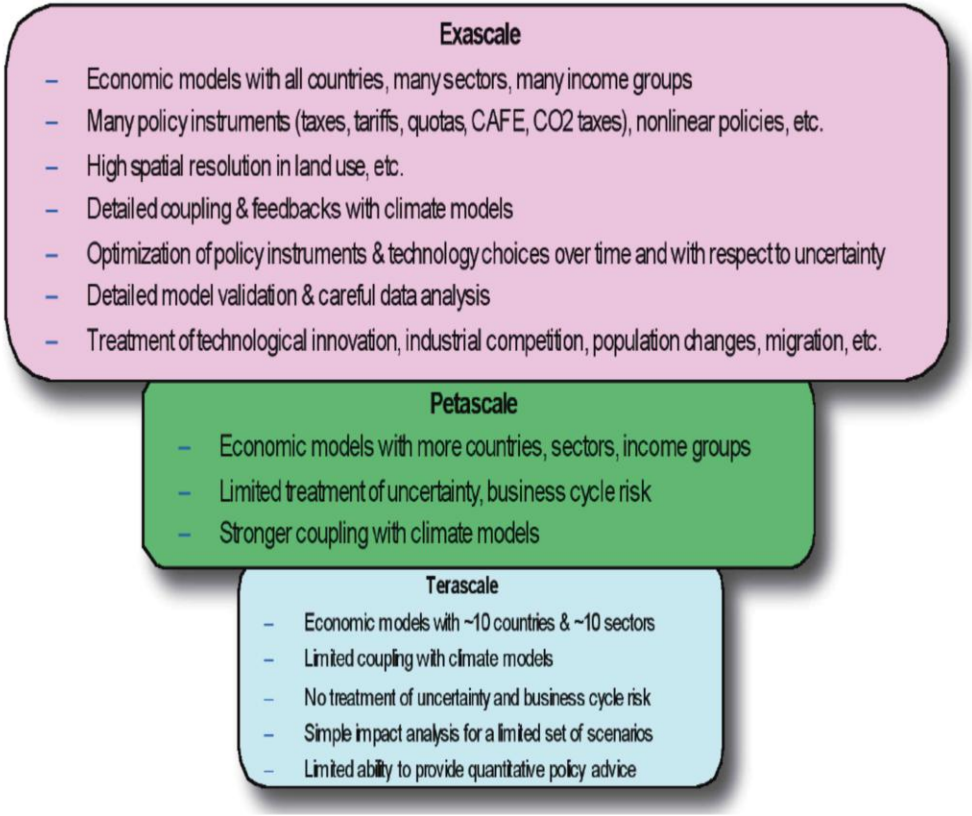
From Earth System Modeling to Computational Socio-Economics



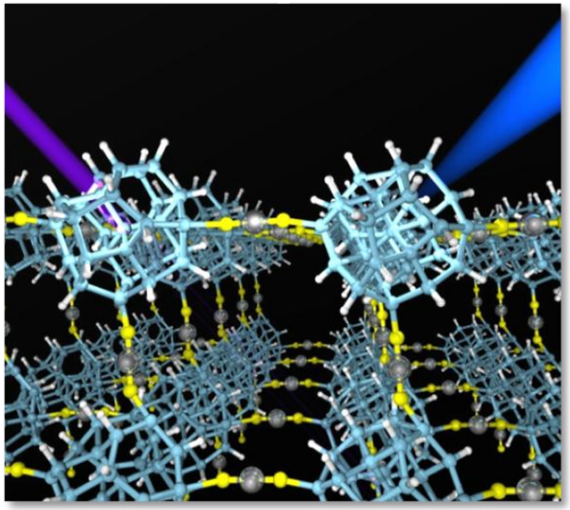
- Earth system modeling has progressed to a point where there is considerable confidence in predictions of continental- and global-scale climate changes over the next 100 years [IPCC 2007]
- Integrated modeling of the social, economic, and environmental system with an extensive treatment of couplings 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 ~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

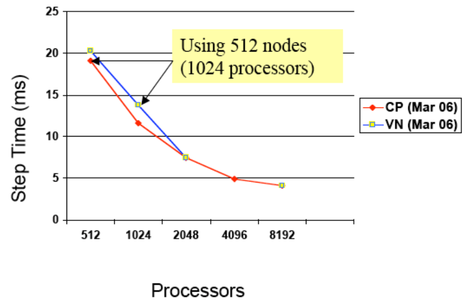
These problems each have very large parameter spaces, so design optimizations take many runs

Petascale Molecular Modeling

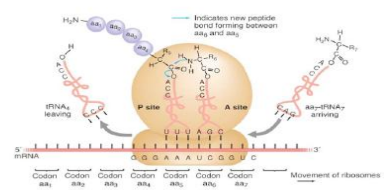
NAMD on BlueGene/L

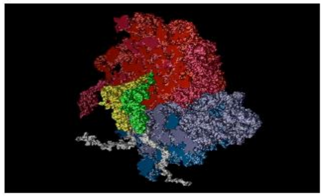
Procs	Time (ms)
1	9000
32	347
128	97.2
512	23.7
1024	13.8
2048	8.6
4096	6.2
8192	5.2

