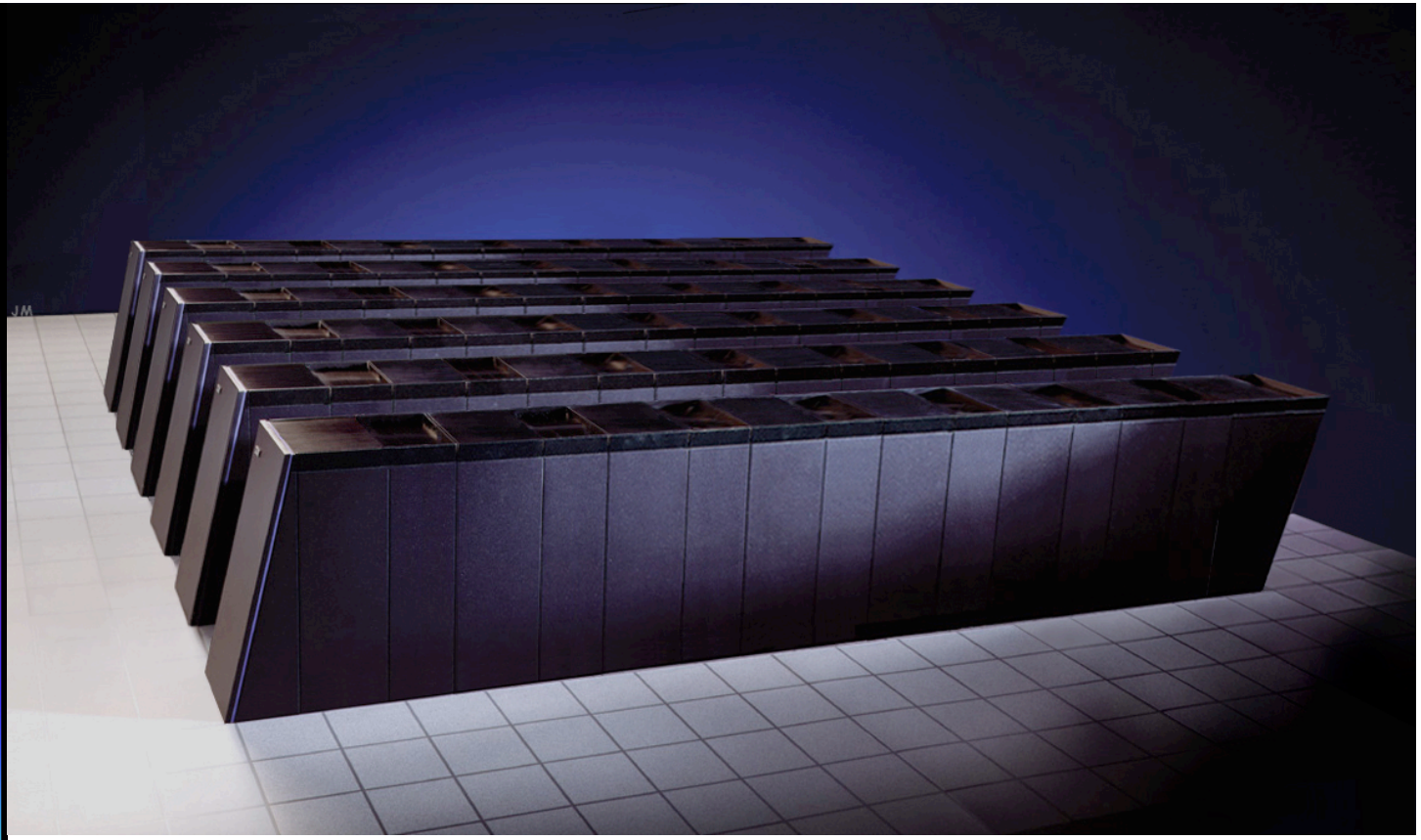
Argonne NATIONAL LABORATORY

... for a brighter future

ALCF

Argonne Leadership Computing Facility



Argonne Leadership Computing Facility

Pete Beckman

Chief Architect
Argonne Leadership Computing Facility

U.S. Department of Energy

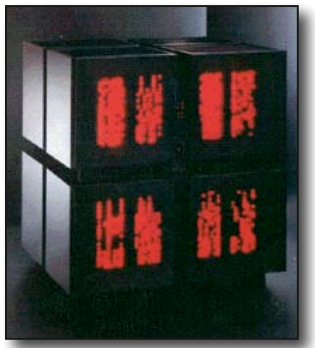
UChicago ▶ Argonne LLC

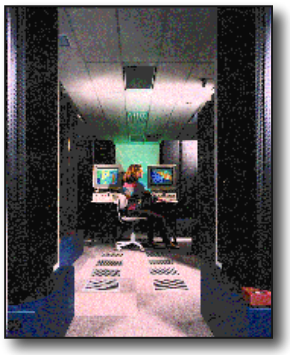
Office of Science
U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

Over 20 years of Advanced Systems for DOE and Others

- **ACRF** period [1983-1992]
  - DOE's founding ACRF
  - Explored many parallel architectures, developed programming models and tools, trained >1000 people
- **HPCRC** period [1992-1999]
  - Production-oriented parallel computing for Grand Challenges in addition to Computer Science.
  - Fielded 1st IBM SP in DOE





- **TeraGrid** [2001-present]
  - Overall Project Lead
  - Defining, deploying and operating the integrated national cyberinfrastructure for NSF
  - 9 sites, 22 systems, 200TF
- **LCRC** [2003-present]
  - Lab-wide production supercomputer service
  - All research divisions, 56 projects, 380 users
- **BlueGene Evaluation** [2005-present]
  - Founded BlueGene Consortium with IBM
    - *67 institutions, >260 members*
    - *Applications Workshop Series*
    - *Systems Software Collaborations*

DOE Leadership Computing Facility (LCF) Strategy

- DOE SC selected the ORNL, ANL and PNNL teams (May 12, 2004) based on a competitive peer review
  - ORNL will deploy a series of systems based on Cray's XT3/4 architectures @ 250TF/s in FY07 and 1000TF/s in FY08/09
  - ANL will develop a series of systems based on IBM's BlueGene @ 100TF/s in FY07 and 250-500TF/s in FY08/FY09 with IBM's Next Generation Blue Gene
  - PNNL will contribute software technology
- DOE SC will make these systems available as capability platforms to the broad national community via competitive awards (e.g. INCITE Allocations)
  - Each facility will target ~20 large-scale production applications teams
  - Each facility will also support development users
- DOE's LCFs complement existing and planned production resources at NERSC
  - Capability runs will be migrated to the LCFs, improving NERSC throughput
  - NERSC will play an important role in training and new user identification

Mission and Vision for the ALCF

Our Mission

Provide the computational science community with a world leading computing capability dedicated to breakthrough science and engineering.

