

BlueGene/L:

the next generation of scalable supercomputer



Lynn Kissel
LLNL BlueGene/L Program Manager

David A. Nowak
Deputy Associate Director for DNT

Mark Seager
ASCI Platforms PI

Kim Yates
LLNL BlueGene/L SW Program Manager

November 18, 2002



Lawrence Livermore National Laboratory , P. O. Box 808, Livermore, CA 94551



Lawrence Livermore
National Laboratory



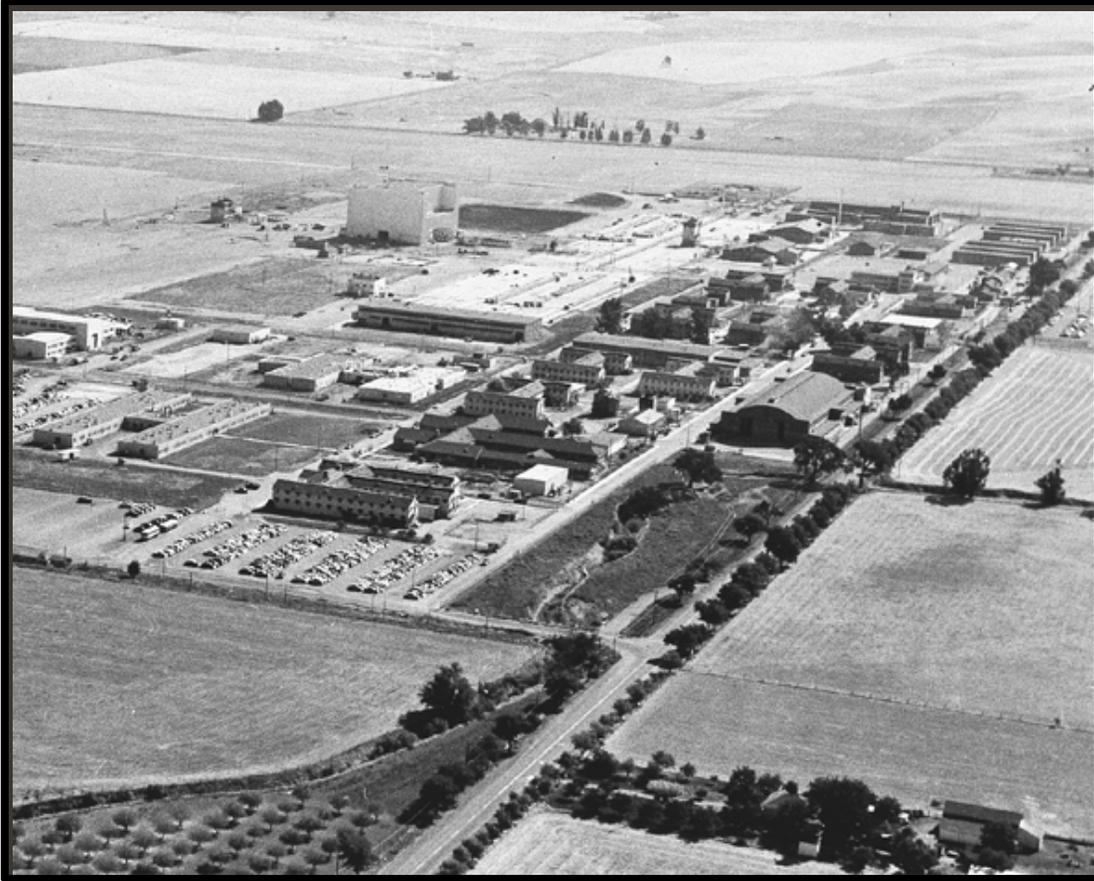
*Making History
Making a Difference*



1952-2002

*Ensure national security
and apply science and
technology to the
important problems of
our time*

Livermore branch of the University of California Radiation Laboratory



In 1952–1953

- **123 employees**
- **Annual budget
~ \$3.5 million**
- **1.2-square-mile
main site**

**Created to meet an urgent national security need by helping to
advance nuclear weapons science and technology**

Today Lawrence Livermore is a vibrant multidisciplined national laboratory

Dedicated to ensure national security and apply science and technology to the important problems of our time