Oct'05

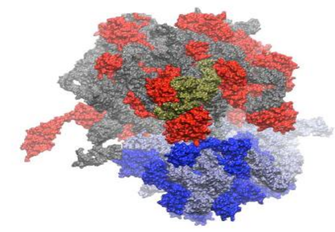


Petascale Project 2: Ribosome



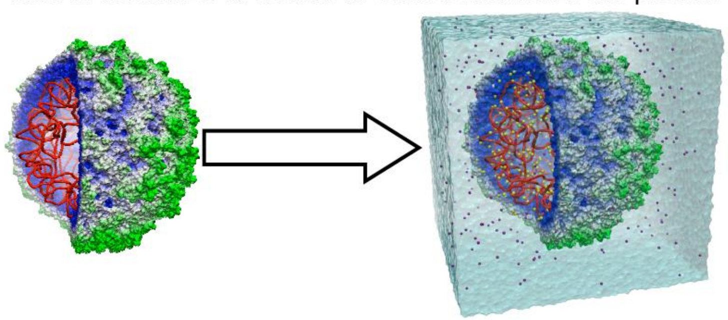


- Very large system size: ~3,000,000 atoms
- Great biological and biomedical relevance
- Simulations with close collaboration with leading experimentalist
- All-atom, coarse graining, and multiscale simulations



Petascale Project 1: Virus Capsid

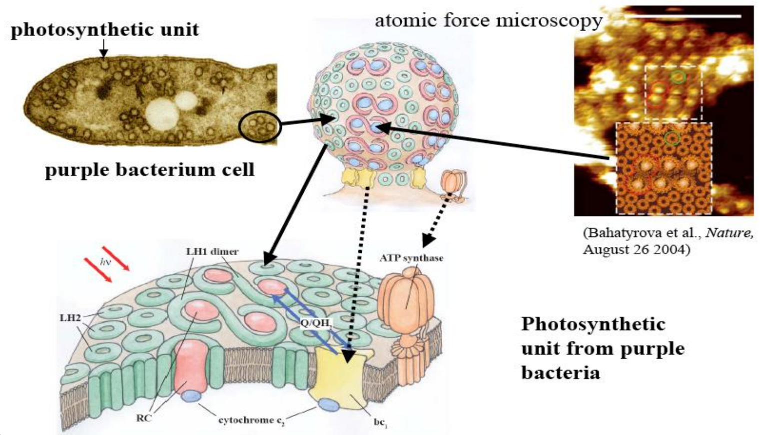
Solvate the virus in a 220Åx220Åx220Å water box, add Mg^{2+} ions to neutralize RNA and Cl^- ions to neutralize the protein



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

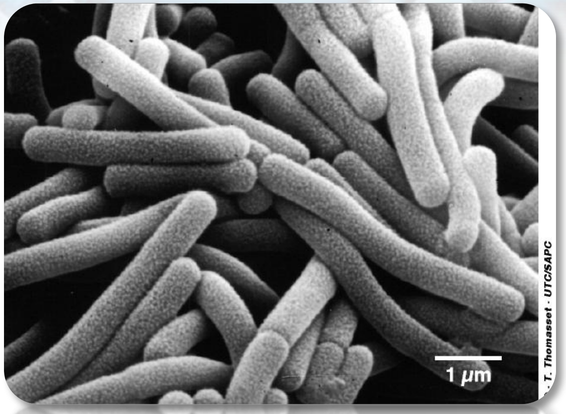


(Bahatyrova et al., *Nature*, August 26 2004)