Our Vision

A world class center for computation driven scientific discovery that has:

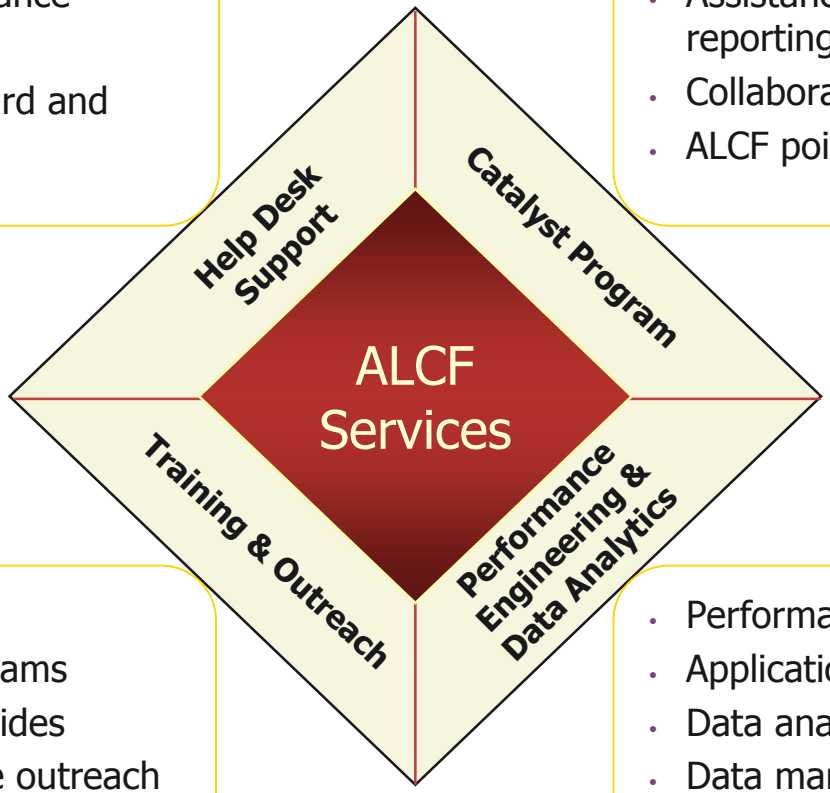
- outstandingly talented people,
- the best collaborations with computer science and applied mathematics,
- the most capable and interesting computers and,
- a true spirit of adventure.

See http://www.alcf.anl.gov/ for additional information

ALCF Service Offerings

- Startup assistance
- User administration assistance
- Job management services
- Technical support (Standard and Emergency)

- ALCF project management
- Assistance with proposals, planning, reporting
- Collaboration within science domains
- ALCF point of coordination



- Workshops & seminars
- Customized training programs
- On-line content & user guides
- Educational and corporate outreach programs

- Performance engineering
- Application tuning
- Data analytics
- Data management services

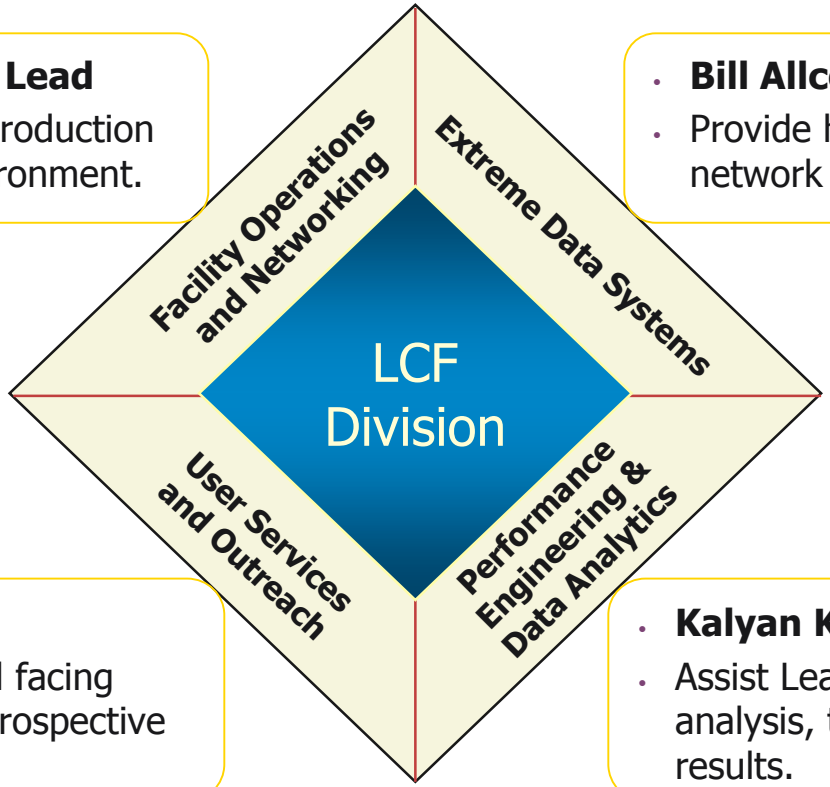
Catalyst Program

What is it?	ALCF Program to establish strategic collaborations with our leading project partners to maximize benefits from the use of various ALCF resources
What are the program features?	■ Full project lifecycle approach to collaboration with our project partners ■ Provide value-added services and support in conjunction with ALCF HW and SW resources ■ Tailor services-delivery program for the unique requirements of each research initiative ■ Maintain close contact with research teams through ongoing interactions with assigned ALCF Project Coordinator
What are the key objectives?	■ 'Jump-start' the use of ALCF resources for major ALCF projects ■ Align availability of ALCF services and compute resources with the needs of researchers through joint project planning based on research goals and timing objectives ■ Establish a spirit of collaboration to maximize the value that ALCF can bring to our project partners

ALCF Organization

- **Susan Coghlan, Acting Lead**
- Provide a stable, secure production capability computing environment.

- **Bill Allcock, Technical Lead**
- Provide high performance data and network services.



- **Paul Davé, Manager**
- Provide the ALCF outward facing interface to current and prospective users and stakeholders.

- **Kalyan Kumaran, Manager**
- Assist Leadership Projects with code analysis, tuning and analyzing their results.

ALCF Timeline

2004

- Formed of the Blue Gene Consortium with IBM
- DOE-SC selected the ORNL, ANL and PNNL teams for Leadership Computing Facility award

2005

- Installed 5 teraflops Blue Gene/L for evaluation

2006

- Began production support of 6 INCITE projects, with BGW
- Joined IBM and LLNL to design and develop BG/P & BG/Q

2007

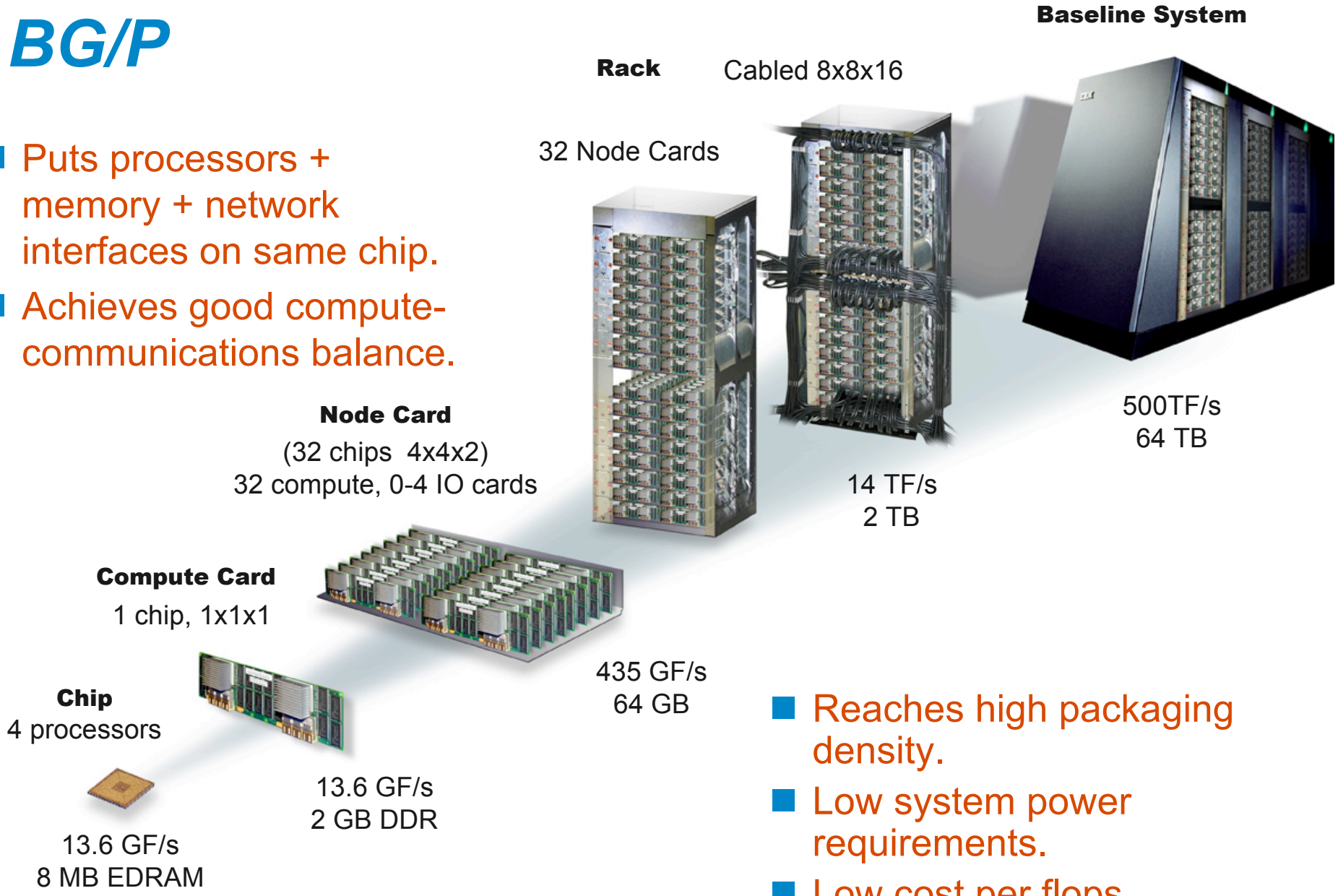
- Increased to 9 INCITE projects; continued development projects
- Install 100 teraflops next gen Blue Gene system (late 2007)

