An Analysis of Application Requirements for Leadership Computing

NATIONAL CENTER
FOR COMPUTATIONAL SCIENCES

Bronson Messer
Scientific Computing Group
NCCS

Oak Ridge National Laboratory
U.S. Department of Energy

Co-authors and collaborators

- Doug Kothe (NCCS Director of Science)
- Ricky Kendall (Scientific Computing Group Leader)
- the rest of the Scientific Computing Group

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

Overview

- What is this exercise?
- Current context: NLCF INCITE projects overview
- How did we gather the data and what are they?
- What did we do with the data?
- What do we now believe?
- What are the ramifications of our conclusions for future systems?

Description

- In an effort to guide future system acquisition, we have started compiling, correlating, and analyzing a number of computational requirements from a variety of application areas

- The original "survey population" was FY 2006 NCCS project teams
  - **Current project list is different, but similar**

- A "living" (read as: incomplete and messy) document exists in draft form - "Computational Science Requirements for Leadership Computing"

- At present, it is mostly a collection and distillation of several data sources on application requirements:
  - **NCCS highlights and quarterly updates from projects**
  - **ASCAC Code Project Questionnaire**
  - **Survey of Acceptance and Early Access Science Applications**
  - **Insider information and educated guesses by Scientific Computing Group members**

ORNL Provides Leadership Computing to 2007 INCITE Program

- The NCCS is providing leadership computing to 28 programs in 2007 under the DOE's Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program.
- Leading researchers from government, industry, and the academic world will use more than 75 million processor hours on the center's Cray leadership computers.
- The center's Cray XT4 (Jaguar) and Cray X1E (Phoenix) systems will provide more than 75% of the computing power allocated for the INCITE program.

Total INCITE Allocations: 45 projects, 95 million hrs

NCCS Allocations: 28 projects, 75 million hrs

Projects

Astrophysics

Multi-Dimensional Simulations of Core-Collapse Supernovae
Anthony Mezzacappa (Oak Ridge National Laboratory)

First Principles Models of Type Ia Supernovae
Stan Woosley (University of California, Santa Cruz)

Via Lactea: A Billion Particle Simulation of the Milky Way's Dark Matter Halo
Piero Madau (University of California, Santa Cruz)

Numerical Relativity Simulations of Binary Black Holes and Gravitational Radiation
Joan Centrella (NASA/Goddard Space Flight Center)

Biology

Next Generation Simulations in Biology: Investigating Biomolecular Structure, Dynamics and Function Through Multi-Scale Modeling
Pratul Agarwal (Oak Ridge National Laboratory)

Gating Mechanism of Membrane Proteins
Benoit Roux (Argonne National Laboratory and University of Chicago)

Chemistry

An Integrated Approach to the Rational Design of Chemical Catalysts
Robert Harrison (Oak Ridge National Laboratory)

Climate

Climate-Science Computational End Station Development and Grand Challenge Team
Warren Washington (National Center for Atmospheric Research)

Eulerian and Lagrangian Studies of Turbulent Transport in the Global Ocean
Synte Peacock (University of Chicago)

Assessing Global Climate Response of the NCAR-CCSM3: CO_2 Sensitivity and Abrupt Climate Change
Zhengyu Liu (University of Wisconsin - Madison)

Computer Science

Performance Evaluation and Analysis Consortium End Station
Patrick Worley (Oak Ridge National Laboratory)

Materials

Predictive Simulations in Strongly Correlated Electron Systems and Functional Nanostructures
Thomas Schulthess (Oak Ridge National Laboratory)

Linear Scale Electronic Structure Calculations for Nanostructures
Lin-Wang Wang (Lawrence Berkeley National Laboratory)

Bose-Einstein Condensation vs. Quantum Localization in Quantum Magnets
Tommaso Roscilde (Max-Planck Gesellschaft)

Projects
(continued)

2007 INCITE PROGRAM

Fusion Energy

Gyrokinetic Plasma Simulation
W.W. Lee (Princeton Plasma Physics Laboratory)

Simulation of Wave-Plasma Interaction and Extended MHD in Fusion Systems
Don Batchelor (Oak Ridge National Laboratory)

Interaction of ITG/TEM and ETG Gyrokinetic Turbulence
Ronald Waltz (General Atomics)

Gyrokinetic Steady State Transport Simulations
Jeff Candy (General Atomics)

High Power Electromagnetic Wave Heating in the ITER Burning Plasma
Fred Jaeger (Oak Ridge National Laboratory)

Geosciences

Modeling Reactive Flows in Porous Media
Peter Lichtner (Los Alamos National Laboratory)

High Energy Physics

Computational Design of the Low-loss Accelerating Cavity for the ILC
Kwok Ko (Stanford Linear Accelerator Center)

Lattice QCD for Hadronic and Nuclear Physics
Robert Edwards (Thomas Jefferson National Accelerator Facility)

Atomic Physics

Computational Atomic and Molecular Physics for Advances in Astrophysics, Chemical Sciences, and Fusion Energy Sciences
Michael Pindzola (Auburn University)

Nuclear Physics

Ab-Initio Nuclear Structure Computations
David J. Dean (Oak Ridge National Laboratory)

Combustion

High-Fidelity Numerical Simulations of Turbulent Combustion - Fundamental Science Towards Predictive Models
Jackie Chen (Sandia National Laboratory)