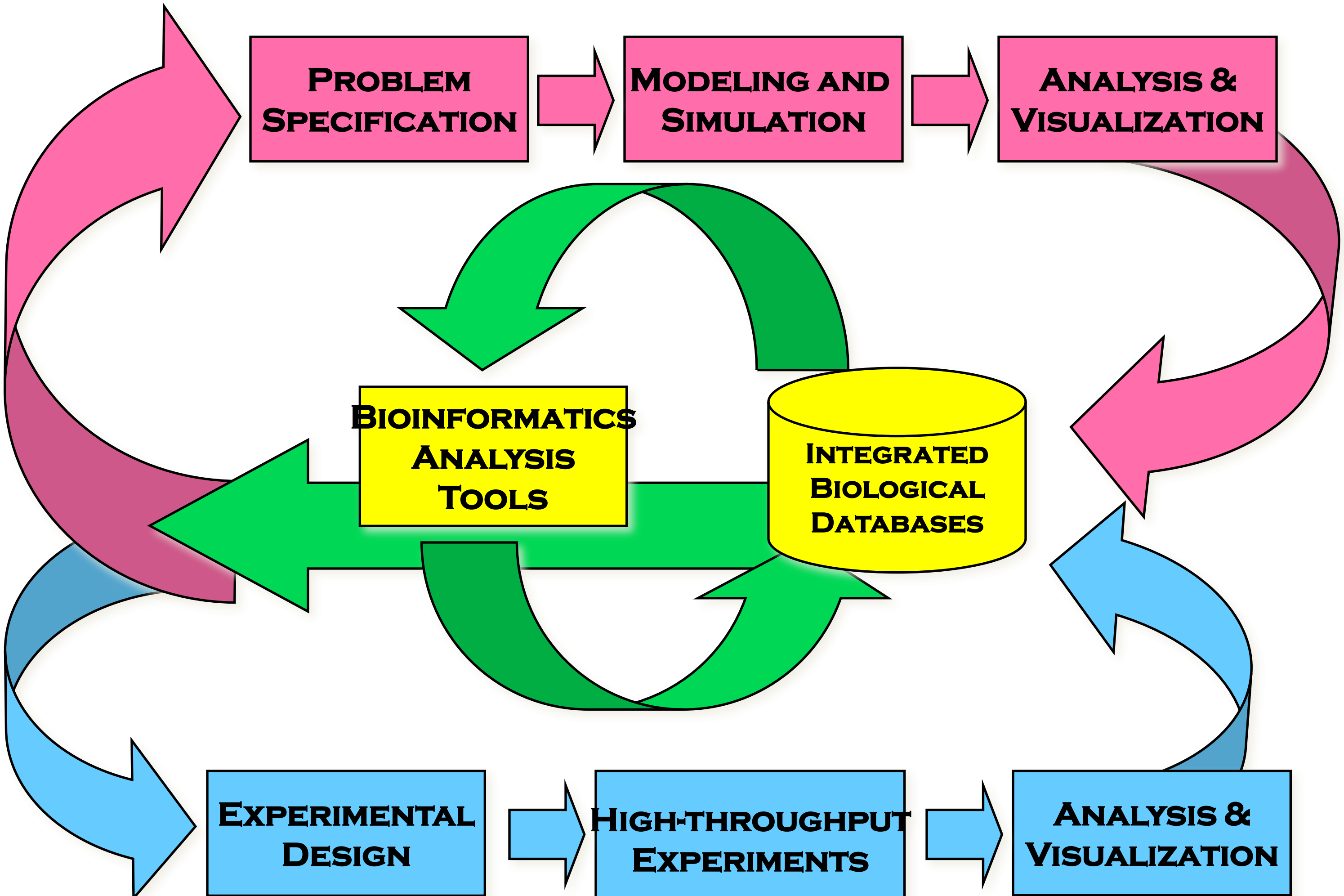
Photosynthetic unit from purple bacteria

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 effective scoped
- **Related experimental support is needed**
  - Validation experiments driven by the simulation and modeling will be required

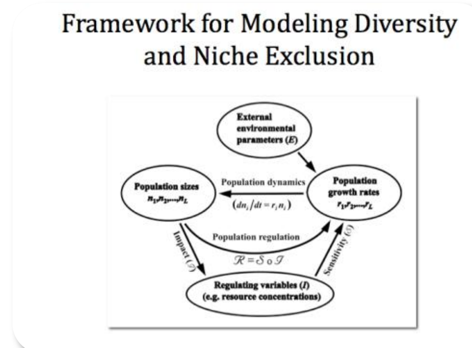


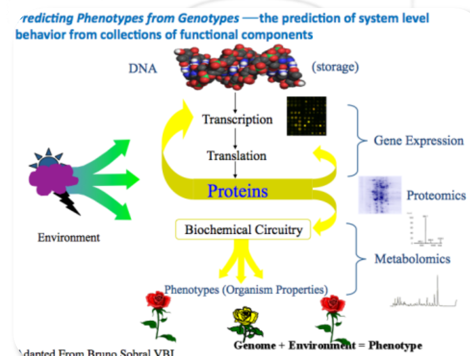
An Integrated View of Modeling, Simulation, Experiment, and Bioinformatics

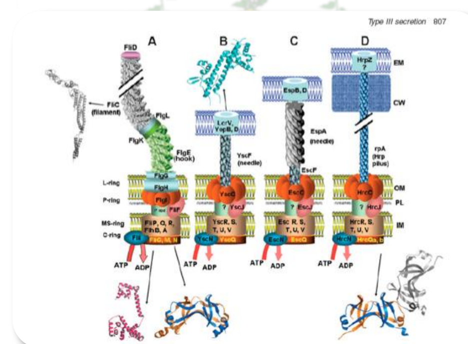


Six Open Problems in Basic Biology Where Computing Can Have an Impact

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 TransferEvents** —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