2008

- Begin support of INCITE projects on next gen Blue Gene
- Add 250-500T teraflops Blue Gene system

BG/P

- Puts processors + memory + network interfaces on same chip.
- Achieves good compute-communications balance.



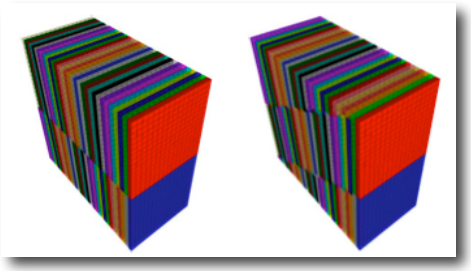
- Reaches high packaging density.
- Low system power requirements.
- Low cost per flops.

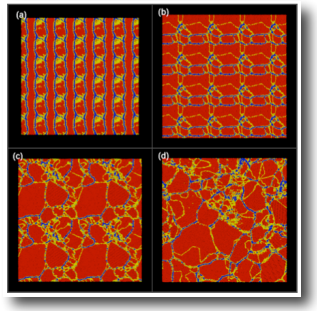
Why We Like Blue Gene

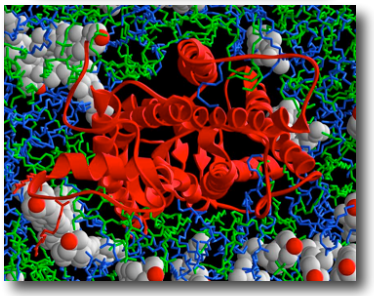
- Blue Gene has been fielded within a factor of 3 of PF goal
  - *No other system is close to this scale (lower risk to scale to PF)*
- Applications community has reacted positively, though the set of codes is still limited, but is larger than expected, and some applications are doing extremely well
  - *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/TF appears to be at least an order of magnitude better than other platforms for which we have data*
- Power consumption is 2x better than other platforms
  - *Lower cost of ownership and window to the future of lower power*
- System Cost
  - *The cost of deploying a balanced system is lower than other platforms*

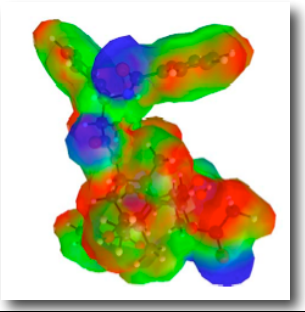
Example Blue Gene Science Applications

- **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 207 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









26 BlueGene/L Systems on 11/06 TOP500 List

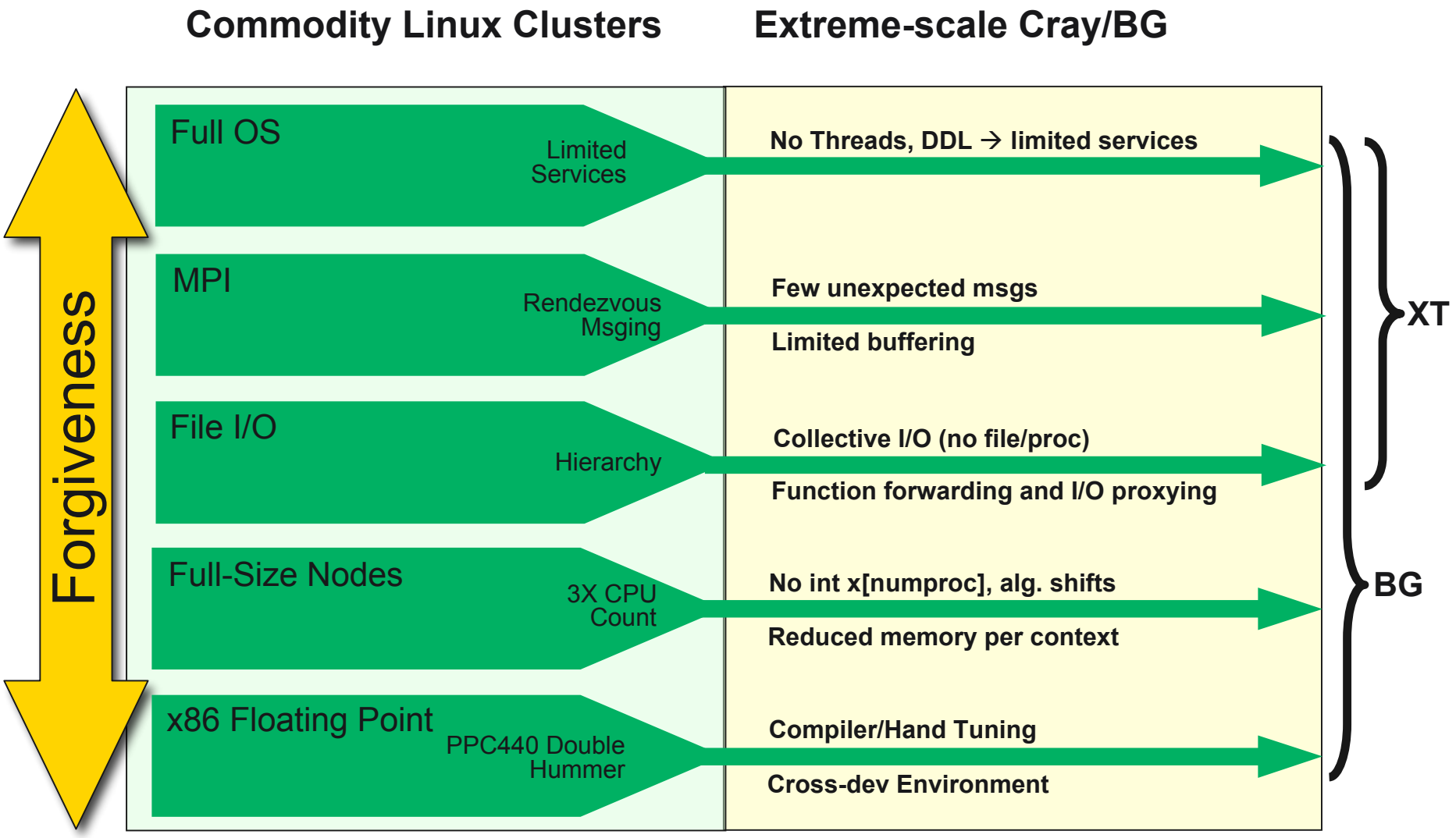
819 teraflops fielded today, with 294,912 processors

