Argonne's Leadership Computing Facility: Petascale Computing for Science

Rick Stevens

Argonne National Laboratory

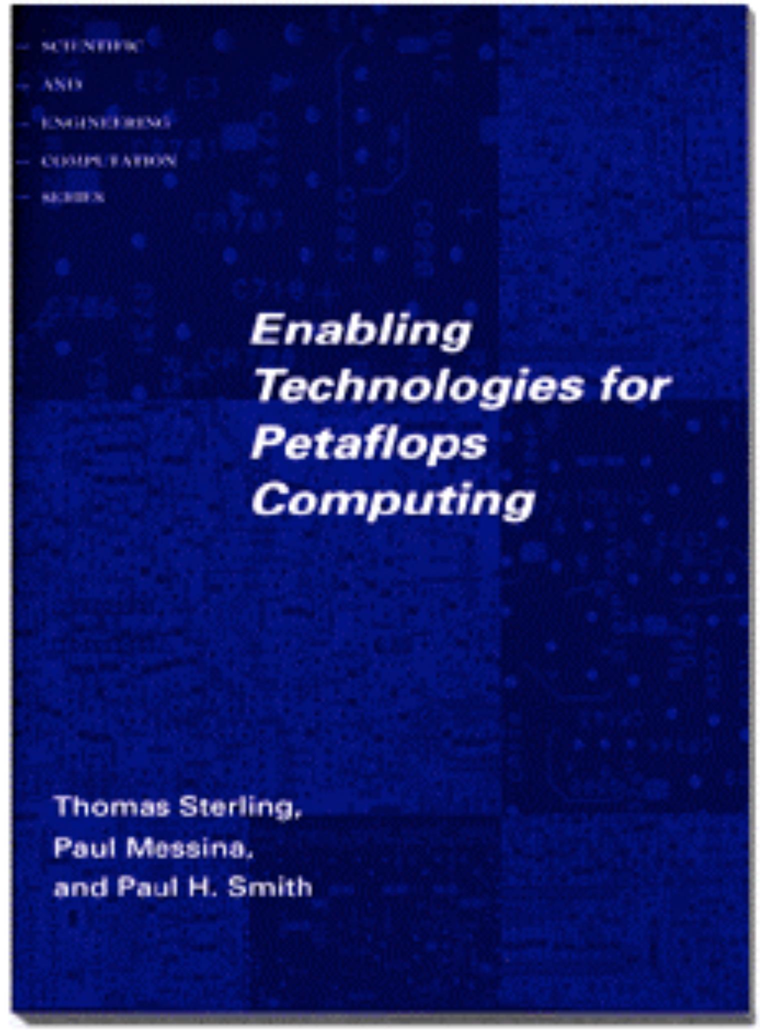
The University of Chicago

THE UNIVERSITY OF CHICAGO
Office of Science
U.S. DEPARTMENT OF ENERGY

Argonne National Laboratory is managed by The University of Chicago for the U.S. Department of Energy

A Brief History of Petaflops Computing

- 1994 Petaflops I (Pasadena)
- 1995 Summer Study (Bodega)
- 1996 Architecture Workshop (Bodega)
- 1996 Software Workshop (Bodega)
- 1996 Petaflops Frontier 2 (Annapolis)
- 1997 Layered SW Workshop (Oxnard)
- 1997 Algorithms Workshop (Williamsburg)
- 1998 Petaflops-sys Operations Workshop
- 1999 Petaflops II (Santa Barbara)
- 2002 WIMPS (Bodega)
- 2003 HECRTF Roadmap (Washington)



National community has been engaged for more than a decade on the problem of petascale computing

Desired Modes of Impact for Petascale Computing

1. Generation of significant datasets via simulation to be used by a large and important scientific community
   - Providing a high-resolution first principles turbulence simulation dataset to the CFD and computational physics community
2. Demonstration of new methods or capabilities that establish feasibility of new computational approaches that are likely to have significant impact on the field
   - Demonstration of the design and optimization of a new catalyst using first principles molecular dynamics and electronic structure codes
3. Analysis of large-scale datasets not possible using other methods
   - Computationally screen all known microbial drug targets against the known chemical compound libraries
4. Solving a science or engineering problem at the heart of a critical DOE mission or facilities design or construction project
   - Designing a passively safe reactor core for the Advanced Burner Reactor Test Facility