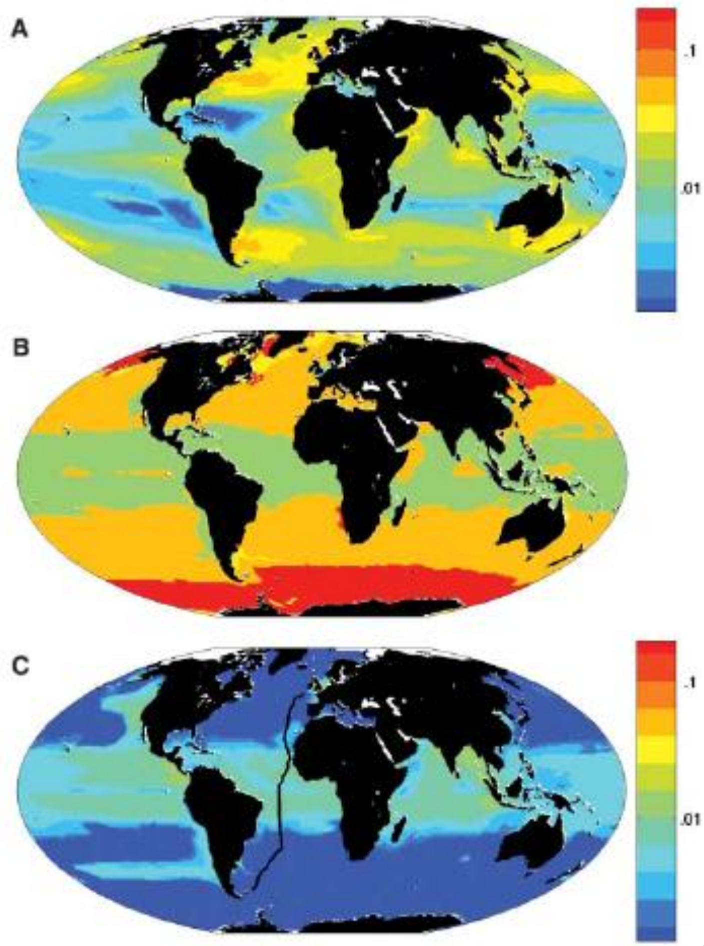




Emergent Biogeography of Microbial Communities in a Model Ocean

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 (μM 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 (μM P, 0 to 50 m average). Black line indicates the track of AMT13.



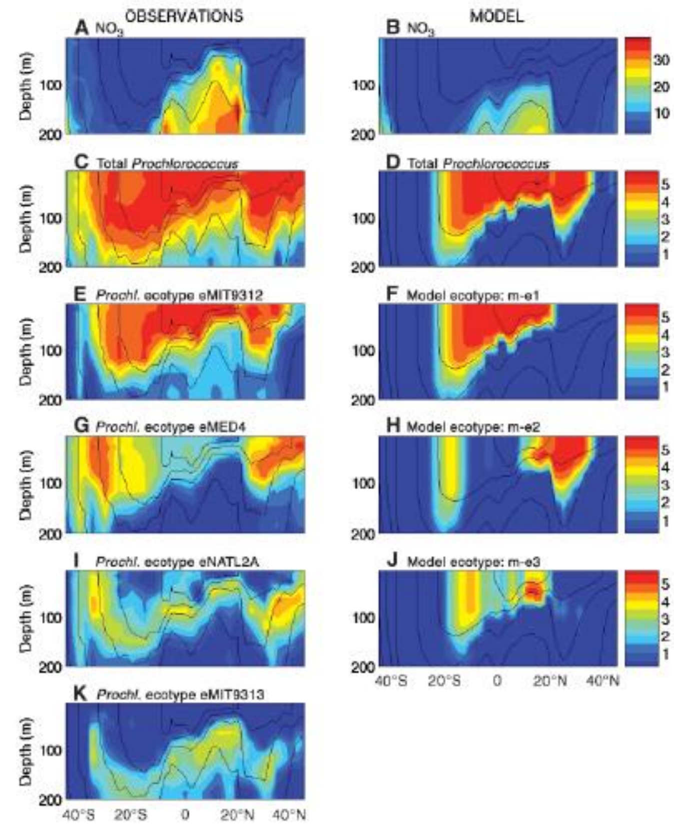
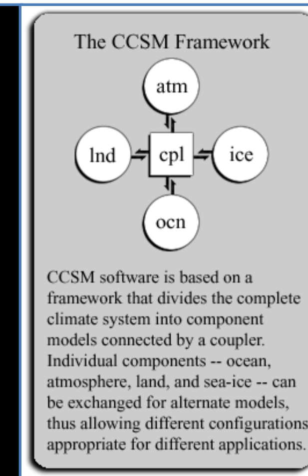


Fig. 2. Observed and modeled properties along the AMT13 cruise track. Left column shows observations (*17*), right column shows results from a single model integration. (**A** and **B**) Nitrate (μmol kg^{-1}); (**C** and **D**) total *Prochlorococcus* abundance [log (cells ml^{-1})]. (**E**, **G**, **I**, and **K**) Distributions of the four most abundant *Prochlorococcus* ecotypes [log (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.