Rank	Site	Country	Processors	RMax	RPeak
1	DOE/NNSA/LLNL	United States	131,072	280,600	367,000
3	IBM Thomas J. Watson Research Center	United States	40,960	91,290	114,688
13	Forschungszentrum Juelich (FZJ)	Germany	16,384	37,330	45,875
17	ASTRON/University Groningen	Netherlands	12,288	27,450	34,406
21	Computational Biology Research Center, AIST	Japan	8,192	18,200	22,938
22	Ecole Polytechnique Federale de Lausanne	Switzerland	8,192	18,200	22,938
23	High Energy Accelerator Research Org (KEK)	Japan	8,192	18,200	22,938
24	High Energy Accelerator Research Org (KEK)	Japan	8,192	18,200	22,938
25	IBM - Rochester Deep Computing	United States	8,192	18,200	22,938
42	UCSD/San Diego Supercomputer Center	United States	6,144	13,780	17,203
52	IBM - Rochester DD1 Prototype	United States	8,192	11,680	16,384
63	High Energy Accelerator Research Org (KEK)	JAPAN	4,096	9,360	11,469
64	IBM - Almaden Research Center	United States	4,096	9,360	11,469
65	IBM Research	Switzerland	4,096	9,360	11,469
66	IBM Thomas J. Watson Research Center	United States	4,096	9,360	11,469
138	Argonne National Laboratory	United States	2,048	4,713	5,734
139	Boston University	United States	2,048	4,713	5,734
140	CERT (Center for Excellence for Applied R&T)	UAR	2,048	4,713	5,734
141	Iowa State University	United States	2,048	4,713	5,734
142	Lawrence Livermore National Laboatory	United States	2,048	4,713	5,734
143	MIT	United States	2,048	4,713	5,734
144	NCAR (National Center for Atmospheric Research)	United States	2,048	4,713	5,734
145	NIWS Co, Ltd	Japan	2,048	4,713	5,734
146	Princeton University	United States	2,048	4,713	5,734
147	Renaissance Computing Institute (RENCI)	United States	2,048	4,713	5,734
148	University of Edinburgh	United Kingdom	2,048	4,713	5,734

Unique and Challenging Features

- Relatively low memory per CPU core (but very large aggregate).
  - BG/P doubles RAM per core (2 Gig / 4 cores)
  - True SMP is possible (sharing data structures)
- Single node optimization
  - Use of “double hummer” requires lots of hand tuning and compiler experimentation
  - Strategies: Good math libs, perf counters, ATLAS, code tools
- Non-linux kernel makes development difficult (see ZeptoOS)
- Scalable I/O strategy required
  - One file per process **strongly** discouraged
    - *Who really wants 300K files per snapshot, with **millions** per large scale run?*
  - PnetCDF and HDF5 are good strategies for effectively using parallel storage system (PVFS)
- Debugging at scale remains challenging
  - Tool groups are helping, but the issue is nevertheless hard

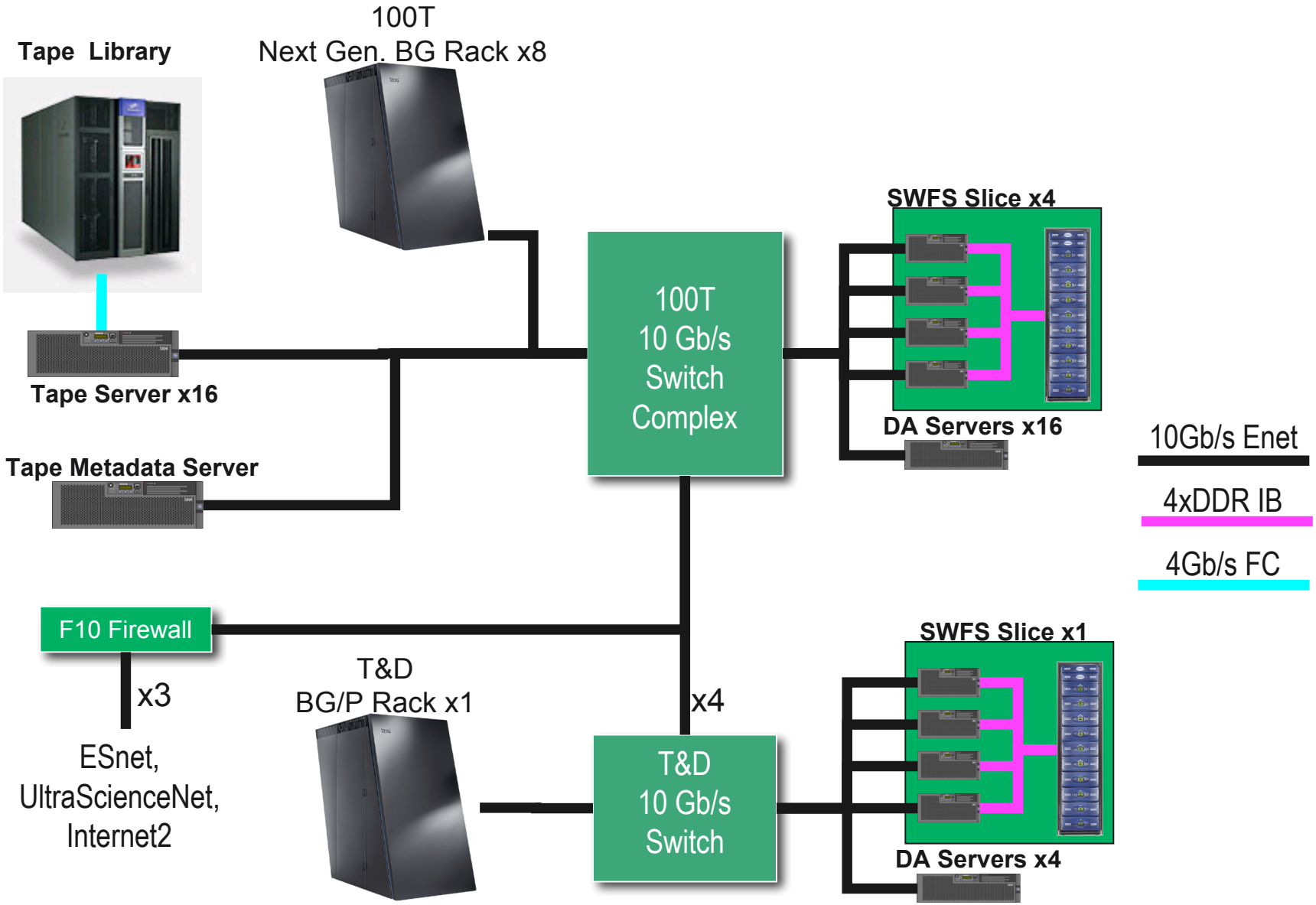
Challenges and Choices to Achieve Leadership-Class Capability