DOE Leadership Computing Facility Strategy

- DOE selected the ORNL, ANL and PNNL team (May 12, 2004) based on a competitive peer review of four proposals to develop the DOE SC Leadership Computing Facilities
  - ORNL will develop a series of systems based on Cray's XT3 and XT4 architectures with systems @ 250TF/s in FY07 and @1000TF/s in FY08/FY09
  - ANL will develop a series of systems based on IBM's BlueGene @ 100TF/s in FY07 and up to 1000TF/s in FY08/FY09 with BG/P
  - PNNL will contribute software technology for programming models (Global Arrays) and parallel file systems
  - The Leadership Class Computing (LCC) systems are likely to be the most powerful civilian systems in the world when deployed
- DOE SC will make these systems available as capability platforms to the broad national community via competitive awards (e.g. INCITE and LCC Allocations)
  - Each facility will target ~20 large-scale production applications teams
  - Each facility will also support order 100 development users
- DOE's LCC facilities will complement the existing and planned production resources at NERSC
  - Capability runs will be migrated to the LCC, improving NERSC throughput
  - NERSC plays an important role in training and new user identification

Why Blue Gene?

- In the National Leadership Computing Facility proposal the ORNL, ANL, PNNL, et. al. team proposed a multi-vendor strategy to achieve national leadership capabilities
- Possible systems capable of 500TF to 1 PF peak performance deployable in FY08/FY09
  - Cray XT3/XT4, IBM Power5/6, IBM Blue Gene L/P
  - Clusters (Intel, AMD, PPC, Cell?)
  - DARPA HPCS design points considered but not available in time
- Decision factors
  - Suitable for DOE applications ⇒ adequate coverage
  - Feasibility demonstrated at scale ⇒ acceptable level of risk
  - Acceptable reliability ⇒ user acceptance and operational efficiency
  - Acceptable power consumption ⇒ acceptable TCO
  - Cost ⇒ acceptable TPC

Leadership Science Platform Mix

- Assumptions
  - DOE will invest in multiple platforms, to avoid risk and unneeded duplication of specific capabilities
  - Users will migrate to platforms were they can get the most science for the least effort
  - We have limited ability to predict the success and ultimate adoption of unfielded systems
  - More specialized (limited application suitability) systems will need to have a cost (TCO) advantage to add value to the fleet of systems
  - The lower the overall risk to the program the better

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.

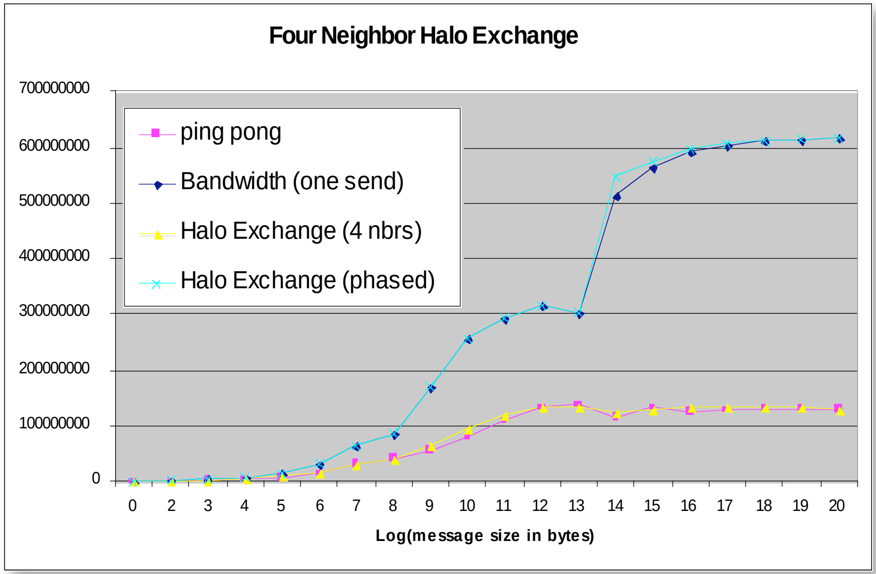
Theory

System Scale TFs	*CPU Type*	*MTBF (Days)*	*Failures per Month per System*	*Failures per Month per TF*
3	IA64	1.3	24	8
10.7	IA64	1.1	28.3	2.7
1.7	x86	4.5	6.7	3.9
17.2	x86	0.7	45.1	2.6
15	Power 5	1.1	19	1.3
114	Blue Gene	6.9	4.3	0.038
365	Blue Gene	7.5	4	0.011
1000	Blue Gene P	7	4.3	0.004