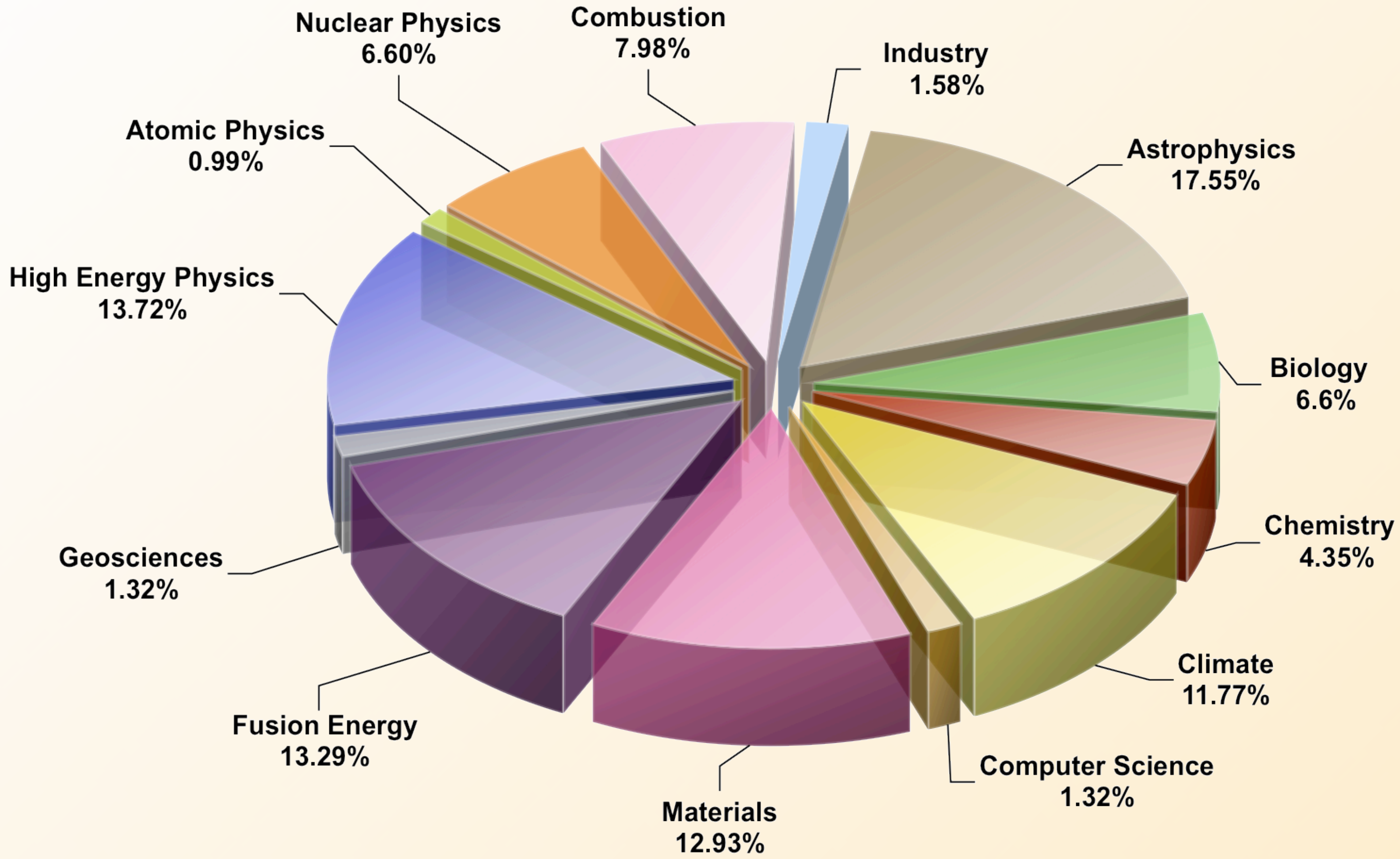
Industry

Real-Time Ray-Tracing
Evan Smyth (Dreamworks Animation)

Development and Correlations of Large Scale Computational Tools for Flight Vehicles
Moeljo Hong (The Boeing Company)

Ab Initio Modeling of the Glass Transition
John Mauro (Corning Incorporated)

INCITE Allocations by Discipline



% of total hours allocated

Science Questions

- Topics of active research represent the union of:
  - **What questions require large scale computing?**
  - **The particular research interests of our users**

- These are NOT always exactly the same thing, though the overlap is, naturally, quite high.