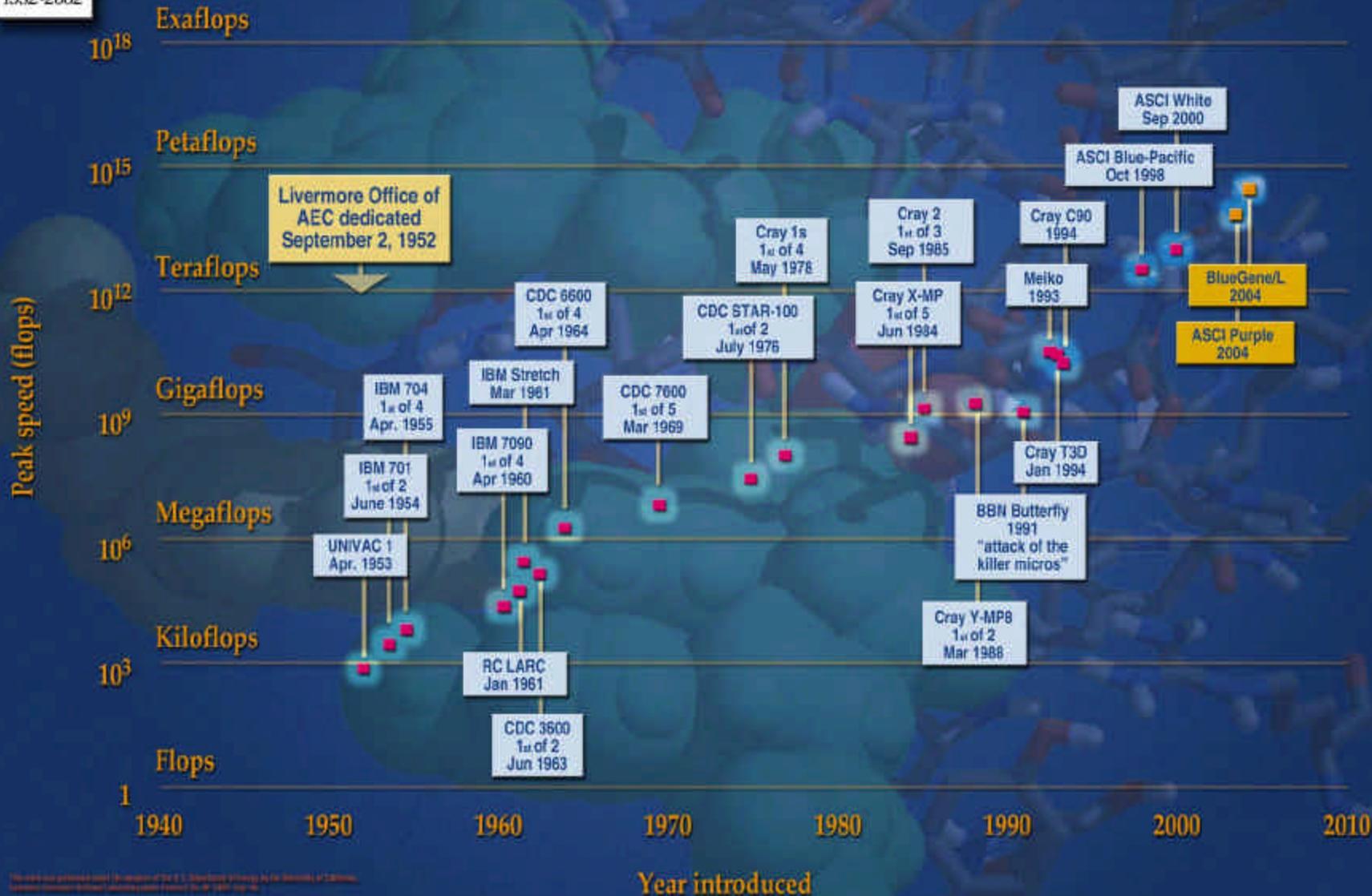


- **8300 employees**
- **Annual budget ~ \$1.5 billion**
- **1.2-square-mile main site**
- **Experimental test site near Tracy**



The Laboratory has been heavily vested in supercomputing since our founding

That tradition continues today



Important elements of our vision

Numerical Simulation is an integral part of the Stockpile Stewardship Program

Numerical Simulation is historically a cornerstone of Lawrence Livermore National Laboratory

ASCI is dedicated to demonstrating the transition

- ✧ **From** theory and experiment with interpolative simulation
- ✧ **To** a balanced trilogy of theory, experiment and simulation
- ✧ Thus, predictive scientific simulation is a critical competency for the Laboratory's future

The full configuration of Blue Pacific featured 5,856 processors and 3.9-teraflop/s capability

- The partnership to build the world's most powerful computer was announced at the White House on July 26, 1996
- The contract was signed on August 12, 1996
- The initial delivery was made in Livermore more than 30 days early on September 20, 1996
- Its full configuration was up and running by September 1998



LLNL's ASCI White is capable of 12.3 trillion floating-point operations per second



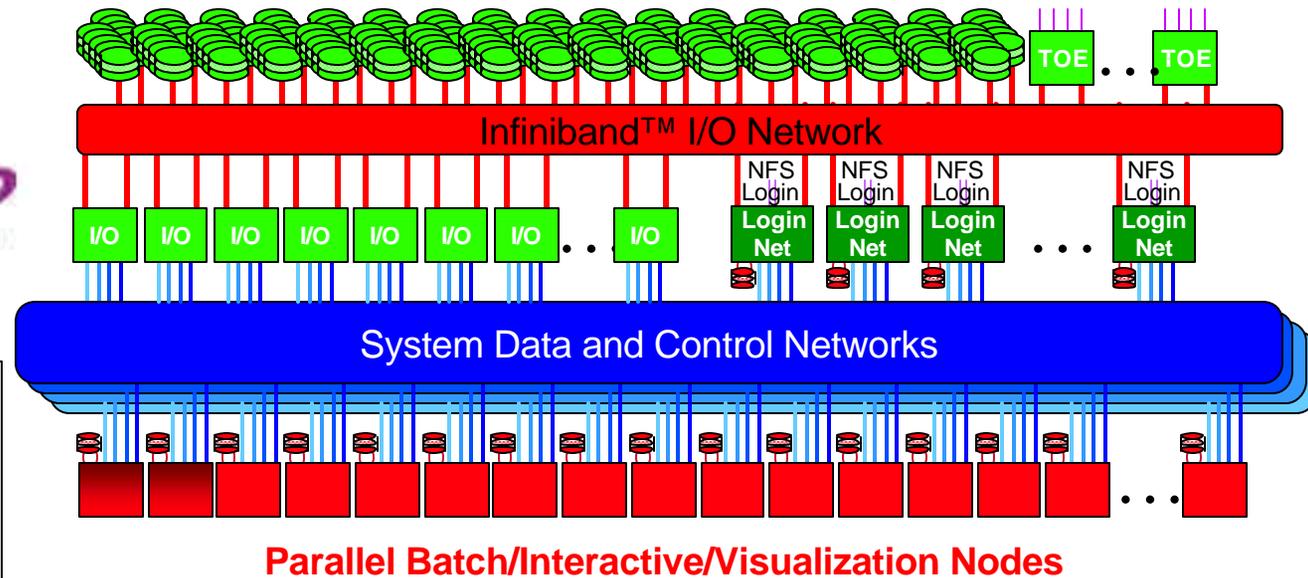
- **ASCI White weighs 106 tons and covers 12,000 square feet of floor space (an area greater than that of two NBA basketball courts)**
- **It contains 8,192 microprocessors in 512 shared memory nodes**
- **Each node contains 16 Power3-II CPUs built with IBM's latest semi-conductor technology (silicon-on-insulator and copper interconnects)**
- **Its 8 TB of memory is 125,000 times that of a 64-MB PC**
- **160 TB of storage in 7000 disk drives provides about 16,000 times the storage capacity of a PC with a 10-GB hard drive**