SCIENCE VOL 315 30 MARCH 2007

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



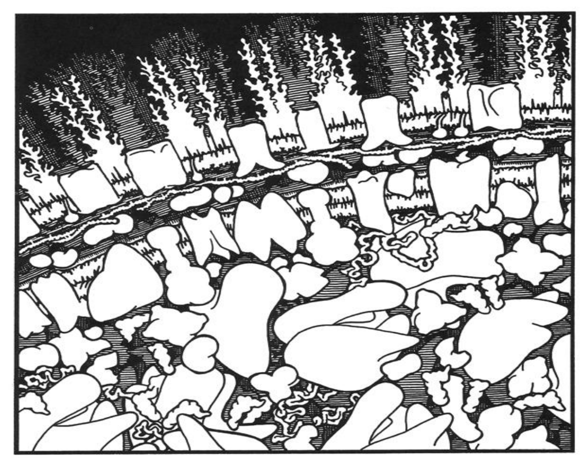
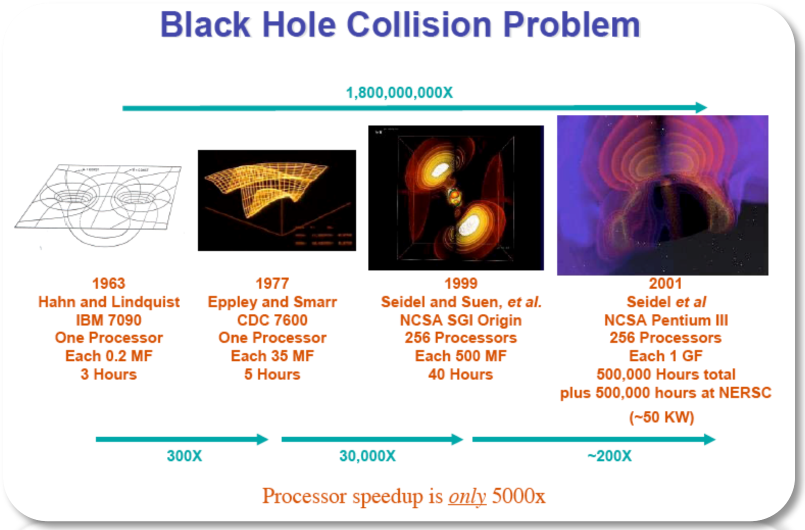
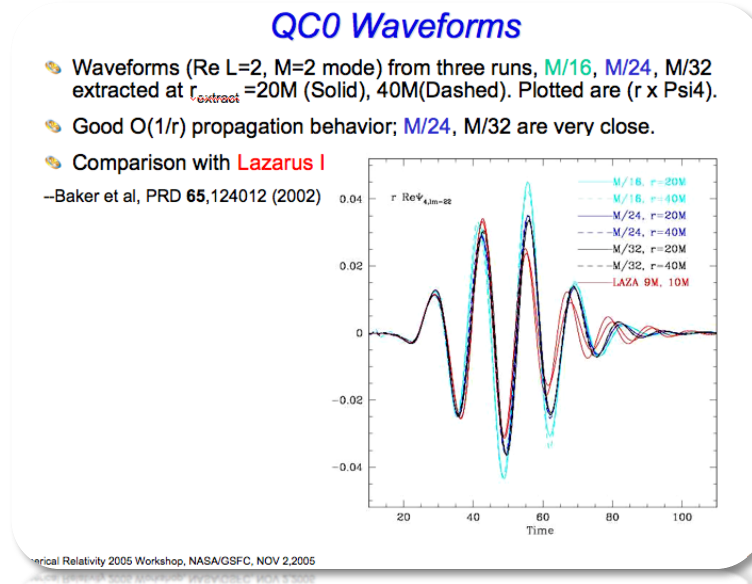
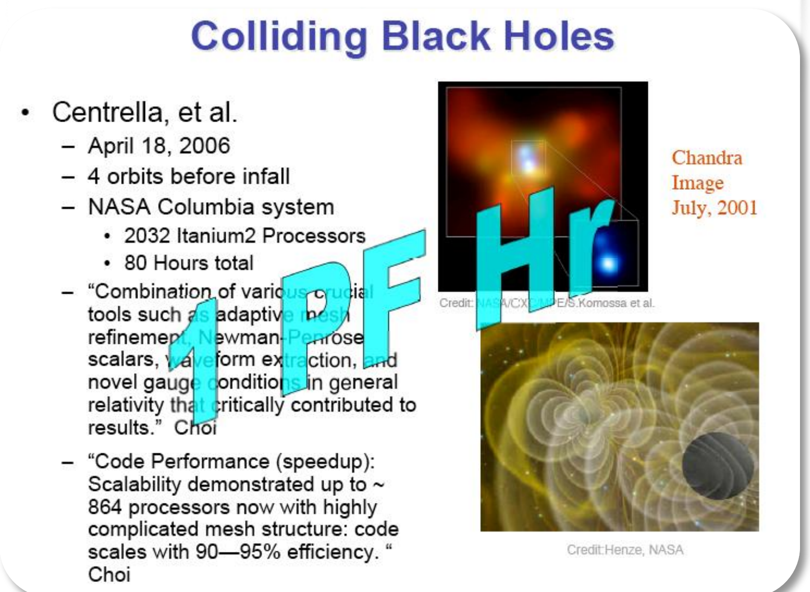