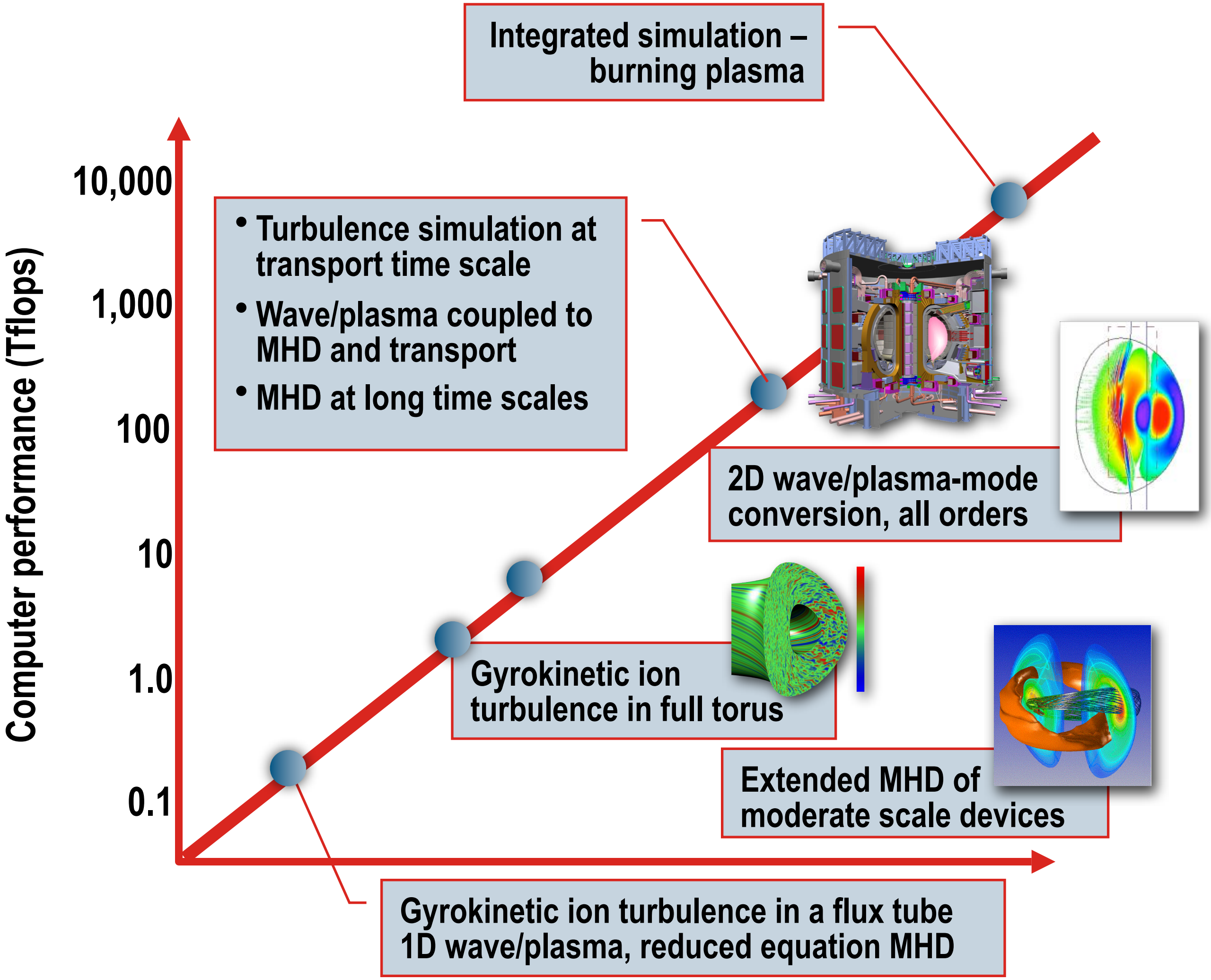
Science Domain	Science Driver
Accelerator Physics	*Evaluate and optimize a new low-loss cavity design for the International Linear Collider (ILC) that has a lower operating cost and higher performance than existing designs.*
Astrophysics	*Determine the explosion mechanism of core-collapse supernovae, one of the universe's most important sites for nucleosynthesis and galactic chemical enrichment.* *Determine details of the explosion mechanism of Type Ia supernovae (thermonuclear explosions of white dwarf stars), helping to determine key characteristics for their use as standard candles for cosmology.*
Biology	*How will the world address the current oil and gasoline crisis?* *One option is ethanol: studying the most efficient means of converting cellulose to ethanol.*
Chemistry	*Catalytic transformation of hydrocarbons; clean energy and hydrogen production and storage; chemistry of transition metal clusters including metal oxide.*
Climate	*Focus on Grand Challenge of climate change science: predict future climates based on scenarios of anthropogenic emissions and other changes resulting from options in energy policies.* *Simulate the dynamic ecological and chemical evolution the climate system.* *Develop, deliver, and support the Community Climate System Model (CCSM).*
Combustion	*Developing cleaner-burning, more efficient devices for combustion.*
Engineering	*Development and correlations/validations of large-scale computational tools for flight vehicles.* *Demonstrating the applicability & predictive accuracy of CFD tools in a production environment.* *Flight vehicle phenomena such as fluid-structure/flutter interaction, and control surface free-plays.*
Fusion	*Fusion reactor technology is in its infancy and fundamental questions must be resolved. The race is on develop analysis tools before ITER comes on line (projected 2015).* *Major hurdle: understand and control plasma turbulent fluctuations that cause particles and energy to travel from the center of the plasma to the edge, causing loss of heat needed to maintain the fusion reaction.*
High Energy Physics	*The Large Hadron Collider (LHC) physics program seeks to find the Higgs particles thought to be responsible for mass, and to find evidence of supersymmetry (SUSY), a necessary element of String Theories that may unify all of nature's fundamental interactions.*
Materials Science	*Predictive simulation of brittle and ductile materials subjected to high-speed loads.* *Understanding the initiation of failure in a local region, the appearance of a macro-crack due to the coalescence of subscale cracks, the localization of deformation due to coalescence of voids, the dynamic propagation of cracks or shear bands, and eventual fragmentation and failure of the solid.*
Nanoscience	*Understanding the quantitative differences in the transition temperatures of high temperature superconductors.* *Understanding and improving colossally magneto-resistive oxides and magnetic semiconductors.* *Developing new switching mechanism in magnetic nanoparticles for ultra high density storage.* *Simulation and design molecular-scale electronics devices.* *Elucidate the physico-chemical factors mechanisms that control damage to DNA.*
Nuclear Energy	*Virtual reactor core, radio-chemical separations reprocessing, fuel rod performance, repository*
Nuclear Physics	*How are we going to describe nuclei whose properties we cannot measure?* *Explore thermal nuclear properties in the mass 80-150 region*

What can be done at 1 PF?