ASCI Purple is targeted for December 2004



See presentation at
the ASCI Booth,
Tues., 11/19, 4:00 pm

“LLNL Platforms”
Mark Seager



* Information from Purple RFP, April 22, 2002

Purple System

- Parallel batch/interactive/visualization nodes
- 2-8 Login/network nodes
- Clustered I/O services for global I/O
- External networking through TOE's
 - Login/network nodes for login/NFS
 - Infiniband™ for parallel FTP
 - All external networking is 1-10Gb/s Ethernet

Programming/Usage Model

- Application launch over all compute nodes
- 1 MPI task/CPU and Shared Memory, full 64b support
- Scalable MPI (MPI_allreduce, buffer space)
- Likely usage
 - multiple MPI tasks/node with 4-16 OpenMP/MPI task
- Single STDIO interface
- Parallel I/O to single file, multiple serial I/O (1 file/MPI task)

ASCI Purple Home Page
<http://www.llnl.gov/asci/purple/>

The requirements are diverse and are overrunning the plan

Stockpile Work

- **Stockpile work requires integrated codes for highly complex multi-physics problems**
 - ✧ **The “Burn-code Milestone” (16 days on 1,024 processors)**
 - ✧ **Memory intensive**
 - ✧ **Communication intensive**
 - ✧ **Problem Solving Environment stressed**

Science Runs

- **Science runs in support of stockpile work have different requirements**
 - ✧ **Floating-point intensive**
 - ✧ **Easily parsed**
 - ✧ **Memory light**
 - ✧ **Communication light**
 - ✧ **Problem Solving Environment light**

The simulation strategy for BlueGene/L

- **Tune system balance to work load**
 - ✧ **Heavy system for Integrated Codes**
 - ✧ **Light system for basic physics**
 - **Turbulence**
 - **Ejecta**
 - **Advanced Materials Simulation**

- **Offload simpler science applications to BlueGene/L to minimize pervasive capacity gridlock**

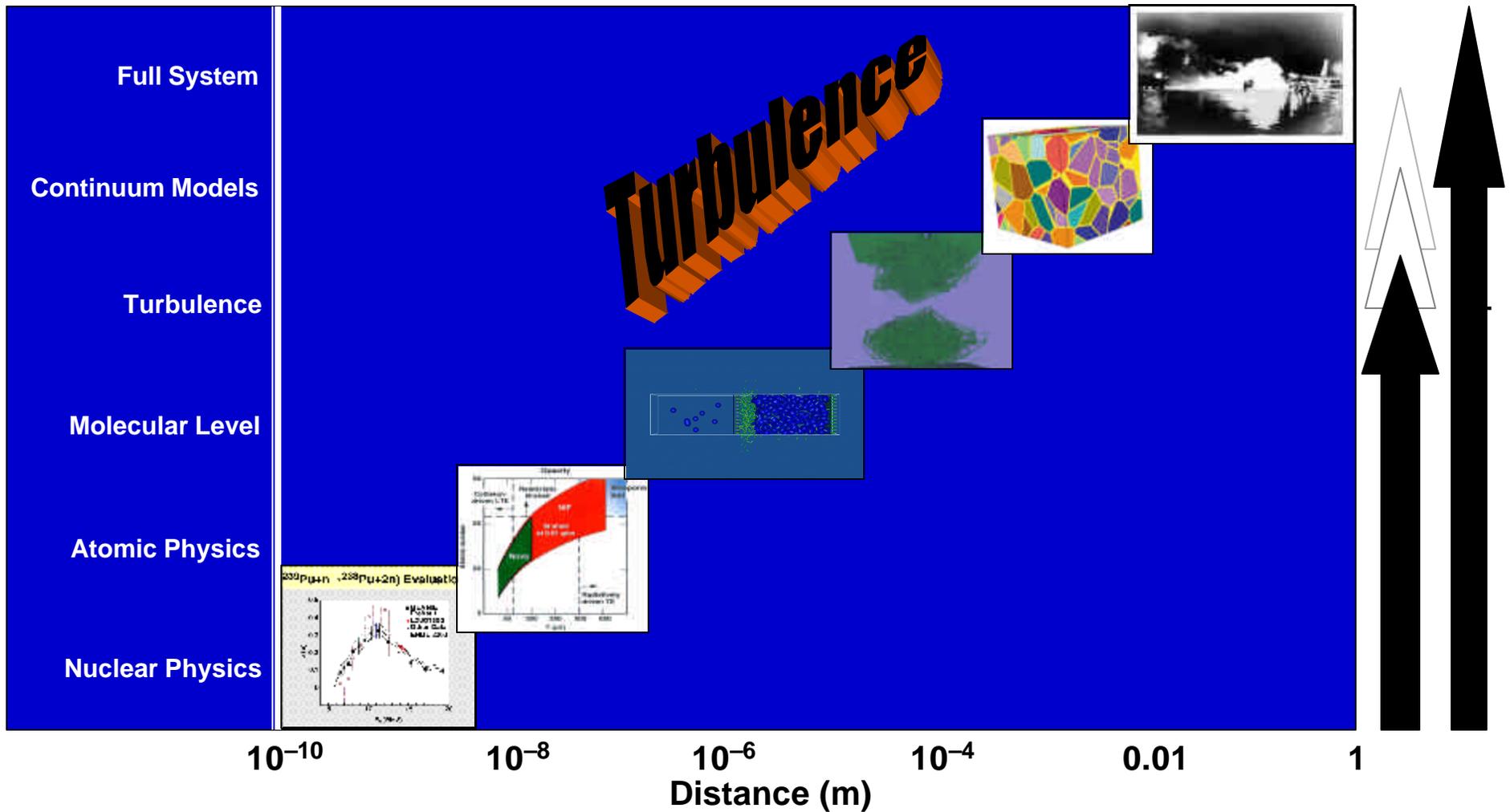
BlueGene/L is an architecture optimized for cost, performance and scalability