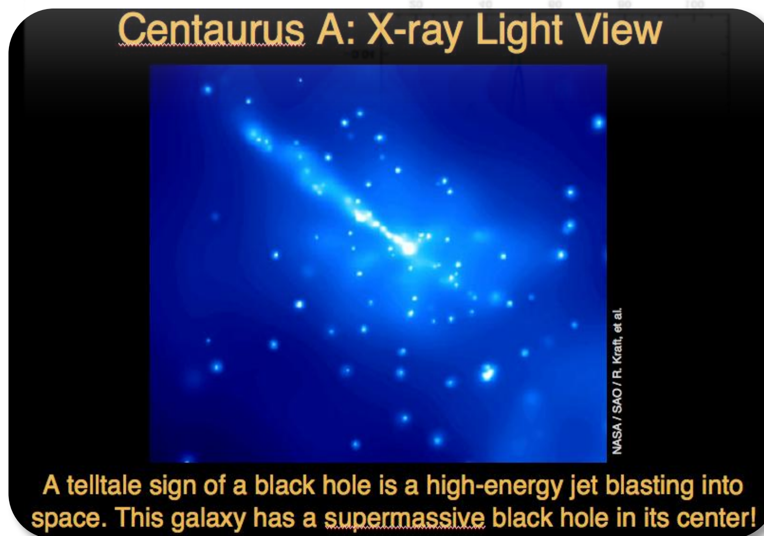
Figure 4.3 Cell Wall

Colliding Black Holes









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]

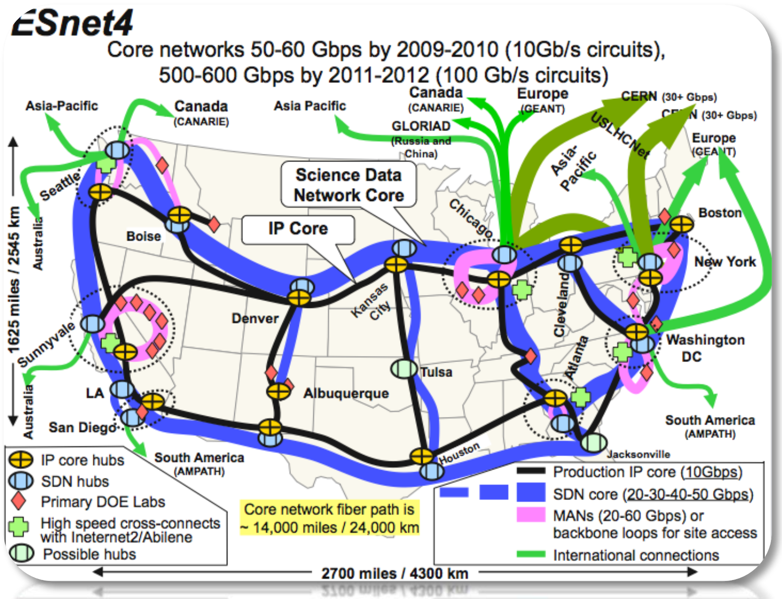
BG/P Configuration Generation Plans

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

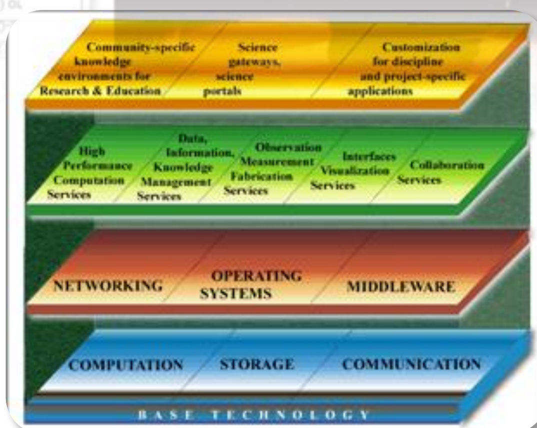
1. Monte Carlo methods generate representative configurations of the QCD ground state -- time intensive
2. Use configurations to calculate a wide variety of quantities of interest in high energy and nuclear physics.

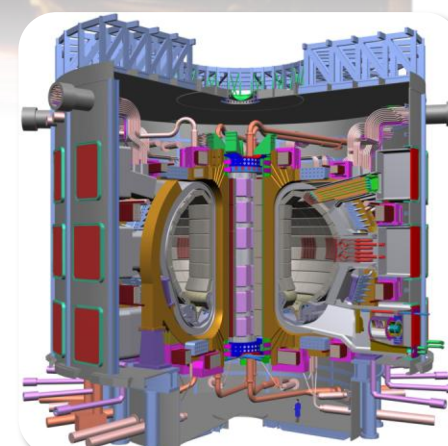
Integrating Leadership Computing Into the International Research Infrastructure