- Future questions come in several flavors
  - Higher fidelity needed for the same problem
  - A new problem that hasn’t been attacked before for lack of computational might
  - “externally determined” e.g. ITER

Science Domain	Code	Science Achievements Possible at 1 PF
Accelerator Physics	T3P	ILC design guidance
Geophysics	PFLOTRAN	Multi-scale, multi-phase, multi-component modeling of a 3D field CO2 injection scenario for CO2 sequestration studies at an identified field site
Materials Science	QBOX	Fundamentals of phase change
Nanoscience	VASP (+WL)	Magnetic properties of FePt nanoparticles
Nuclear Energy	NEWTRNX	6D multi-group, multi-angle neutron transport in an entire reactor core consisting of ~10,000,000 fuel pins
Nuclear Physics	CCSD	Properties (mass, transition strengths) of medium mass nuclei
Chemistry	NWCHEM	Nanotube catalysis
Chemistry	ORETRAN	Molecular electronics and transport in nanostructures
QCD	MILC/CHROMA	Determine exotic meson spectrum down to pion mass
Nanoscience	CASINO	Photodissociation of water on titanium dioxide surface; hydrogen storage in organic and solid state nanostructures
Nanoscience	LSMS (+WL)	Determination of temperature-dependent free energy for magnetic FePt nanoparticles, allowing exploitation of its magnetic properties for designing high density (>10 TB/in2) magnetic storage media
High-Temperature Conductivity	QMC/DCA	High-temperature superconductivity with multi-band Hubbard model using material-specific band structures
Astrophysics	CHIMERA	First 3D core-collapse supernova simulation with realistic neutrino transport, magnetic fields, general relativistic gravity, and realistic nuclear burning
Climate	POP/CICE	Fully-coupled, eddy-resolving simulation of North Atlantic current systems with sea ice. Aim to understand polar ice cap melting scenarios. Eddy-resolving, two-hemisphere Atlantic computations (including the Antarctic Circumpolar connection), with the goal of understanding the factors controlling the oceanic poleward heat transport.
Combustion	S3D	Flame stabilization of high-pressure n-heptane jets in low temperature mixing-controlled diesel combustion
Fusion	GTC	Understand anomalous particle transport for electrons in the presence of electromagnetic effects for long-time scale ITER-size simulations
Fusion	GYRO	Steady-state temperature and density profile predictions for ITER (ions and electrons) given pedestal boundary conditions.
Chemistry	MADNESS	Exact simulation of the dynamics of a fully interacting few-electron system (He, H2, H3+, Li, LiH) in a general external field
Fusion	AORSA	Complete simulation of mode conversion heating in ITER with a realistic antenna geometry and non-Maxwellian alpha particles
Biology	LAMMPS	Millisecond simulation of million-atom protein breathing motions and enzyme complexes such as lactate dehydrogenase
Biology	CHARMM, NAMD	Simulation of motor function on the physical time scale

Software and science: Fusion



Integrated simulation – burning plasma

- Turbulence simulation at transport time scale
- Wave/plasma coupled to MHD and transport
- MHD at long time scales