- **Partnership between IBM and DOE/NNSA, LLNL is lead lab**
 - ✧ **Address a limited but large set of applications**
- **Low power, low cost, high performance**
 - ✧ **New generation of embedded processors**
 - ✧ **Large on chip DRAM**
 - ✧ **High speed, low latency, low power serial links**
 - ✧ **Simplified OS**
- **Proven methodology**
 - ✧ **Columbia QCDSP Gordon Bell Prize**
- **Attacks the distance to memory problem**
 - ✧ **Low latency memory**
 - ✧ **High bandwidth memory**
- **Significant savings in facilities cost**
- **Scalable to petaFLOP/s systems**

Detailed algorithm analysis shows that ASCI science applications map well to BlueGene/L architecture

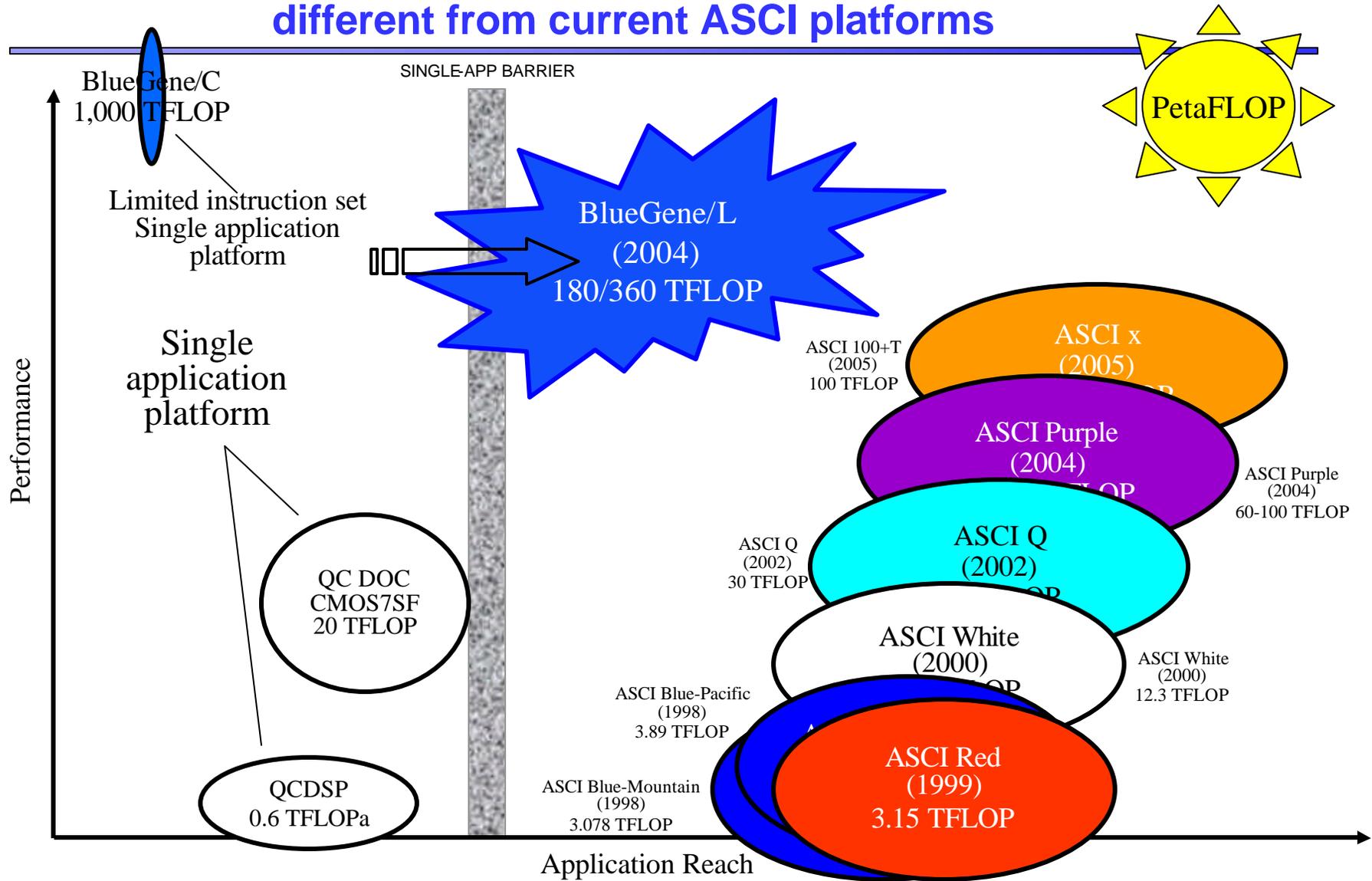
Science Applications	Scientific Importance	Mapping to BlueGene/L Architecture
Ab initio molecular dynamics (multiple impacts in materials and computational biology)	A	A
Three-Dimensional Dislocation Dynamics	A	A
Multi-scale materials modeling	A	A
Atomistic-dynamic simulations extended to macroscopic timescales	A	B
Computational gene discovery	A	B
Turbulence: Rayleigh-Taylor instability	A	A
Shock turbulence	A	A
Turbulence and instability modeling (sPPM)	B	A
Hydrodynamic instability	A	A