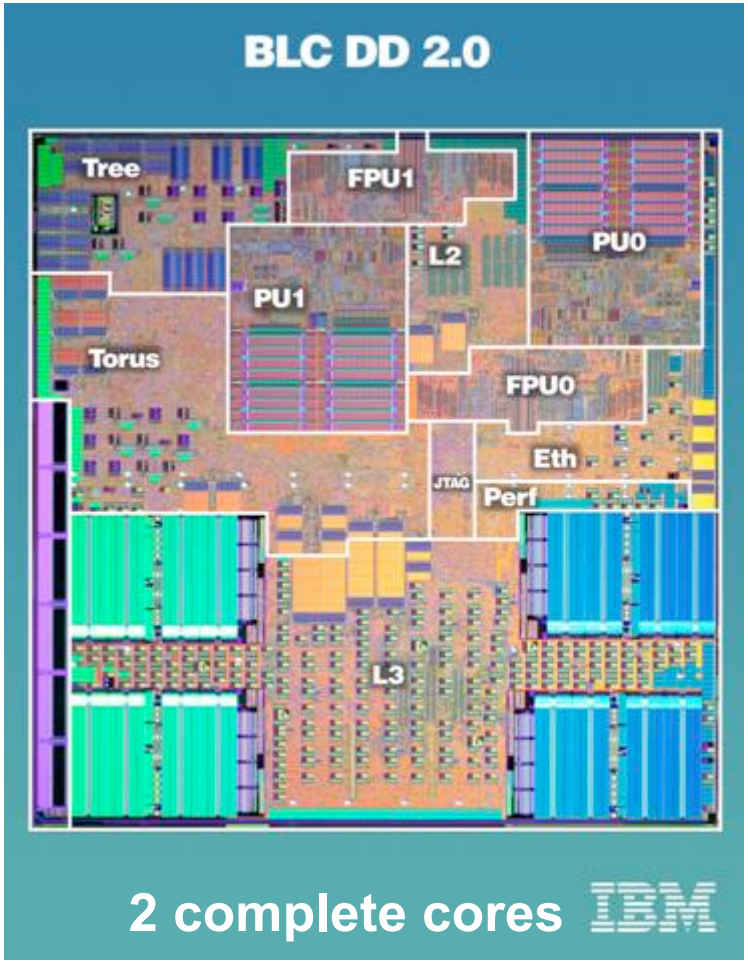
Understanding Blue Gene Usability

- Blue Gene/L is a “PureMPI” software environment
- Current kernel limits the use of some technologies:
  - DLLs, sockets, Java, threads, fork(), Python, etc
- ***BG/P will improve functionality***
- Porting ultra-scale PureMPI programs is generally straightforward
  - The IBM xl* you know and love
  - Some *build* environments can be complicated (cross-compilers)
  - Reduced memory per node and sheer node count forces some algorithmic changes
  - Achieving good double-hummer requires extra work

ALCF Q3 FY2008 *At start of INCITE Production*



BlueGene/L Chip



Just add DRAM

Processor

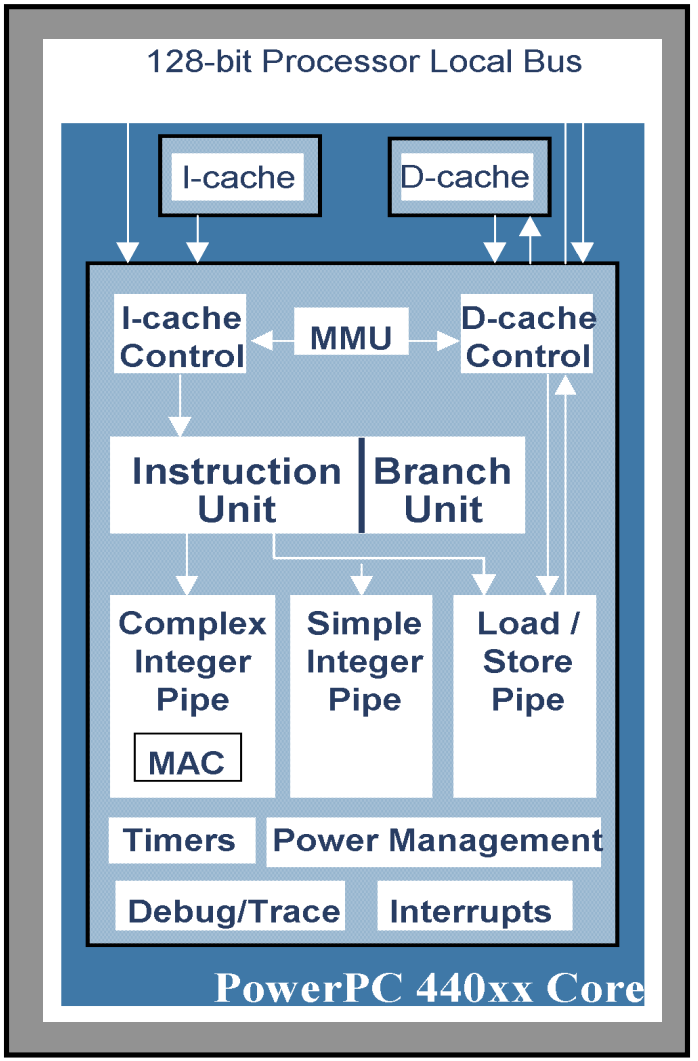
- PPC440x5 Processor Core – 700 MHz
  - *Superscalar: 2 instructions per cycle*
  - *Out of order issue and execution*
  - *Dynamic branch prediction, etc.*
- Two 64-bit floating point units
  - *SIMD instruct. over both register files*
  - *Parallel (quadword) loads/stores*
  - *2.8 GFLOPS/processor*

Interconnect

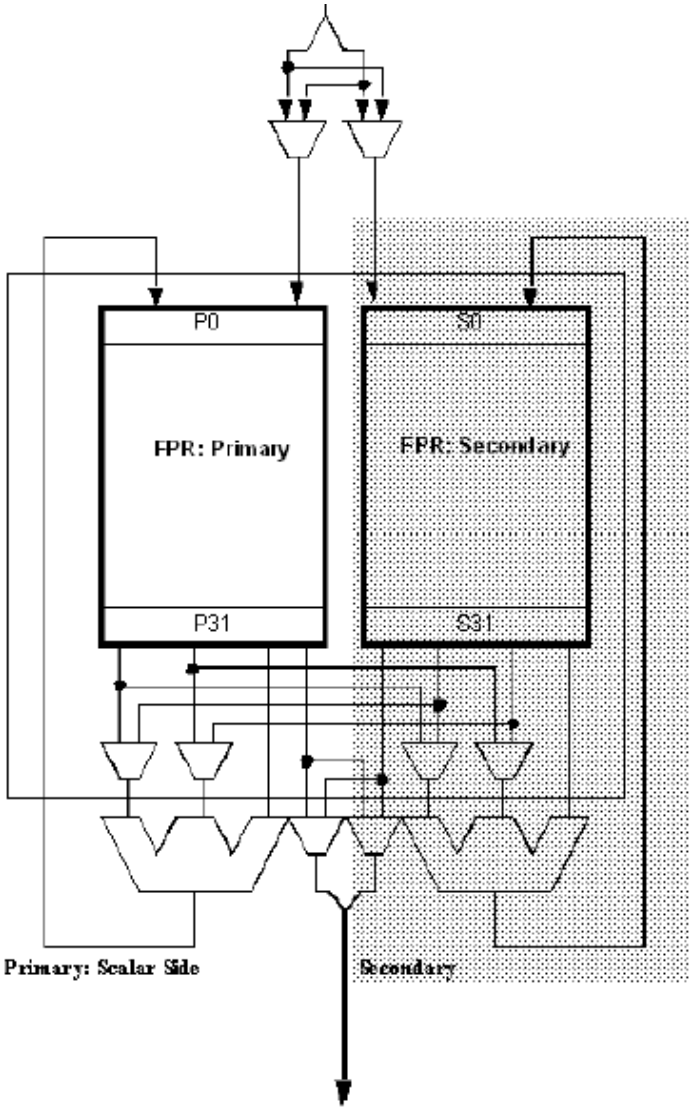
- 3 Dimensional Torus
  - *Virtual cut-through hardware routing*
  - *1.4Gb/s on all 12 node links*
  - *1 µs latency bet. neighbors, 5 µs to farthest*
- Global Tree
  - *One-to-all broadcast, reduction functionality*
  - *2.8 Gb/s of bandwidth per link*
  - *Latency of one way tree traversal 2.5 µs*
- Low Latency Global Barrier and Interrupt
  - *Latency of round trip 1.3 µs*
- Ethernet
  - *All external comm. (file I/O, control, etc.)*
- Control Network