2D wave/plasma-mode conversion, all orders

Gyrokinetic ion turbulence in full torus

Extended MHD of moderate scale devices

Gyrokinetic ion turbulence in a flux tube
1D wave/plasma, reduced equation MHD

Expected Outcomes

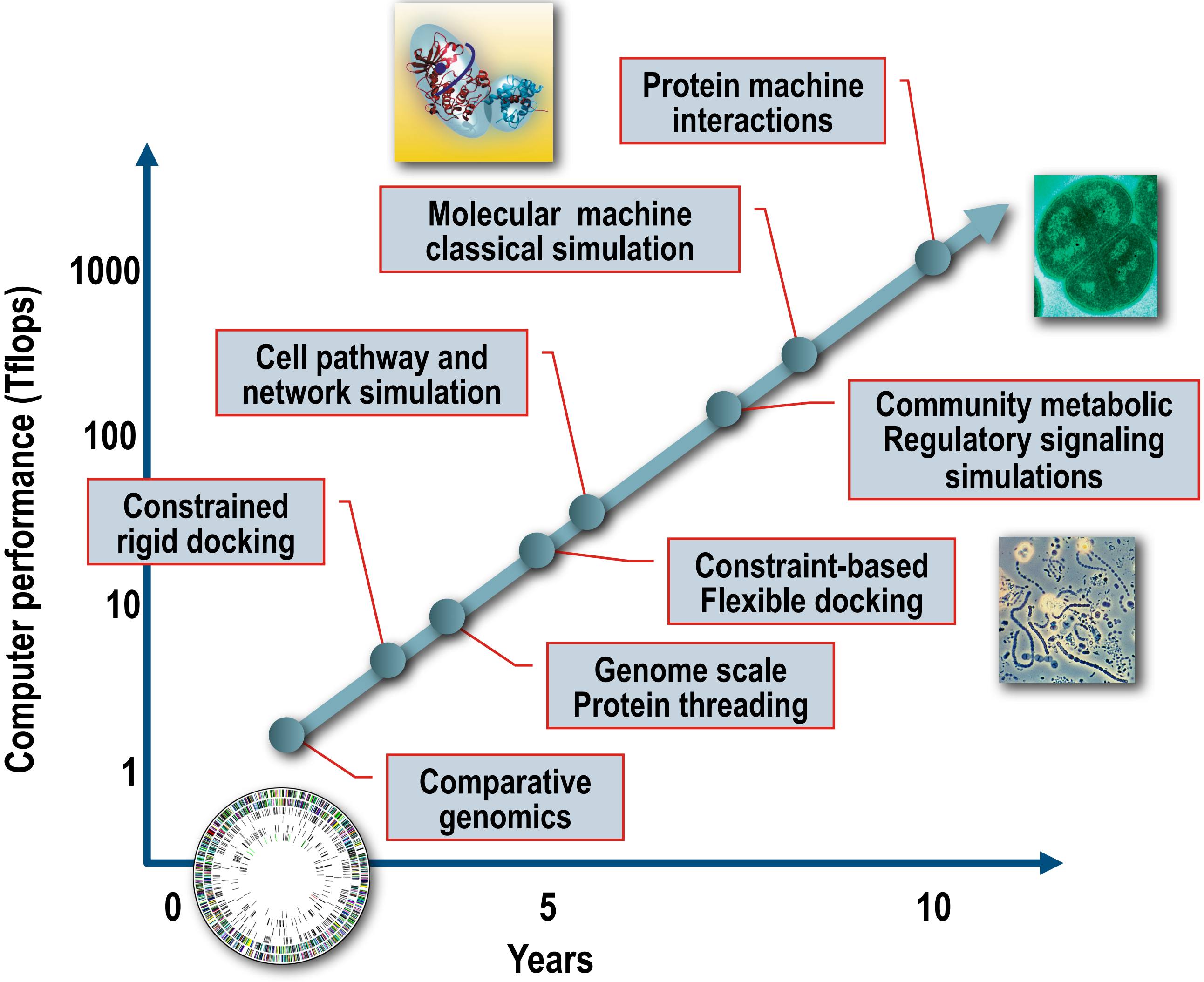
5 Years

- Full-torus, electromagnetic simulation of turbulent transport with kinetic electrons for simulation times approaching transport time-scale
- Develop understanding of internal reconnection events in extended MHD, with assessment of RF heating and current drive techniques for mitigation

10 years

- Develop quantitative, predictive understanding of disruption events in large tokamaks
- Begin integrated simulation of burning plasma devices – multi-physics predictions for ITER

Software and science: Biology



Comparative genomics

Constrained rigid docking

Genome scale Protein threading

Constraint-based Flexible docking

Cell pathway and network simulation

Community metabolic Regulatory signaling simulations

Molecular machine classical simulation

Protein machine interactions

Expected outcomes

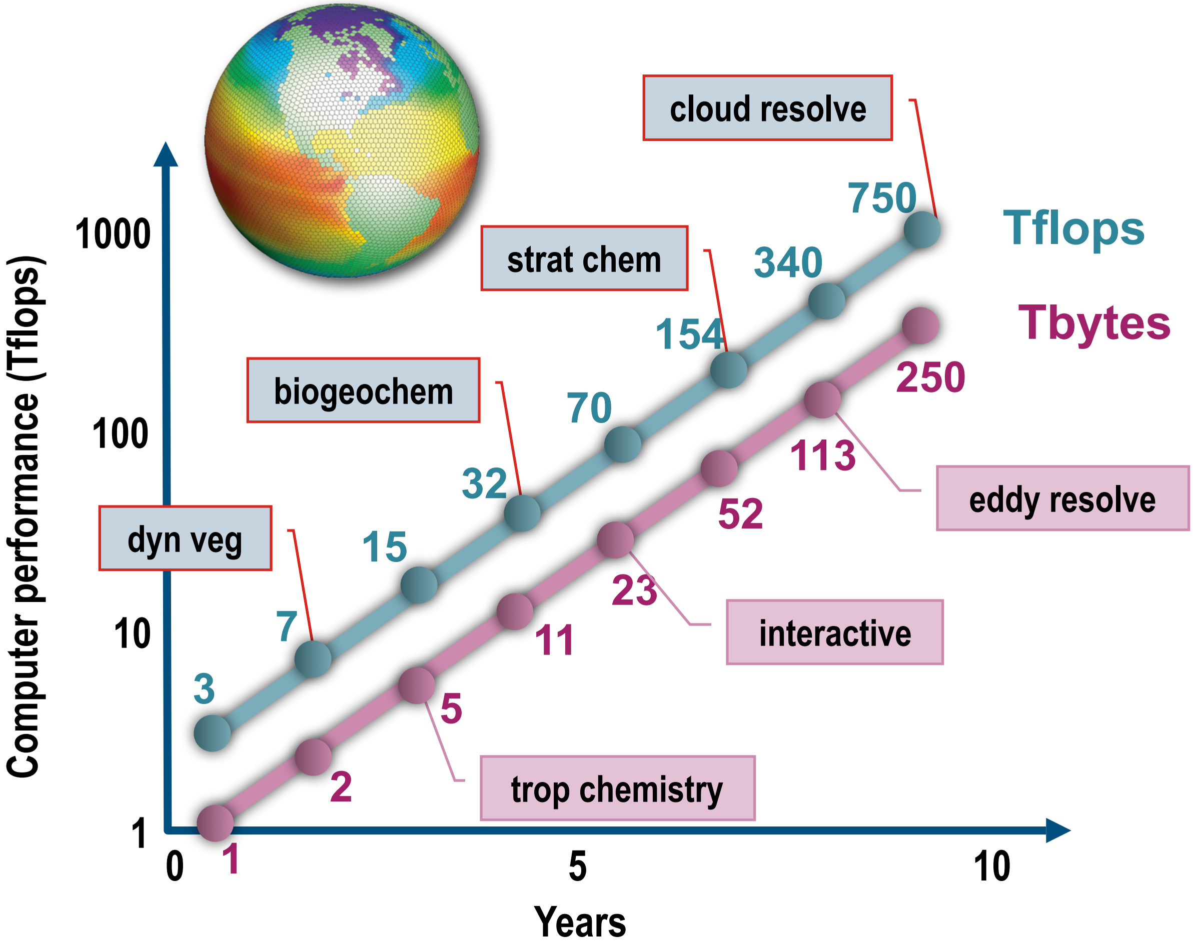
5 years

- Metabolic flux modeling for hydrogen and carbon fixation pathways
- Constrained flexible docking simulations of interacting proteins

10 years

- Multiscale stochastic simulations of microbial metabolic, regulatory, and protein interaction networks
- Dynamic simulations of complex molecular machines