Experiment

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

Decision to choose Blue Gene is Supported by

- Blue Gene has been fielded within a factor of 3 of PF goal
  - *No other system is close to this scale (lower risk to scale to PF)*
- Applications community has reacted positively, though the set is still limited it is larger than expected, and some applications are doing extremely well
  - *For those applications that can make the transition, the BG platform provides outstanding scientific opportunity, many can, some can't*
- Blue Gene has been remarkably reliable at scale
  - *The overall reliability appears to be several orders of magnitude better than other platforms for which we have data*
- Power consumption is 2x-4x better than other platforms
  - *Lower cost of ownership and window to the future of lower power*
- System Cost
  - *The cost of the system is significantly lower than other platforms*

BlueGene/P has a strong family resemblance

- Puts processors + memory + network interfaces on the same chip

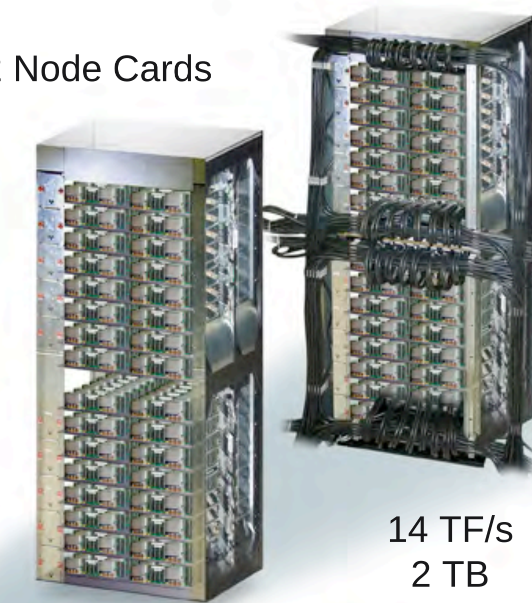
System
72 Racks

1 PB/s
144 TB

Rack

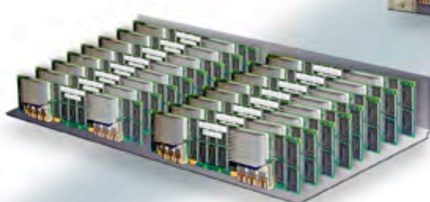
32 Node Cards

Cabled 8x8x16



14 TF/s
2 TB

Node Card
(32 chips 4x4x2)
32 compute, 0-4 IO cards



435 GF/s
64 GB

Compute Card
1 chip, 1x1x1

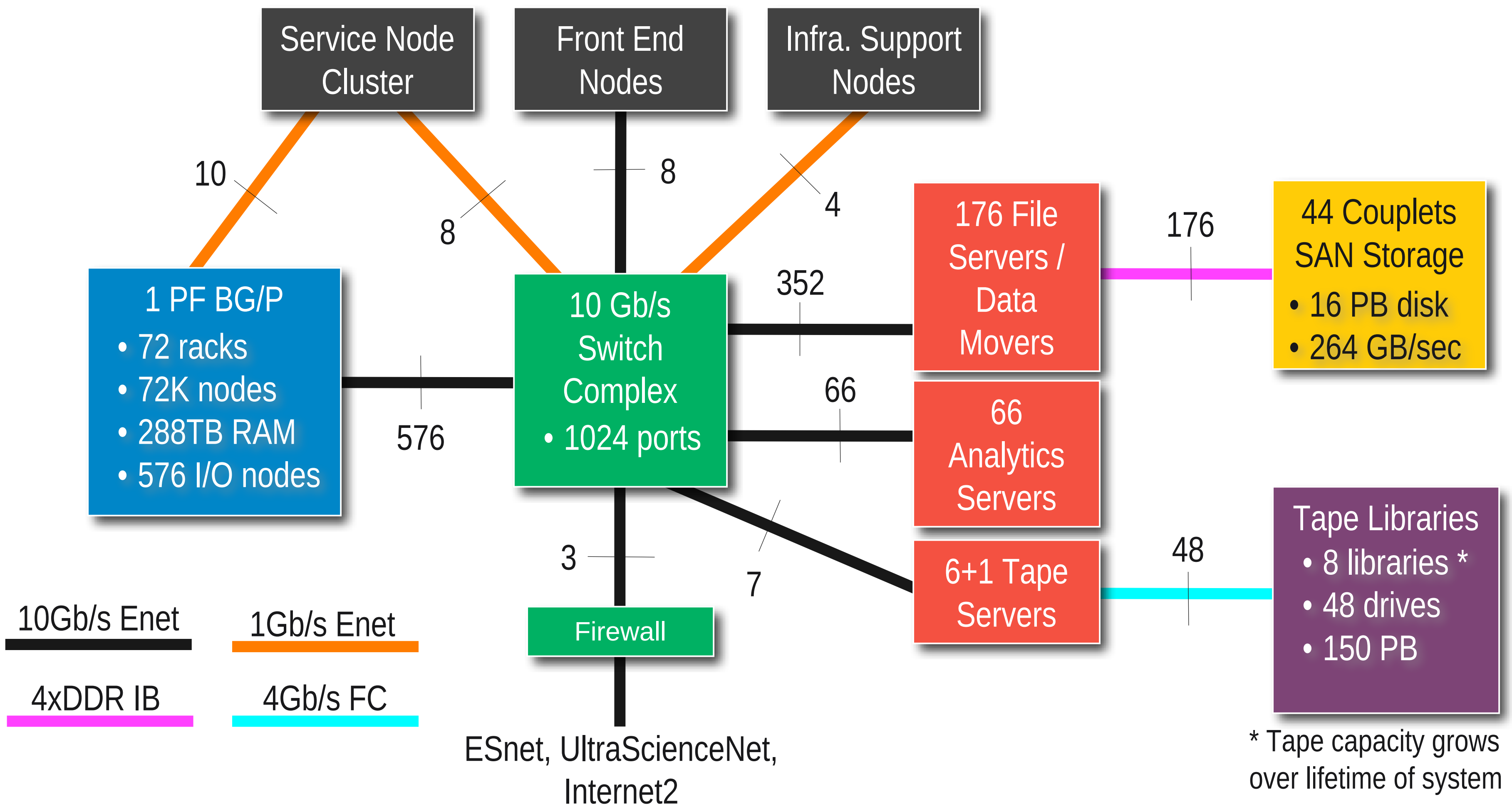
13.9 GF/s
2 GB DDR

Chip
4 processors

13.6 GF/s
8 MB EDRAM

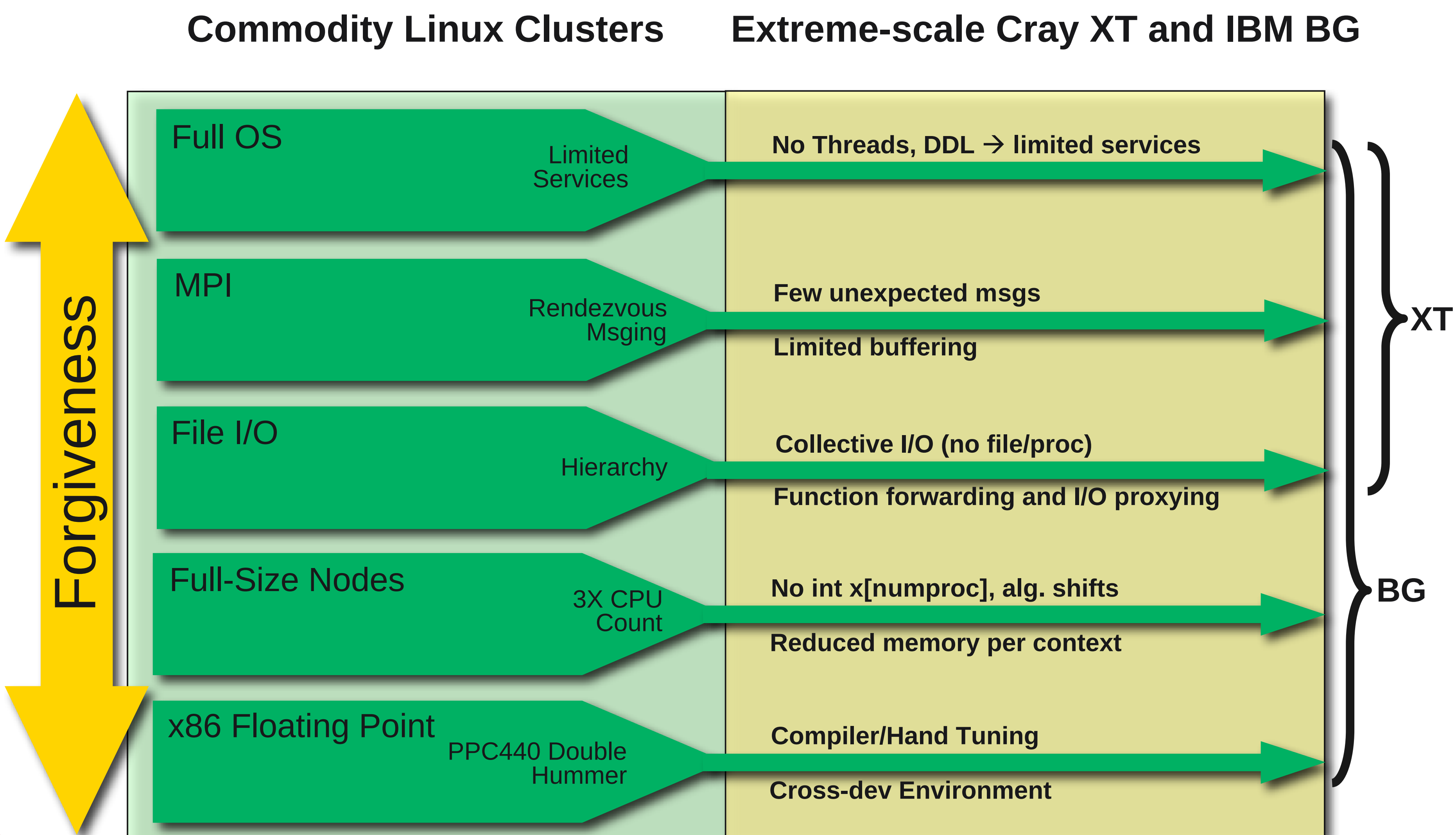
- High packaging density
  - 13.9TF in 15 sq ft of floor space (1 rack)
- Low system power requirements
  - 31KW per rack