PPC440x5 Processor Core Features

- High performance embedded PowerPC core
- 2.0 DMIPS/MHz
- Book E Architecture
- Superscalar: Two instructions per cycle
- Out of order issue, execution, and completion
- 7 stage pipeline
- 3 Execution pipelines
  - Combined complex, integer, & branch pipeline
  - Simple integer pipeline
  - Load/store pipeline.
- Dynamic branch prediction
- Single cycle multiply
- Single cycle multiply-accumulate
- Real-time non-invasive trace
- 128-bit CoreConnect Interface

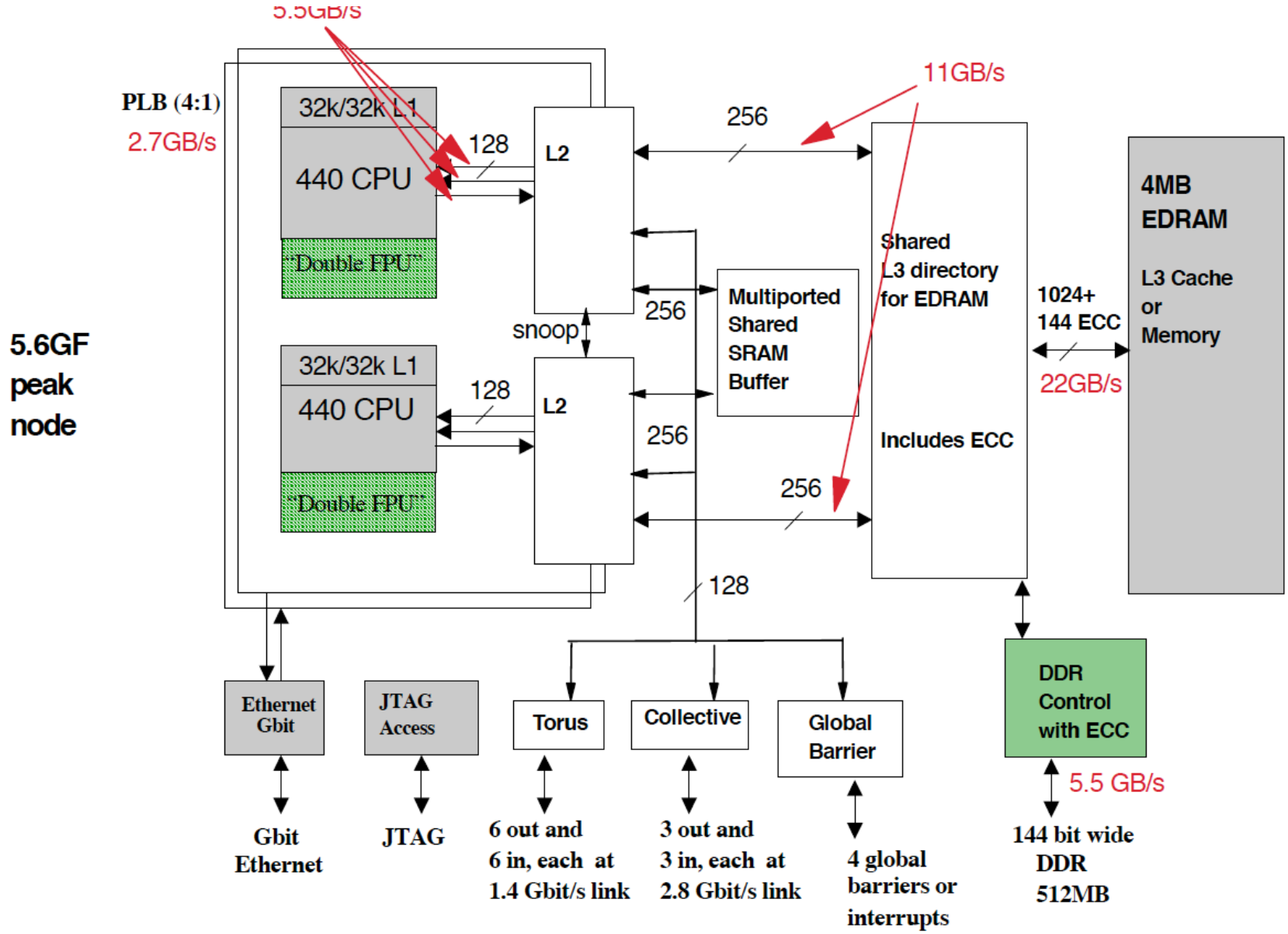


Dual FPU Architecture

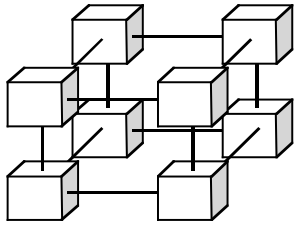


- Two 64 bit floating point units
- Designed with input from compiler and library developers
- SIMD instructions over both register files
  - FMA operations over double precision data
  - More general operations available with cross and replicated operands
    - *Useful for complex arithmetic, matrix multiply, FFT*
- Parallel (quadword) loads/stores
  - Fastest way to transfer data between processors and memory
  - Data needs to be 16-byte aligned
  - Load/store with swap order available
    - *Useful for matrix transpose*

BlueGene/L Compute System on a Chip ASIC

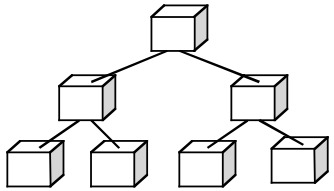


BlueGene/L - Five Independent Networks



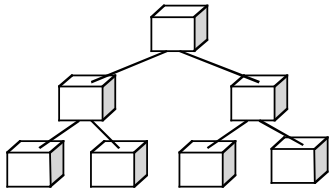
3 Dimensional Torus

- Point-to-point



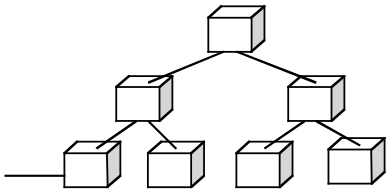
Global Tree

- Global Operations



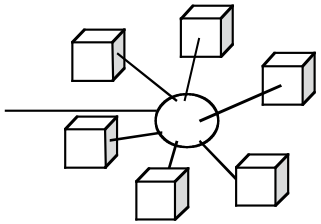
Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