Software and science: Climate



Expected outcomes

5 years

- Fully coupled carbon-climate simulation
- Fully coupled sulfur-atmospheric chemistry simulation

10 years

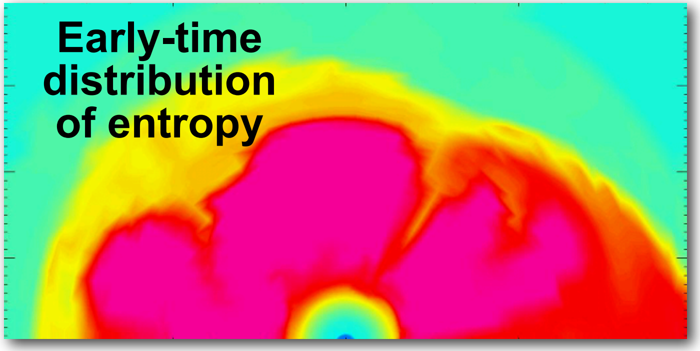
- Cloud-resolving, 30-km spatial resolution atmosphere climate simulation
- Fully coupled physics, chemistry, biology earth system model

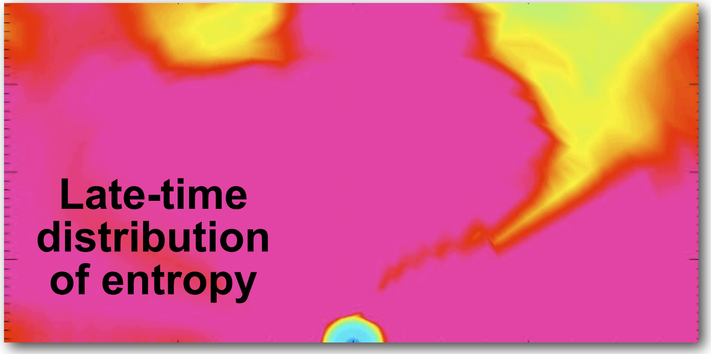
Evolution of supernovae

Supernova models must incorporate all known physics (in spite of computational demands) to capture phenomena.

- Explosions obtained for 11 and 15 solar mass progenitors
  - **recently reported at a flurry of SN1987a anniversary meetings**
- Explosions seem to be contingent on simulating **all** of the following:
  - **Multidimensional hydro**
  - **Good transport (MGFLD)**
  - **Nuclear burning**
  - **Long physical times (equivalent to long run times)**
- New result builds on earlier SASI findings
  - **Longer time scales required to observe explosion**
- Near-future simulations will include less-schematic nuclear burning, GR, and magnetic fields

Exploding core





Implementations

- Fortran still winning

- NetCDF and HDF5 use is widespread, but their parallel equivalents are used much less

- Widespread use of BLAS & LAPACK

Science Domain	Code	Programming Language	Programming Model	I/O Libraries	Math Libraries
Accelerator Design	T3P	C/C++	MPI	NetCDF	MUMPS, ParMETIS, Zoltan
Astrophysics	CHIMERA	F90	MPI	HDF5, pNetCDF	LAPACK
Biology	LAMMPS	C/C++	MPI		FFTW
Chemistry	MADNESS	F90	MPI		BLAS
Chemistry	NWChem	F77, C/C++	MPI, Global Arrays, ARMCI		BLAS, ScaLAPACK, FFTPACK
Chemistry	OReTran	F95	MPI		LAPACK
Climate	CAM	F90, C, CAF	MPI, OpenMP	NetCDF	SciLib
Climate	POP/CICE	F90, CAF	MPI, OpenMP	NetCDF	
Climate	MITgcm	F90, C	MPI, OpenMP	NetCDF	
Combustion	S3D	F90	MPI		
Fusion	AORSA	F77, F90		NetCDF	ScaLAPACK, FFTPACK
Fusion	GTC	F90, C/C++	MPI, OpenMP	MPI-IO, HDF5, NetCDF, XML	PetSC
Fusion	GYRO	F90, Python	MPI	MPI-IO, NetCDF	BLAS, LAPACK, UMFPACK, MUMPS, FFTW, SciLib, ESSL
Geophysics	PFLOTRAN	F90	MPI		BLAS, PetSC
Materials Science	LSMS	F77, F90, C/C++	MPI2	HDF5, XML	BLAS, LAPACK
Materials Science	QBOX	C/C++	MPI	XML	LAPACK, ScaLAPACK, FFTW
Materials Science	QMC	F90	MPI		BLAS, LAPACK, SPRNG
Nanoscience	CASINO	F90	MPI		BLAS
Nanoscience	VASP	F90	MPI		BLAS, ScaLAPACK
Nuclear Energy	NEWTRNX	F90, C/C++, Python		HDF5	LAPACK, PARPACK
Nuclear Physics	CCSD	F90	MPI	MPI-IO	BLAS
QCD	MILC, Chroma	C/C++	MPI		

Snapshot of Runtime and Data Rates (and other things)

Science Domain	Code	Code Attributes	Job Size (nodes, time)	Local and Archive Storage Capacity Needs	Node Memory Capacity Needs	Number of Queue Dwell Times Needed for Full Simulation
Accelerator Design	Omega3D	9 years old, 173K C++ LOC, 12 developers	128-256 24 hours	1 TB 12 TB	8 GB	3-4
Astrophysics	CHIMERA	Components 10-15 years old, 5 developers, F90	128-256 24 hours	300 GB 2 TB	≥2 GB	10-15
Climate	CCSM	Components 20 years old, 690K Fortran LOC, over 40 developers	250 24 hours	5 TB 10 TB	2 GB	10-30
Combustion	S3D	16 years old, 100K Fortran LOC, 5 developers	4000 24 hours	10-20 TB 300 TB	1 GB	7-10
Fusion	GTC	7 years old, ~30 developers	4800 24 hours	10 TB 10 TB	2 GB	4-5
Nuclear Physics	CCSD	3 years old, 10 developers, F90	200-1000 4-8 hours	300 GB 1 TB	2 GB	1