Petascale System Architecture



* Tape capacity grows over lifetime of system

Challenges and Choices to Achieve Leadership-Class Capability



Software Environment

Compute / Development

Community

IBM/Vendor

Resource Mgmr / Scheduler / Workflow	Cobalt, Kepler
User Mgmt, Ticket system, Accounting	ANL UserBase/Accting System
Other compilers, IDEs	UPC, Eclipse
IBM Math Libraries, Tools, Compilers	ESSL, MASS/V, HPC Toolkit, IBM xl*
Community Math Libraries	FFTW, PETSc, BLAS, LAPACK
Performance, & Debugging Tools	TAU, Kojak, PAPI
Parallel I/O Libraries	HDF5, pNetCDF
MPI, MPI-IO, GAs	MPICH, ROMIO, ARMCI
Low-level MSG Layer & Collectives	IBM, MPICH Nemesis
Low-level HW Drivers	IBM
CN & ION Kernels; CIOD	ZeptoOS (Linux) and ZOID, IBM CN and coid
Home Directory File System	GPFS

Blue Gene Applications Analysis Strategy

- Over 80 applications have been ported to BG/L
- In many cases the application runs within 1 or 2 days
- Typical issues
  - **Memory footprint** [512MB node on BG/L ⇒ 4GB node on BG/P]
  - **Scalability** [impact of collectives, torus loading, load balancing, I/O]
  - **Libraries** [FFT, BLACS, etc.]
  - **Single node performance** [compiler optimization, double hummer]
  - **Memory hierarchy management** [blocking, prefetch, fusing ops, etc.]
- Initial tests are done to confirm correctness, then weak scaling and then strong scaling limits determined, etc.
  - Work then focuses on improving scaling and performance
- We believe applications are self-selecting for BG
  - Highly portable, well understood codes, aggressive user/developers
- In a multi-architecture DOE environment we believe user driven application self-selection is the most efficient path forward
- Due to the effort required to achieve leadership level performance we believe general HPC benchmarks are of extremely limited utility

Example Applications Ported to BG/L

- The following lists codes ported to date on BG/L evidencing the strong community interest and potential scientific ROI.