BlueGene/L addresses some important physics issues that have been outside the reach of direct numerical simulation



BlueGene/L is clearly appropriate to address levels 1-4 – application to simulation levels 5-6 is hopeful

BlueGene/L occupies a niche of the ultra-computer landscape different from current ASCI platforms



BlueGene/L characteristics differ from recent ASCI acquisitions

	ASCI White	ASCI Q	Earth Simulator	BlueGene/L [†]
Machine Peak Speed (Tflop/s)	12.3	30	40	180 / 360*
Total Memory (Tbytes)	8	33	10	16–32
Footprint (ft.² / m²)	10,000 / 930	20,000 / 1,900	34,000 / 3,200	2,500 / 230
Total Power (MW)	1.0	3.8	10.0	1.2
Cost (M\$)	~100	~200	~350	Much Less
Installation Date	9/2000	~9/2002	2/2002	~12/2004
No. of Nodes	512	4,096	640	65,536
CPUs per Node	16	4	8	2
Clock Frequency (MHz)	375	1,000	500	700
Power Dissipation/Node (W)	1,953	922	6,400	15
Peak Speed/Node (Gflop/s)	24.0	7.3	64.0	2.8
Memory/Node (GiB)	16	8	16	0.25–0.5
Memory Bandwidth (TB/s)	8	19	164	360
Memory Latency (cycles)	140	330	–	70
MPI Latency (ms)	25	4.5	6–20	7
Interconnect Bandwidth (B:F)	0.042	0.085	0.13	0.75
Bi-Section Bandwidth (B:F)	0.04	0.04	0.03	0.008

[†] target specifications

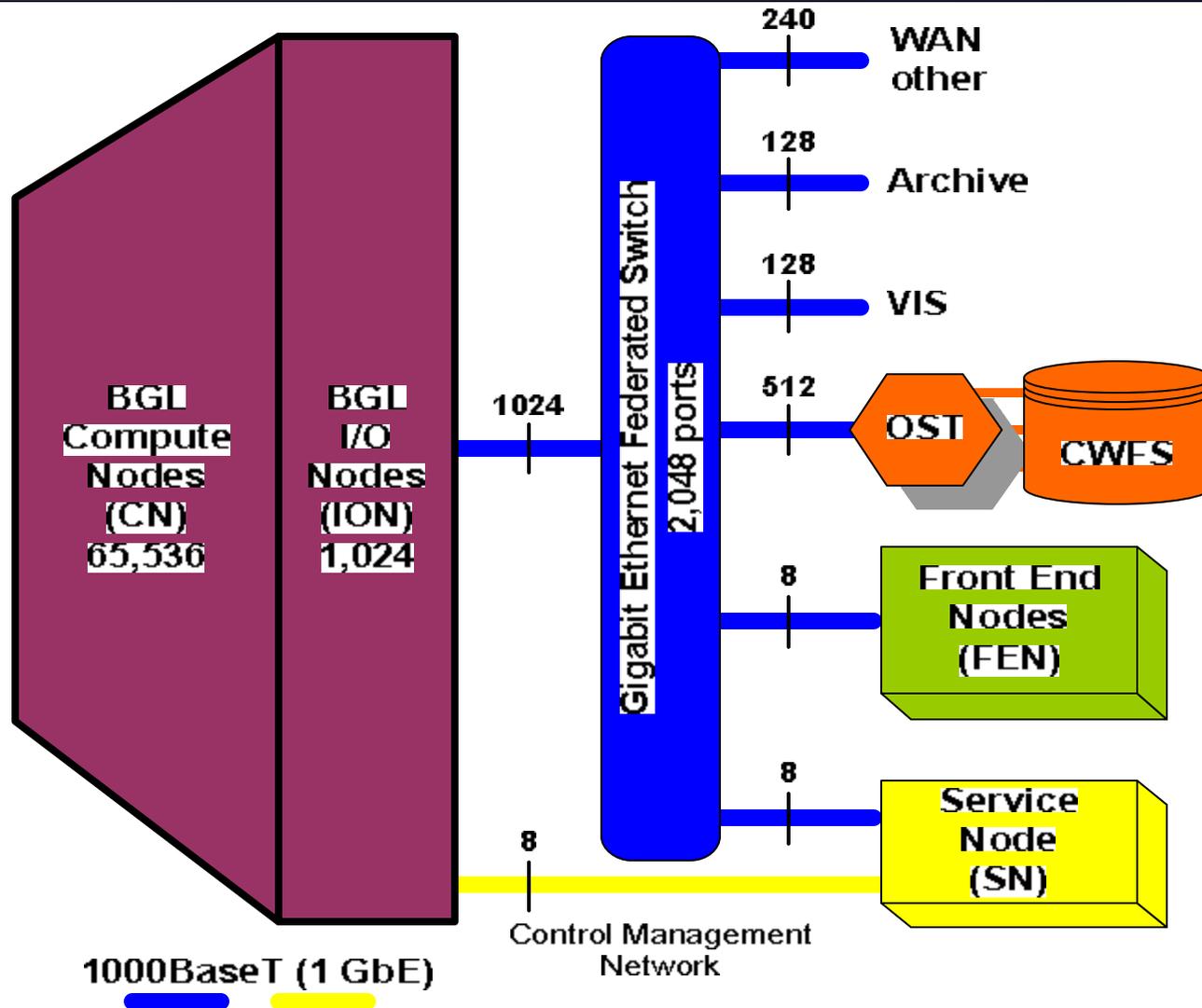
* comm. co-processor mode / symmetric mode