Current Models and Their Future

Science Domain	Code	Current Physical Model Attributes	Physical Model Attributes @ >1 PF
Astrophysics	Chimera	Deterministic nonlinear integro-PDEs. 63 variables.	High resolution energy and angle phase space and 200 species nuclear network. >1000 variables.
Climate	CCSM	Deterministic nonlinear PDEs. 5-10 prognostics and ~100 diagnostic variables.	Deterministic nonlinear PDEs. Could add another ~100 diagnostic variables for biogeochemical processes.
Climate	MITgcm	Deterministic nonlinear PDEs. 3 prognostic and 2 diagnostic variables.	Could add stochastic component. 5 prognostic and 1 diagnostic variables. Can vary key forcing parameters to study the response to changed climate scenarios.
Combustion	S3D	Deterministic nonlinear PDEs. 16 variables.	Deterministic nonlinear PDEs. 75 variables
Fusion	GTC	Vlasov equation in Lagrangian coordinates as ODEs, Maxwell equations in Eulerian coordinates as PDEs, and collisions as stochastic Monte Carlo processes. 2 field equations and 5 phase variables per particle.	5 field equations and 6 phase variables per particle.
Fusion	GYRO	2 field, no feedback	3 field with profile feedback

“Seven Dwarfs” (a lá Collela) Mapping

Science Domain	Code	Structured Grids	Unstructured Grids	FFT	Dense Linear Algebra	Sparse Linear Algebra	Particles	Monte Carlo
Accelerator Physics	*T3P*		*X*			*X*		
Astrophysics	*CHIMERA*	*X*			*X*	*X*	*X*	
Astrophysics	*VULCAN/2D*		*X*		*X*			
Biology	*LAMMPS*			*X*			*X*	
Chemistry	*MADNESS*		*X*		*X*			
Chemistry	*NWCHEM*			*X*	*X*			
Chemistry	*OReTran*	*X*		*X*	*X*			
Climate	*CAM*	*X*		*X*			*X*	
Climate	*POP/CICE*	*X*				*X*	*X*	
Climate	*MITgcm*	*X*				*X*	*X*	
Combustion	*S3D*	*X*						
Fusion	*AORSA*	*X*		*X*	*X*			
Fusion	*GTC*	*X*				*X*	*X*	*X*
Fusion	*GYRO*	*X*		*X*	*X*	*X*		
Geophysics	*PFLOTRAN*	*X*	*X*			*X*		
Materials Science	*QMC/DCA*				*X*			*X*
Materials Science	*QBOX*			*X*	*X*		*X*	
Nanoscience	*CASINO*						*X*	*X*
Nanoscience	*LSMS*	*X*			*X*			
Nuclear Energy	*NEWTRNX*		*X*		*X*	*X*		
Nuclear Physics	*CCSD*				*X*			
QCD	*MILC*	*X*						*X*

Translating Application Requirements to System Requirements

LC System Attribute	Application Algorithms Driving a Need for this Attribute	Application Behaviors Driving a Need for this Attribute
Node Peak Flops	*Dense Linear Algebra, FFT, Sparse Linear Algebra, Monte Carlo*	*Scalable and required spatial resolution low; would benefit from a doubling of clock speed; only a problem domain that has strong scaling, completely unscalable algorithms; embarrassingly parallel algorithms (e.g., SETI at home)*
Mean Time to Interrupt	*Particles, Monte Carlo*	*Naïve restart capability; large restart files; large restart R/W time*
WAN Bandwidth		*Community data/repositories; remote visualization and analysis; data analytics*
Node Memory Capacity	*Dense Linear Algebra, Sparse Linear Algebra, Unstructured Grids, Particles*	*High DOFs per node, multi-component/multi-physics, volume visualization, data replication parallelism, restarted Krylov subspace with large bases, subgrid models (PIC)*
Local Storage Capacity	*Particles*	*High frequency/large dumps, out-of-core algorithms, debugging at scale*
Archival Storage Capacity		*Large data (relative to local storage) that must be preserved for future analysis, for comparison, for community data (e.g., EOS tables, wind surface and ozone data, etc.); expensive to recreate; nowhere to store elsewhere*
Memory Latency	*Sparse Linear Algebra*	*Data structures with stride-one access patterns (e.g., cache-aware algorithms); random data access patterns for small data*
Interconnect Latency	*Structured Grids, Particles, FFT, Sparse Linear Algebra (global), Monte Carlo*	*Global reduction of scalars; explicit algorithms using nearest-neighbor or systolic communication; interactive visualization; iterative solvers; pipelined algorithms*
Disk Latency		*Naïve out-of-core memory usage; many small I/O files; small record direct access files;*
Interconnect Bandwidth	*Dense Linear Algebra (global), Sparse Linear Algebra (global), Unstructured Grids*	*Big messages, global reductions of large data; implicit algorithm with large DOFs per grid point;*
Memory Bandwidth	*Sparse Linear Algebra, Unstructured Grids*	*Large multi-dimensional data structures and indirect addressing; lots of data copying; lots of library calls requiring data copies; if algorithms require data retransformations; sparse matrix operations*
Disk Bandwidth		*Reads/writes large amounts of data at a relatively low frequency; read/writes lots of large intermediate temporary data; well-structured out-of-core memory usage*

What's Most Important to You?