Gbit Ethernet

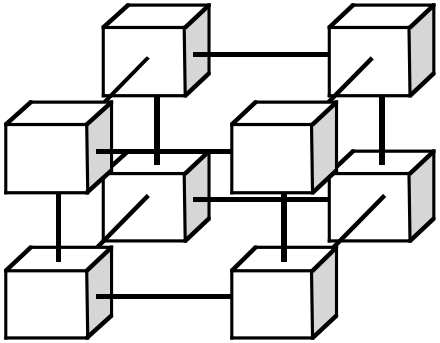
- File I/O and Host Interface

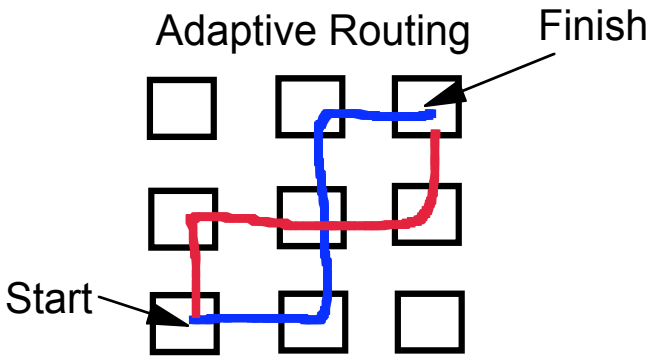


Control Network

- Boot, Monitoring and Diagnostics

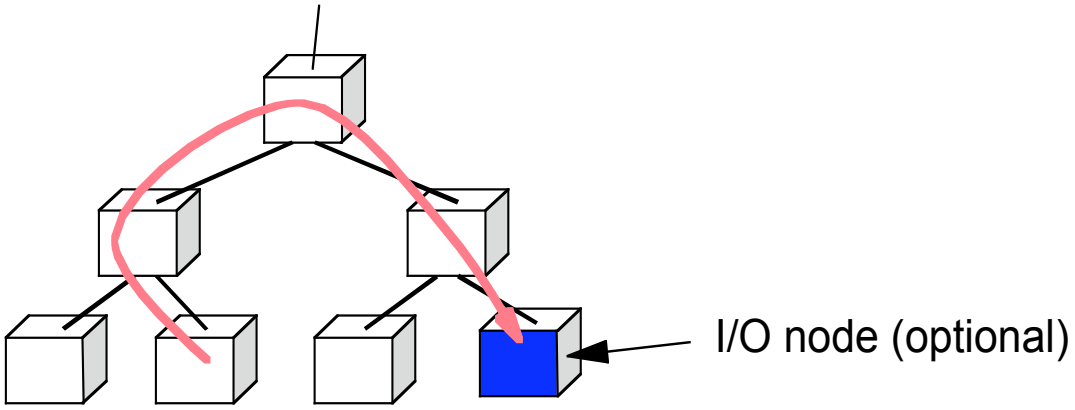
3-D Torus Network





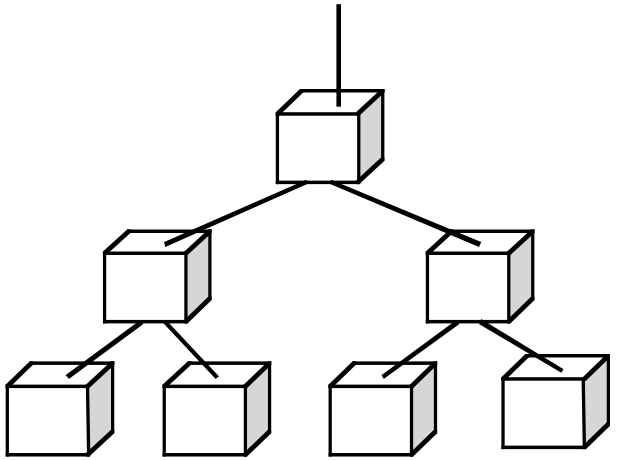
- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/node)
- 64k * 6 * 1.4Gb/s = 68 TB/s total torus bandwidth
- 4 * 32 *32 * 1.4Gb/s = 5.6 Tb/s Bisectional Bandwidth
- Worst case hardware latency through node ~ 69nsec
- Virtual cut-through routing with multipacket buffering on collision
  - Minimal
  - Adaptive
  - Deadlock Free
- Class Routing Capability (Deadlock-free Hardware Multicast)
  - Packets can be deposited along route to specified destination.
  - Allows for efficient one to many in some instances
- Active messages allows for fast transposes as required in FFTs.
- Independent on-chip network interfaces enable concurrent access.

Tree Network



- High Bandwidth one-to-all
  - 2.8Gb/s to all 64k nodes
  - 68TB/s aggregate bandwidth
- Arithmetic operations implemented in tree
  - Integer/ Floating Point Maximum/Minimum
  - Integer addition/subtract, bitwise logical operations
- Latency of tree less than 2.5usec to top, additional 2.5usec to broadcast to all
- Global sum over 64k in less than 2.5 usec (to top of tree)
- Used for disk/host funnel in/out of I/O nodes.
- Minimal impact on cabling
- Partitioned with Torus boundaries
- Flexible local routing table
- Used as Point-to-point for File I/O and Host communications

Fast Barrier Network

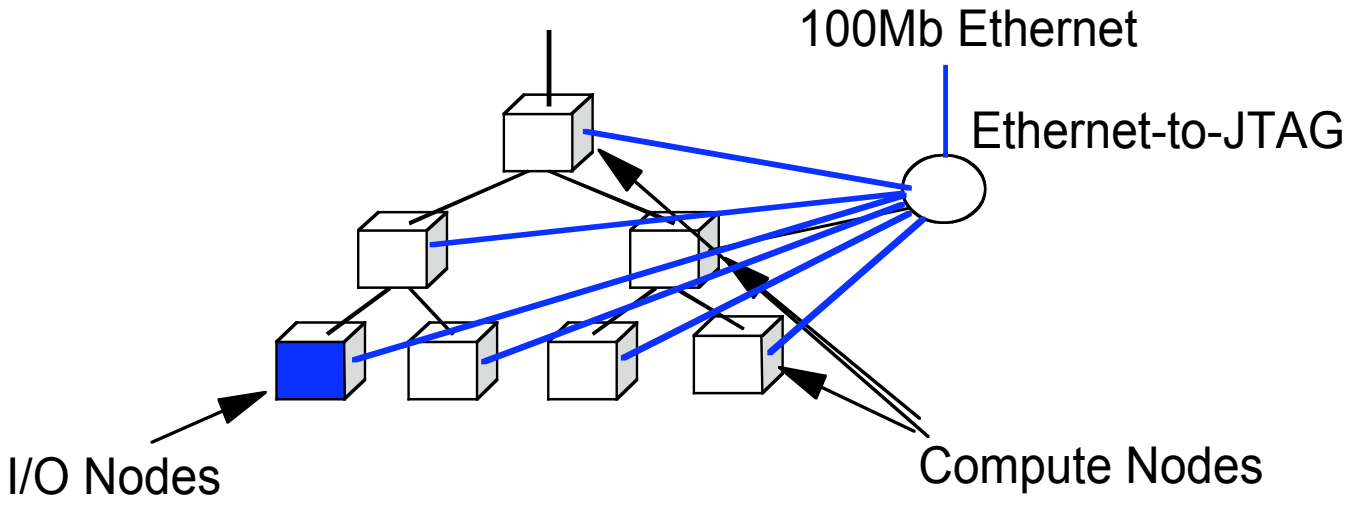


- Four Independent Barrier or Interrupt Channels
  - Independently Configurable as "or" or "and"
- Asynchronous Propagation
  - Halt operation quickly (current estimate is 1.3usec worst case round trip)
    > 3/4 of this delay is time-of-flight.
- Sticky bit operation
  - Allows global barriers with a single channel.
- User Space Accessible
  - System selectable
- Partitions along same boundaries as Tree, and Torus
  - Each user partition contains it's own set of barrier/ interrupt signals

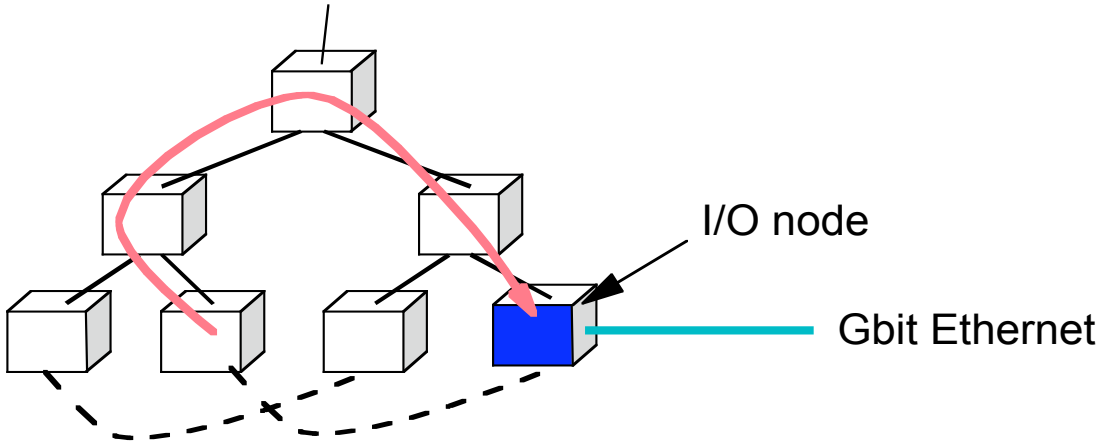
Control Network

JTAG interface to 100Mb Ethernet

- direct access to all nodes.
- boot, system debug availability.
- runtime noninvasive RAS support.
- non-invasive access to performance counters
- Direct access to shared SRAM in every node



Ethernet Disk/Host I/O Network



Gb Ethernet on all I/O nodes

- Gbit Ethernet Integrated in all node ASICs but only used on I/O nodes.
- Funnel via global tree.
- I/O nodes use same ASIC but are dedicated to I/O Tasks.
- I/O nodes can utilize larger memory.