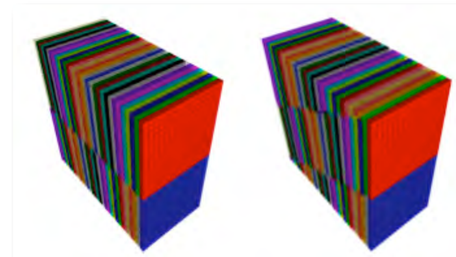
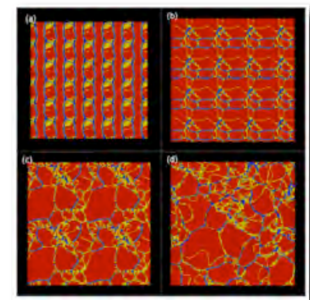
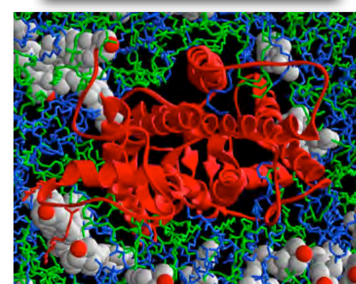
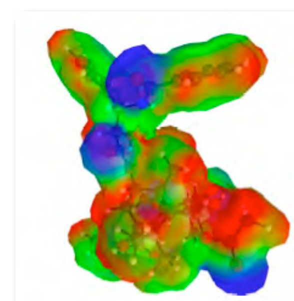
General Domain	Code	Institution	General Doman	Code	Institution
Astro Physics	**Enzo**	UCSD/SDSC	Material Sciences	**ALE3D**	LLNL
Astro Physics	**Flash**	UC/Argonne	Material Sciences	**LSMS**	LLNL
Basic Physics	**CPS**	Columbia	Molecular Biology	**mpiBLAST**	Argonne
Basic Physics	**QCD kernel**	IBM	Molecular 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 Stucture	**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	**TLBE**	LLNL	Molecular dynamics	**Paradis**	LLNL
CFD	**sPPM**	LLNL	Nano-Chemistry	**DL_POLY**	Argonne
CFD	**mpcugles**	LLNL	Neuroscience	**pNEO**	Argonne
CFD	**Nek5**	Argonne	neutron transport	**SWEEP3D**	LArgonne
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 propogatio	**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

DOE Applications Drivers and Example Codes

- Computational Materials Science and Nanoscience
  - Electronic structure, First Principles ⇒ Qbox, LSMS, QMC
  - (mat) Molecular dynamics ⇒ CPMD, LJMD, ddcMD, MDCASK
  - Other materials ⇒ ParaDIS
- Nuclear Energy Systems
  - Reactor core design and analysis ⇒ NEK5, UNIC
  - Neutronics, Materials, Chemistry ⇒ QMC, Sweep3D, GAMESS
- Computational Biology/Bioinformatics
  - (bio) Molecular dynamics ⇒ NAMD, Amber7/8, BlueMatter
  - Drug Screening ⇒ DOCK5, Autodock
  - Genome-analysis ⇒ mpiBLAST, mrBayes, CLUSTALW-mpi
- Computational Physics and Hydrodynamics
  - Nuclear Theory ⇒ GFMC
  - Quantum chromo dynamics ⇒ QCD, MILC, CPS
  - Astrophysics/Cosmology ⇒ FLASH, ENZO
  - Multi-Physics/CFD ⇒ ALE3D, NEK5, Miranda, SAGE

Example Leadership Science Applications

- **Qbox** — FPMD solving Kohn-Sham equations, strong scaling on problem of 1000 molybdenum atoms with 12,000 electrons (86% parallel efficiency on 32K cpus @ SC05), achieved 190 TFs recently on BG/L
- **ddcMD** — many-body quantum interaction potentials (MGPT), 1/2 billion atom simulation, 128K cpus, achieved > 107 TFs on BG/L via fused dgemm and ddot
- **BlueMatter** — scalable biomolecular MD with Lennard-Jones 12-6, P3ME and Ewald, replica-exchange 256 replicas on 8K cpus, strong scaling to 8 atoms/node
- **GAMESS** — *ab initio* electronic structure code, wide range of methods, used for energetics,spectra, reaction paths and some dynamics, scales O(N^5-N^7) in number of electrons, uses DDI for communication and pseudo-shared memory, runs to 32,000 cpus
- **FLASH3** — produced largest weakly- compressible, homogeneous isotropic turbulence simulation to date on BG/L, excellent weak scaling, 72 million files 156 TB of data

Communication Needs of the "Seven Dwarves"