BlueGene/L's characteristics suggest new metrics to emphasize its dramatic departure from recent supercomputers

	ASCI White	ASCI Q	Earth Simulator	BlueGene/L [†]
Memory-Space Effectiveness (GiB/m²)	8.6	17	3.1	140
Speed-Space Effectiveness (GF/s/m²)	13	16	13	1600
Speed-Power Effectiveness (GF/s/kW)	12	7.9	4.0	300
Speed-Cost Effectiveness (GF/s/M\$)	~100	~100	~100	> 1,000 <10,000

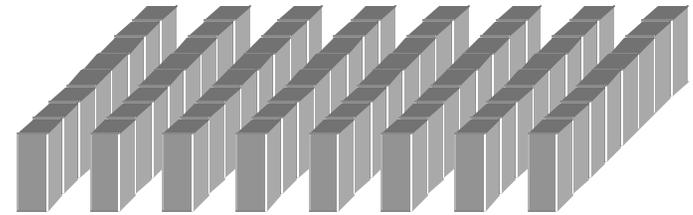
† target specifications

BlueGene/L is stimulating the development of a scalable storage area network for LLNL's Open Computing Facility



The high-level of integration results is a compact footprint –
the low part count results in improved reliability

Building BlueGene/L

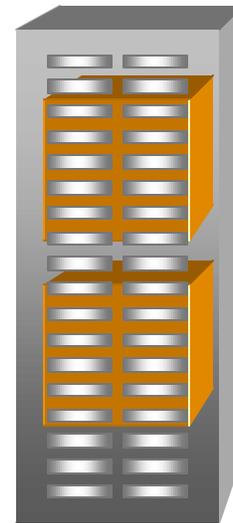


System

64 cabinets
(32x32x64)
180/360 TF/s
16 TiB*
~1 MW
2500 sq.ft.

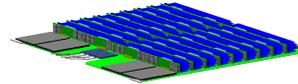
Cabinet

32 node boards
(8x8x16)
2.9/5.7 TF/s
256 GiB* DDR
15-20 kW



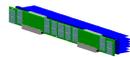
Midplane

SU (scalable unit)
16 node boards
(8x8x8)
1.4/2.9 TF/s
128 GiB* DDR
7-10 kW



Node Card

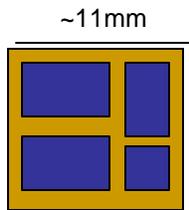
32 compute chips
16 compute cards
(4x4x2)
90/180 GF/s
8 GiB* DDR