Dedicated DMA controller for transfer to/from Memory

Configurable ratio of Compute to I/O nodes

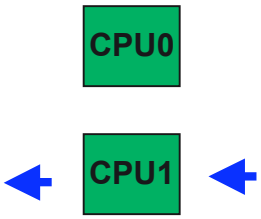
- I/O nodes are leaves on the tree network

BG/P Programming Environment

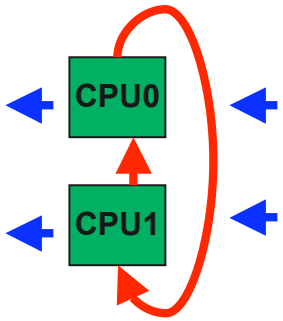
- Fortran, C, C++ with MPI
- Linux: User accesses system through Front End nodes for compilation, job submission, debugging
- Compute Node OS: very small, selected services, I/O forwarding
- OpenMP, Pthreads (up to 4)
- Space sharing - one parallel job (user) per partition of machine, one process per processor of compute node
- Single executable image is replicated on each node
- Virtual memory limited to physical memory
- Maybe dynamically linking

Applications Developer's View of BG/L

- Two CPU cores per node at 700 MHz
  - Each CPU can do 2 Float multiply-adds per cycle
- Mode 1 (Co-processor mode - CO)
  - CPU0 does all the computations (512MB memory)
  - CPU1 does the communications
  - Communications overlap with computation
  - Peak compute performance is 5.6/2 = 2.8 GFlops

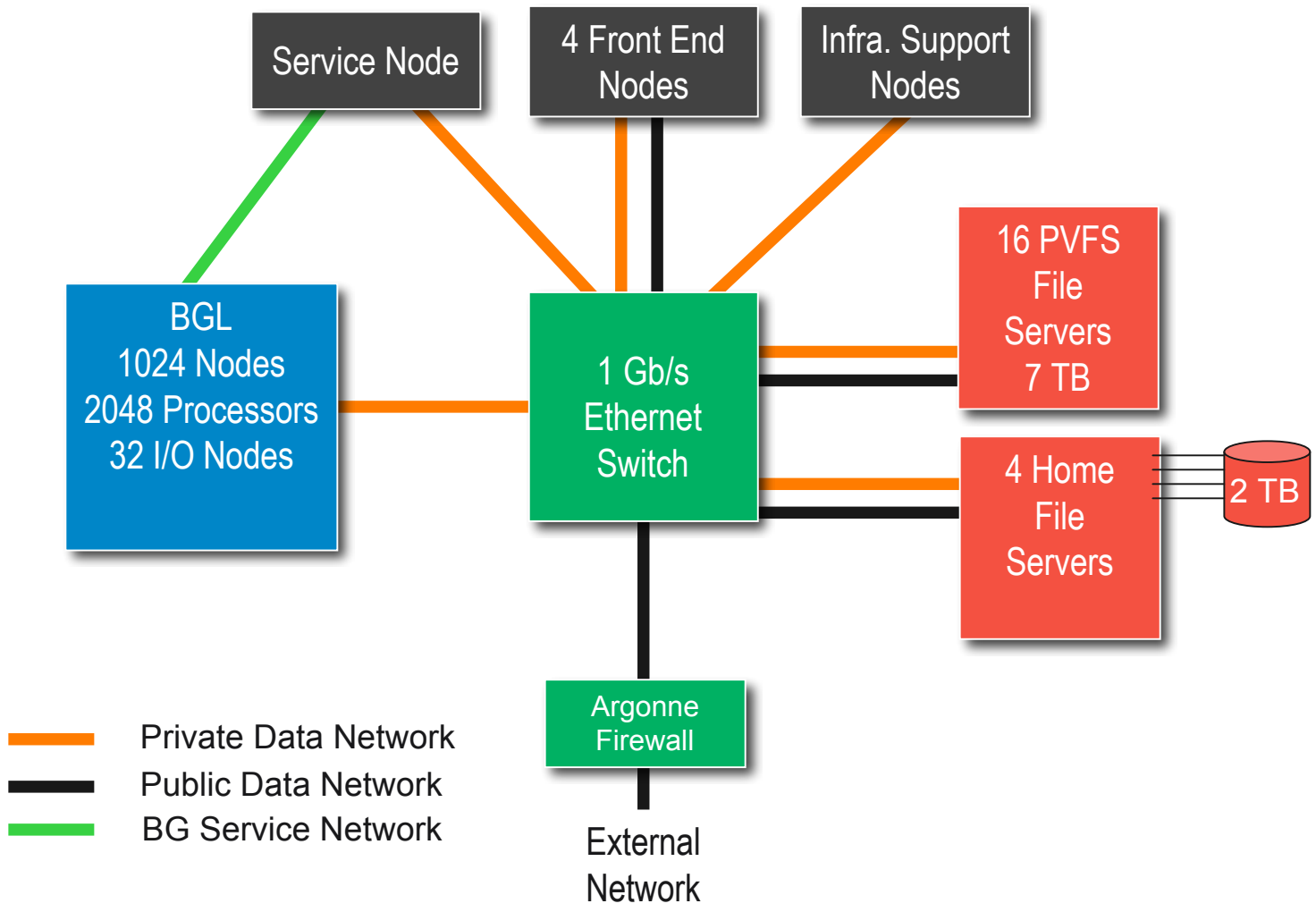


- Mode 2 (Virtual node mode - VN)
  - CPU0, CPU1 independent "virtual tasks" (256MB each)
  - Each does own computation and communication
  - The two CPU's talk via memory buffers
  - Computation and communication cannot overlap
  - Peak compute performance is 5.6 GFlops



- 3-D torus network with virtual cut-through routing
  - (point to point: MPI_ISEND, MPI_IRECV)
- Global combine/broadcast tree network
  - (collectives: MPI_GATHER, MPI_SCATTER)

Argonne BGL System Architecture



U.S. Department of Energy

Office of Science

Since 2004

INCITE

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
- Provides large computer time & data storage allocations
  - to a small number of projects for 1-3 years

INCITE Project Proposals

- **Scientific Discipline Peer Review**
  - Scientific quality
  - Proposed impact of the science
  - Ability of the PI and team
  - Computational plan
  - Relation to the Office of Science mission-related research
- **Computational Readiness Review**
  - Reasonableness and appropriateness of resource request
  - Appropriateness of approach
  - Technical readiness - has code run at scale on target system?
  - Progress in previous year (for renewals)
- **Nonproprietary Research**
  - Must sign user agreement (non-negotiable)
- **Proprietary Research is permitted**
  - Full cost recovery; user agreement required; data protection considerations

INCITE 2008

- **Call for proposals issued May 16**
  - **Proposals due August 8**
  - **See http://hpc.science.doe.gov**

- **Spans 250M hours of computing at**

Argonne	IBM Blue Gene	www.alcf.anl.gov
ORNL	Cray X1e and XT4	www.nccs.gov
NERSC/LBNL	Opteron Cluster, SGI Altix, IBM Power 3+5	www.nersc.gov
PNNL	HP-MPP	mscf.emsl.pnl.gov

- For guidance on submitting a proposal, contact
  - Paul Davé, Manager, ALCF User Services and Outreach
    Dave@alcf.anl.gov