System Attribute	Climate	Astrophysics	Fusion	Chemistry	Combustion	Accelerator Physics	Biology	Materials Science
Node Peak Flops	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Mean Time to Interrupt (MTTI)	⬜	⬜	🟨	⬜	🟨	⬜	🟨	⬜
WAN Network Bandwidth	🟨	🟨	⬜	⬜	⬜	⬜	⬜	⬜
Node Memory Capacity	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟨
Local Storage Capacity	⬜	🟨	🟨	🟩	🟩	🟨	🟩	🟨
Archival Storage Capacity	🟨	⬜	⬜	⬜	🟨	⬜	⬜	🟨
Memory Latency	🟨	🟨	⬜	🟨	⬜	🟨	⬜	🟩
Interconnect Latency	🟩	⬜	🟩	🟩	🟨	🟨	🟩	🟩
Disk Latency	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Interconnect Bandwidth	🟩	🟩	🟩	🟨	🟩	🟩	🟨	🟨
Memory Bandwidth	🟩	🟩	🟨	🟨	🟨	🟩	🟨	🟩
Disk Bandwidth	🟨	🟨	🟨	🟨	⬜	🟨	🟨	⬜

🟩 Most important

🟨 Important

⬜ Least Important

What’s Most Important to You?

System Attribute	Climate	Astrophysics	Fusion	Chemistry	Combustion	Accelerator Physics	Biology	Materials Science	
Node Peak Flops	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	←
Mean Time to Interrupt (MTTI)	⬜	⬜	🟨	⬜	🟨	⬜	🟨	⬜	
WAN Network Bandwidth	🟨	🟨	⬜	⬜	⬜	⬜	⬜	⬜	
Node Memory Capacity	⬜	🟩	🟩	🟩	🟩	🟩	🟩	🟨	←
Local Storage Capacity	⬜	🟨	🟨	🟩	🟩	🟨	🟩	🟨	
Archival Storage Capacity	🟨	⬜	⬜	⬜	🟨	⬜	⬜	🟨	
Memory Latency	🟨	🟨	⬜	🟨	⬜	🟨	⬜	🟩	
Interconnect Latency	🟩	⬜	🟩	🟩	🟨	🟨	🟩	🟩	←
Disk Latency	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	
Interconnect Bandwidth	🟩	🟩	🟩	🟨	🟩	🟩	🟨	🟨	←
Memory Bandwidth	🟩	🟩	🟨	🟨	🟨	🟩	🟨	🟩	
Disk Bandwidth	🟨	🟨	🟨	🟨	⬜	🟨	🟨	⬜	

🟩 Most important

🟨 Important

⬜ Least Important

Scaling Disk Bandwidth

Variable	Description	Typical Values
M	Total system memory	100-400 TB for a 1 PF system
f	Fraction of application runtime memory captured and written out per restart	20-80%
T	Runtime intervals between successive restart file outputs	1-3 hours when MTTI is 12-24 hours or maximum queue runtimes ~24 hours
O	Maximum allowable fraction of total runtime devoted to I/O operations	10%
B	Required bandwidth to local storage	$= \frac{fM}{TO}$

Restart File Size / Total System Memory	Restart Period (hours)	Allowable I/O Overhead	Required Local Storage Bandwidth (GB/s)
0.10	1	5% 10%	111 56
	2	5% 10%	56 28
0.20	1	5% 10%	222 111
	2	5% 10%	111 56
0.80	1	5% 10%	888 444
	2	5% 10%	444 222

Scaling Total Disk Store

Variable	Description	Typical Values
M	*Total system memory*	*100-400 TB for a 1 PF system*
P	*Total number of projects with LC allocations annually*	*20-40*
F	*Fraction of application runtime memory captured and written out per restart*	*20-80%*
R	*Average number of simulations per project whose output is retained on local storage*	*10-20*
C	*Required local storage capacity*	$= fMPR$

Number of Projects	Restart File Size / Total System Memory	Number of Runs Per Project Retained on Local Storage	Required Local Storage Capacity (PB)
10	*0.20*	*2* *10*	*0.8* *4.0*
	0.80	*2* *10*	*3.2* *16.0*
20	*0.20*	*2* *10*	*1.6* *8.0*
	0.80	*2* *10*	*6.4* *32.0*
40	*0.20*	*2* *10*	*3.2* *16.0*
	0.80	*2* *10*	*12.8* *64.0*

Planned total disk store in 2008: 10 PB

Another prescription…

- Analysis by Shane Cannon & Sarp Oral (NCCS Technology Integration)
- Scale current usage with projected future system attributes
- Total memory of system, memory bandwidth, and total peak FLOP rate are, e.g., attributes that might provide scale of increased needs
- For example, scaling for **archival storage**:

System Attribute Assumed to Govern Archival Storage Requirements	*Estimated Capacity Needs by end of CY06*	*Estimated 250 TF Capacity Needs*	*Estimated 1 PF Capacity Needs*
Memory	*2.8 PB*	*4.6 PB*	*15.9 PB*
Memory Bandwidth	*3.8 PB*	*10.8 PB*	*36.0 PB*
Peak Flop Rate	*3.6 PB*	*7.1 PB*	*18.5 PB*

Planned HPSS capacity in 2009: 18 PB

Summary

- Current best understanding points to a widespread need for a balanced system (FLOPs, memory, memory bandwidth)
  - **There is a strong call for large memory!**
- Most application teams know the algorithms and the implementations they plan to use for the foreseeable future
- No one kind of algorithm dominates our portfolio, i.e. we shouldn’t deploy a GRAPE farm
- I/O rates and system reliability (MTTI) could make or break a lot of science.
- We need more data!