Compute Card

FRU
25mmx32m
m 2
compute
chips
(2x1x1)
2.8/5.6

GF/s 256
MiB* DDR
15 W



~11mm

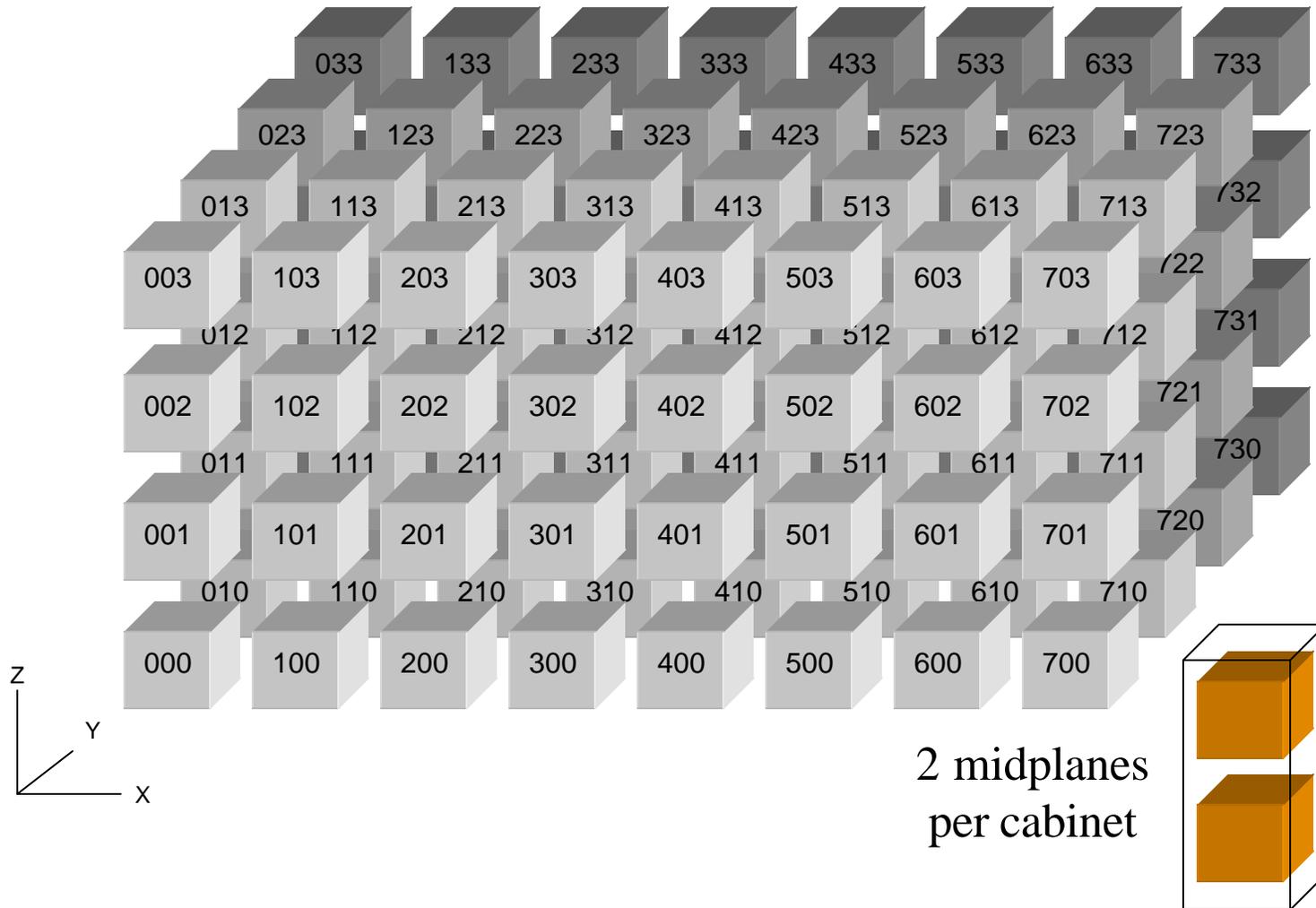
Compute Chip

2 processors
2.8/5.6 GF/s
4 MiB* eDRAM

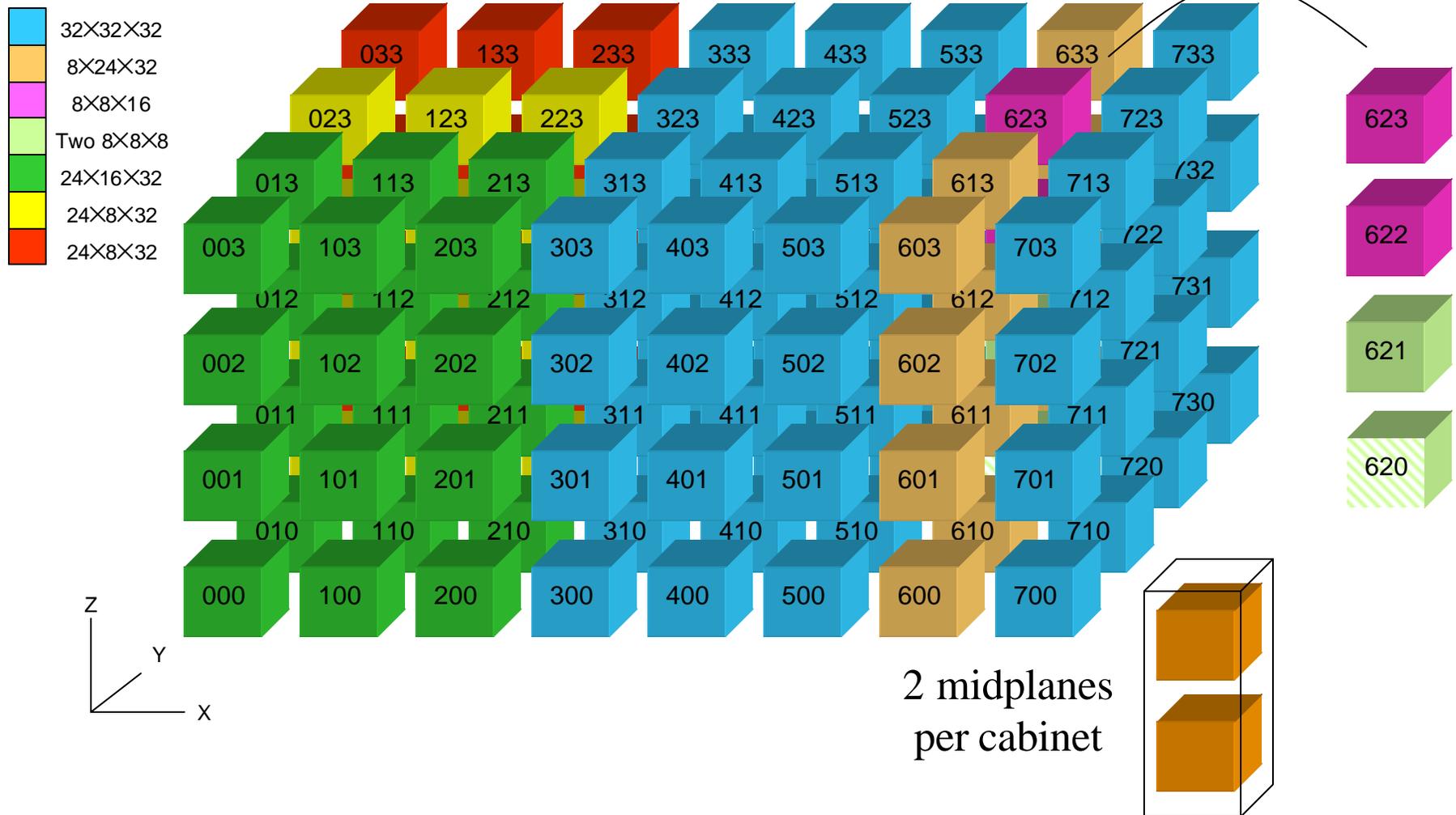
(compare this with a 1988
Cray YMP/8 at 2.7 GF/s)