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

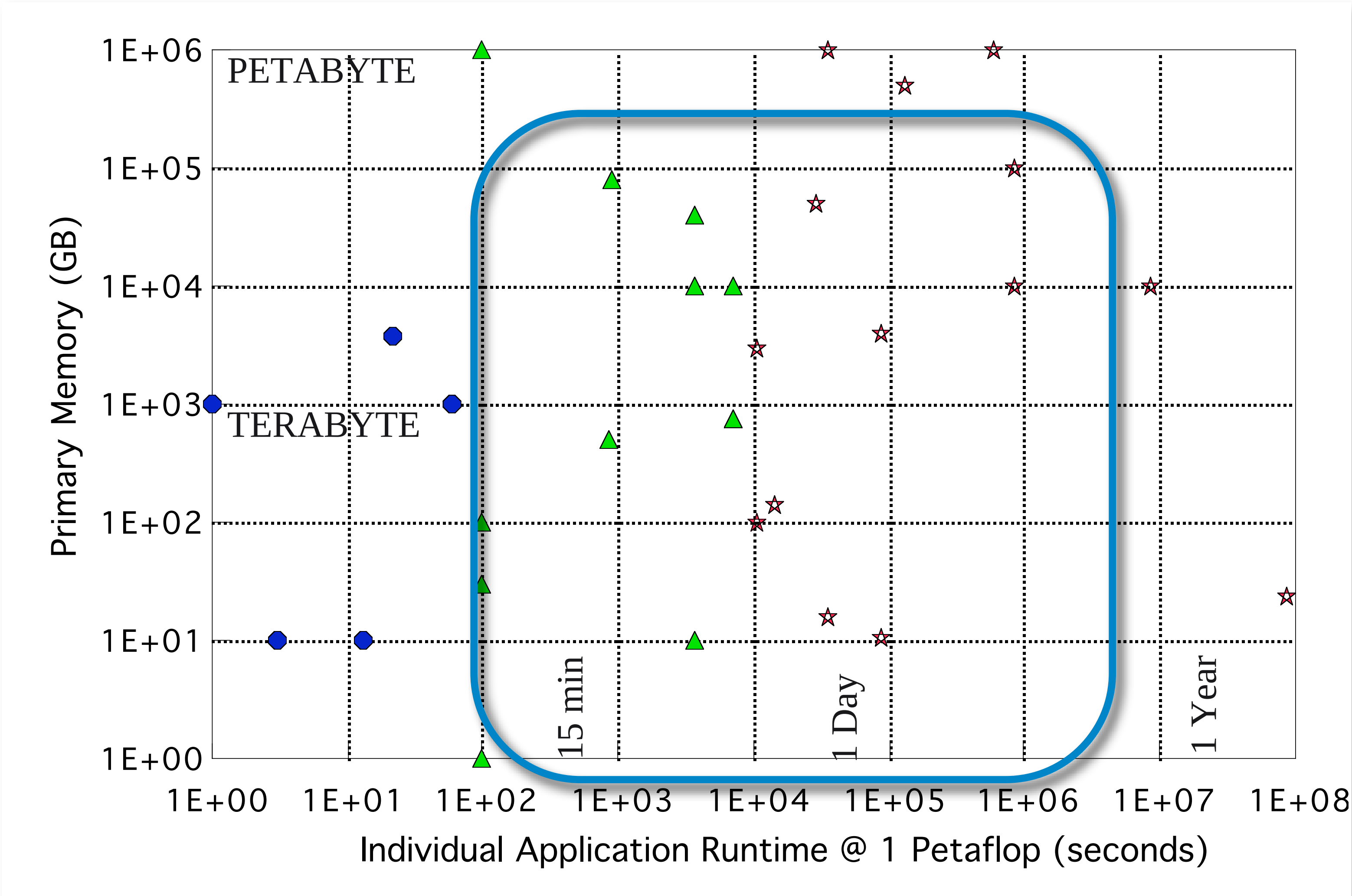
1. **Molecular dynamics (mat)**
2. **Electronic structure**
3. **Reactor analysis/CFD**
4. **Fuel design (mat)**
5. **Reprocessing (chm)**
6. **Repository optimizations**
7. **Molecular dynamics (bio)**
8. **Genome analysis**
9. **QMC**
10. **QCD**
11. **Astrophysics**