Emerging Areas and Directions

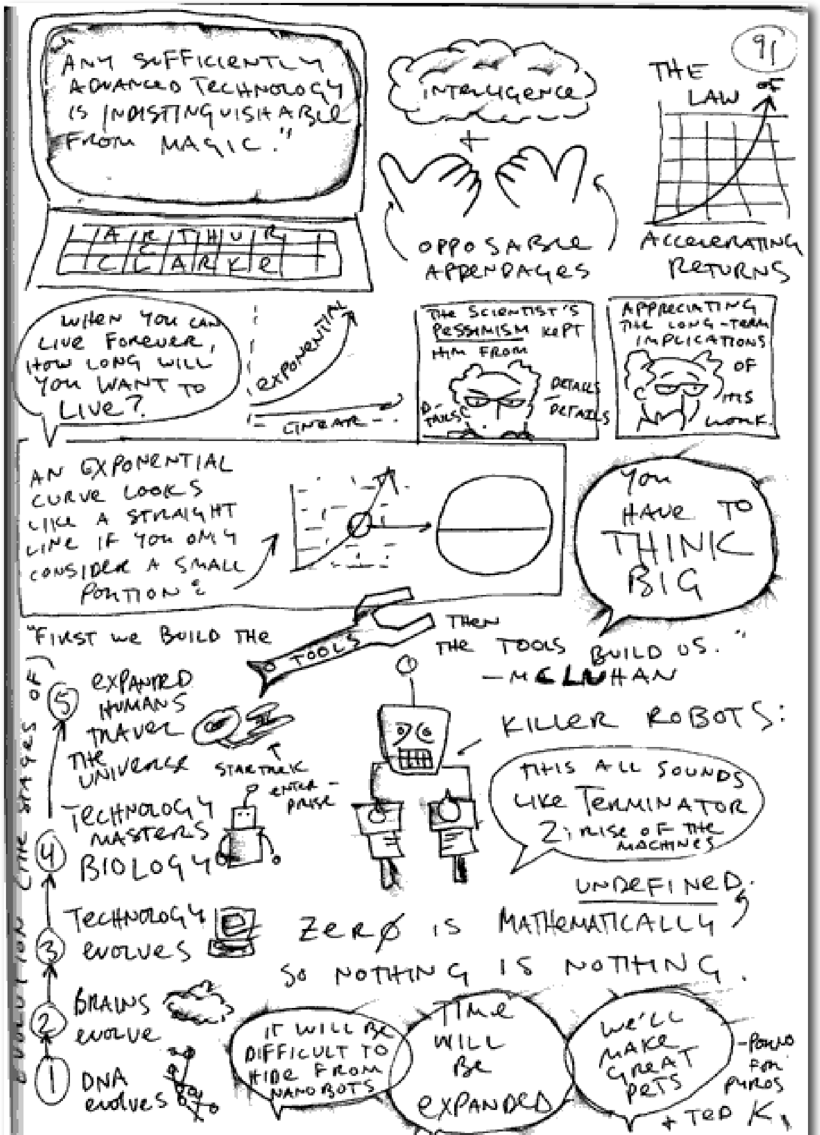
- Economics
- Large-scale Optimization
- Large-scale Data Analysis
- Agent-Based Models (cyber security, evolution, social organizations)
- Parallel Symbolic Computation
- Interactive Exploratory Analysis
- Sensor Network Data Assimilation
- Comparative Genomics
- Cell Network Models