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

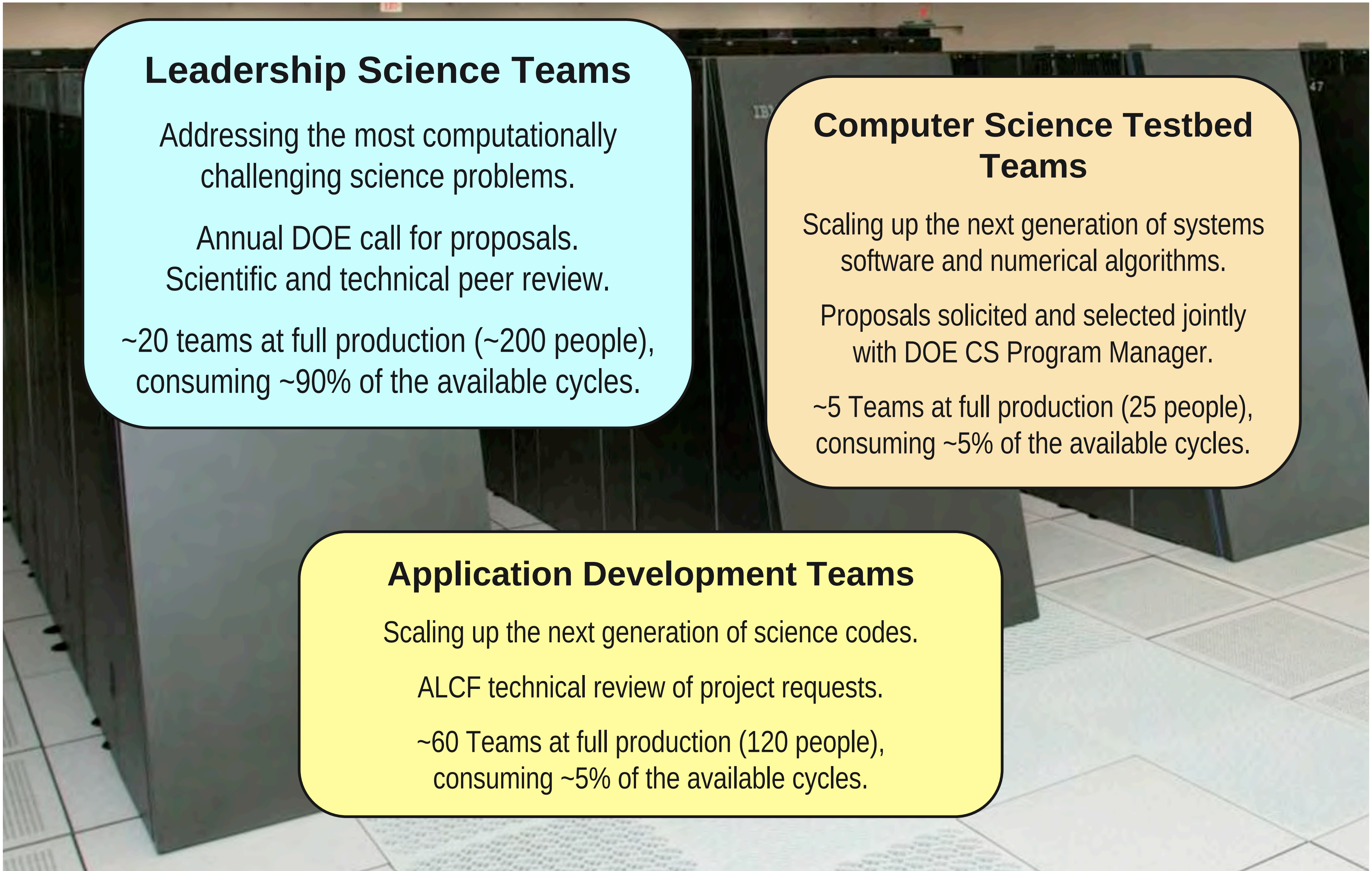
Petaflops Applications Coverage



Scalable Software Testbed

- The ALCF BG system provides a unique opportunity for the computer science community to test ideas for next generation operating systems and scalable systems software
- ALCF could allocate a fraction (up to 5%) for competitively awarded computer science proposals aimed at advancing petascale software projects
- ALCF will be configured to permit testbed users to try new operating systems and file systems
- It is anticipated that the software environment on the ALCF will be open source and available to the development community for enhancement and improvement

ALCF Science Community



Leadership Science Teams

Addressing the most computationally challenging science problems.

Annual DOE call for proposals.
Scientific and technical peer review.

~20 teams at full production (~200 people), consuming ~90% of the available cycles.

Computer Science Testbed Teams

Scaling up the next generation of systems software and numerical algorithms.

Proposals solicited and selected jointly with DOE CS Program Manager.

~5 Teams at full production (25 people), consuming ~5% of the available cycles.

Application Development Teams

Scaling up the next generation of science codes.

ALCF technical review of project requests.

~60 Teams at full production (120 people), consuming ~5% of the available cycles.