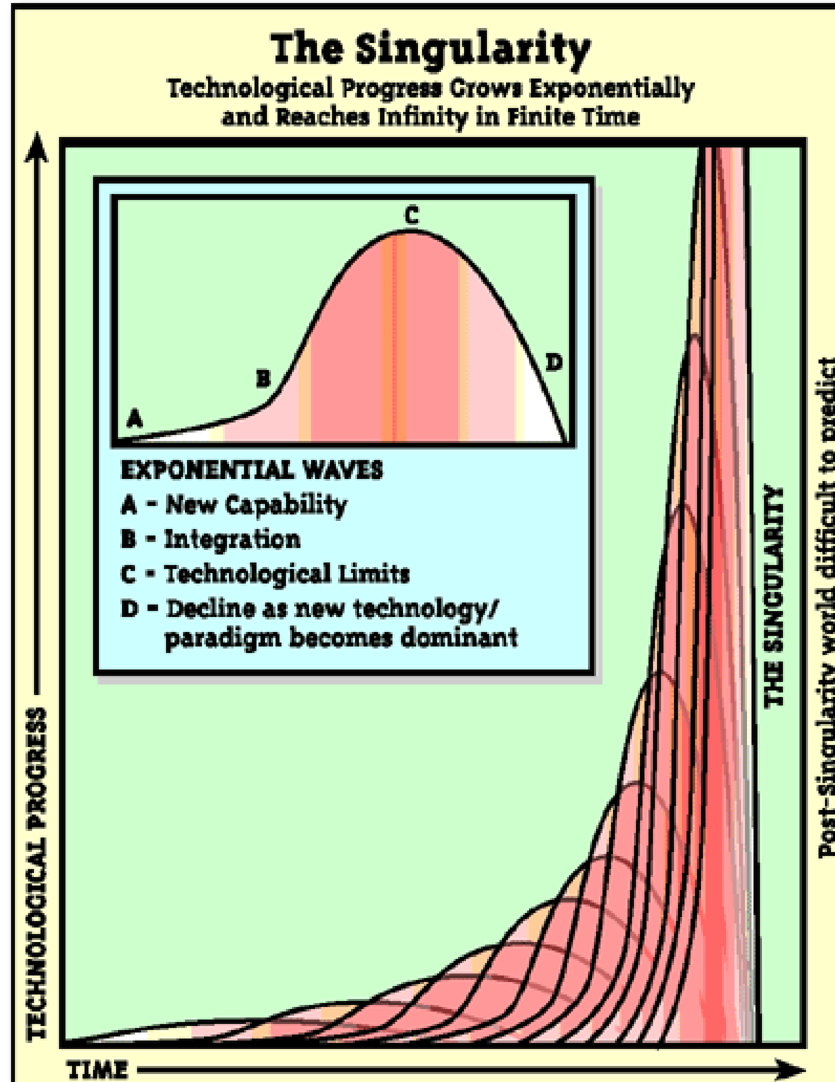




Some Final Words

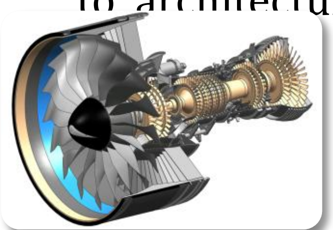
- Scientific breakthroughs require flexibility and abundance of computing resources for serendipity and insight to work.
  - One must be able to make lots of mistakes.. therefore cost matters to make mistakes affordable
- High-capability platforms require considerable quantities of capacity platforms 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 country 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..

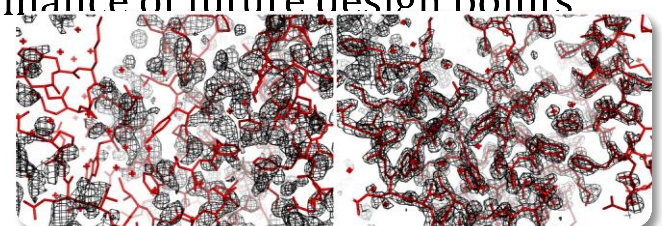

"ANY SUFFICIENTLY ADVANCED TECHNOLOGY IS INDISTINGUISHABLE FROM MAGIC."
ARTHUR CLARKE
INTELLIGENCE
+
OPPOSABLE APPENDAGES
THE LAW OF ACCELERATING RETURNS
91
WHEN YOU CAN LIVE FOREVER, HOW LONG WILL YOU WANT TO LIVE?
EXPONENTIAL
LINEAR
THE SCIENTIST'S PESSIMISM KEPT HIM FROM
DETAILS DETAILS
APPRECIATING THE LONG-TERM IMPLICATIONS OF HIS WORK.
AN EXPONENTIAL CURVE LOOKS LIKE A STRAIGHT LINE IF YOU ONLY CONSIDER A SMALL PORTION
YOU HAVE TO THINK BIG
"FIRST WE BUILD THE TOOLS THEN THE TOOLS BUILD US." —MCLUHAN
EVOLUTION (THE STAGES OF)
5 EXPANDED HUMANS TRAVEL THE UNIVERSE
STAR TREK ENTERPRISE
4 TECHNOLOGY MASTERS BIOLOGY
3 TECHNOLOGY EVOLVES
2 BRAINS EVOLVE
1 DNA EVOLVES
KILLER ROBOTS:
THIS ALL SOUNDS LIKE TERMINATOR 2: RISE OF THE MACHINES
ZERO IS MATHEMATICALLY UNDEFINED.
SO NOTHING IS NOTHING.
IT WILL BE DIFFICULT TO HIDE FROM NANOBOTS
TIME WILL BE EXPANDED
WE'LL MAKE GREAT PETS
—PORNO FOR PYROS
+ TED K.

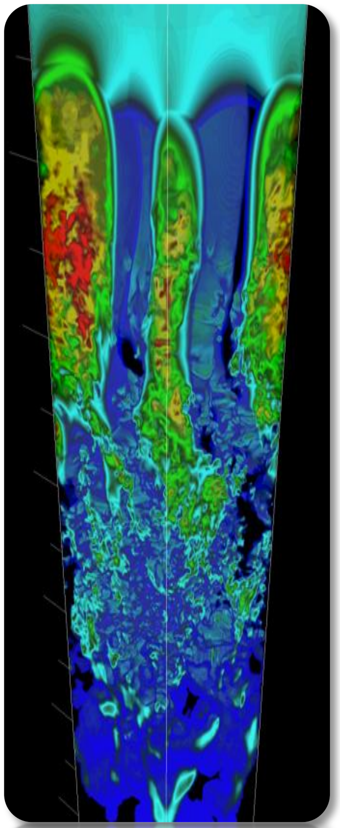

The Singularity
Technological Progress Grows Exponentially and Reaches Infinity in Finite Time
A
B
C
D
EXPONENTIAL WAVES
A - New Capability
B - Integration
C - Technological Limits
D - Decline as new technology/paradigm becomes dominant
TECHNOLOGICAL PROGRESS
TIME
THE SINGULARITY
Post-Singularity world difficult to predict

Some Conclusions

- We understand the role of leadership class computing in science
- Building a long-term engagement with the best basic science communities is critical to enable LCC to have maximum scientific impact.
- Each lab can effectively do this for a relatively small set of areas Argonne's focus: Fundamental Physics, Biology, Multi-Physics CFD, Large-Scale Optimization
- It critical for the community to have multiple computing platforms to enable the most cost effective science and to mitigate risk
- Understanding the arch-app coupling is critical for effective decision making
- Significant effort is needed to determine the best match of algorithms to architectures and to estimate performance of future design points







A push to the exascale is a ten year vision to keep the US at the forefront of what is possible in high-end computing. The challenges are many and it is likely that it will need to be a global effort in both research and development and the development of codes.