* <http://physics.nist.gov/cuu/Units/binary.html>

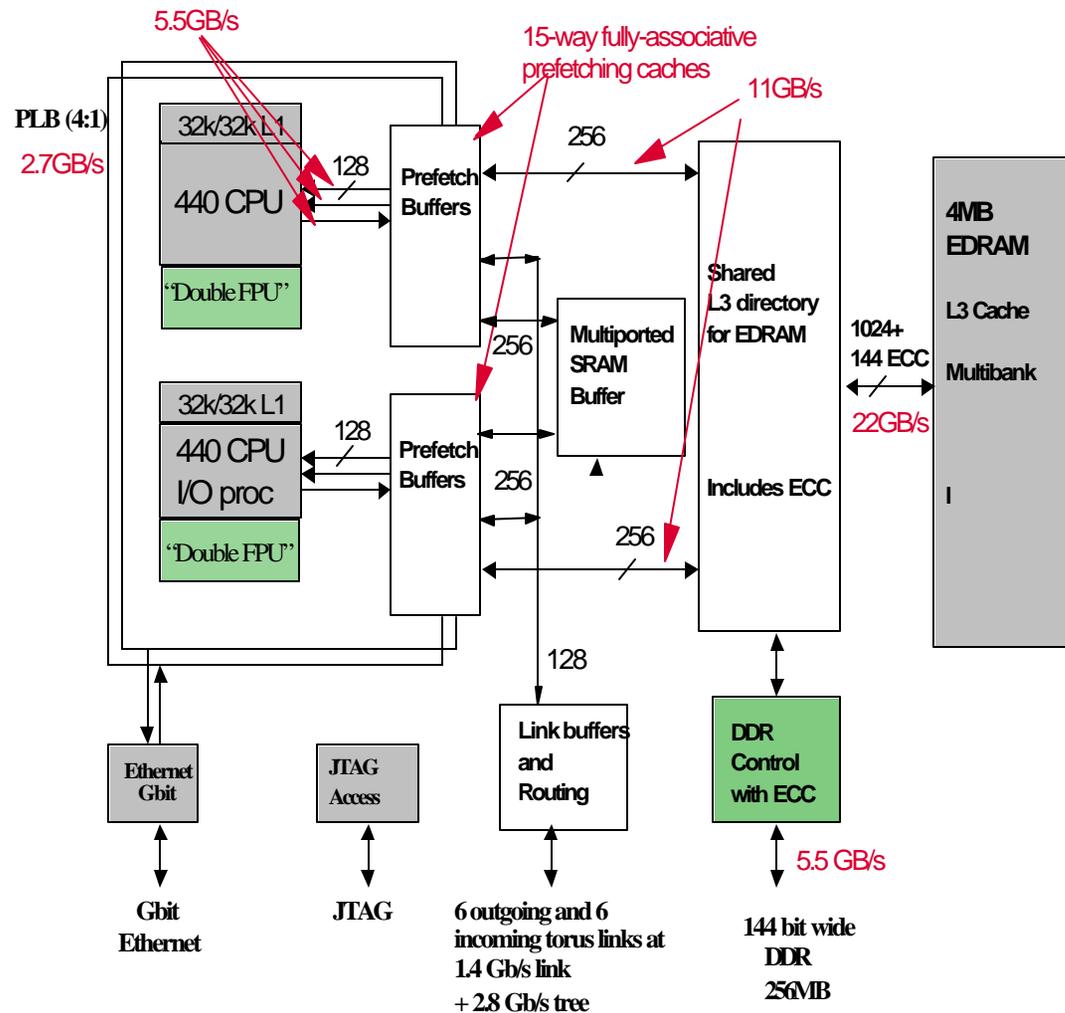
A midplane contains 512 nodes (8x8x8, 2.9TF/s) and is the scalable unit, either connected or isolated from neighbors



Each partition is a separate electrically-isolated machine with individual torus, tree and barrier networks



Each ~15W BlueGene/L compute node is composed of a single ASIC and 9 SDRAM-DDR memory chips – that's it!

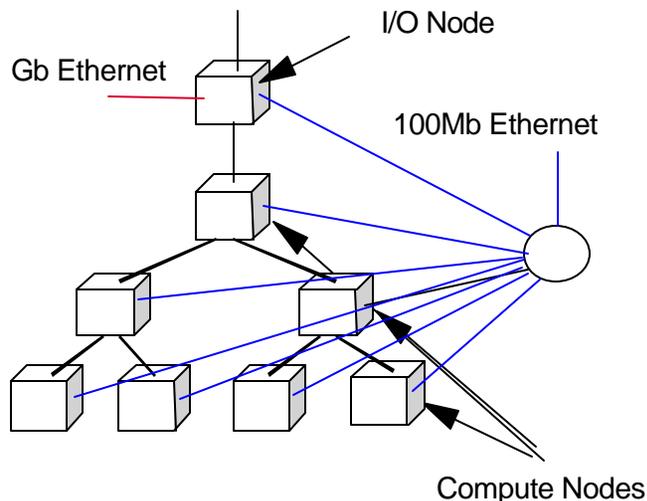
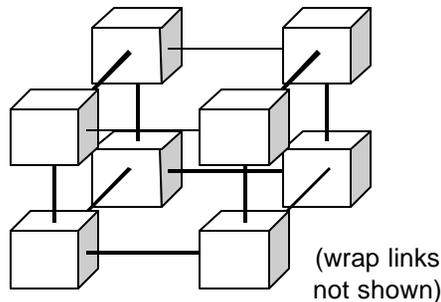


See technical paper at SC2002 in Hardware Architectures session, Thursday, 11/21, 3:30-5:00 pm

"An Overview of the BlueGene/L Supercomputer"

The **BlueGene/L** compute ASIC uses IBM CMOS CU-11 0.13µm technology. In this diagram, the *gray blocks* are standard System-On-A-Chip offerings from IBM's ASIC library. The *white blocks* require a new design effort, while the *green blocks* are developed from existing designs. All this on a ~11mmx11mm Si die.

Architectural features of BlueGene/L promote application efficiency and scaling – nodes connected by 5 networks



- **3D torus is the main node-node communications network**
 - Each node supports 6 independent bi-directional nearest-neighbor links with a target aggregate bandwidth of 4.2GB/s
- **Binary combining tree extends over the entire machine**
 - Allows data to be sent from any node to all others (broadcast), or a subset of nodes, with a target latency of about 2 ms
 - Tree can be utilized for global broadcast of data rather than shipping it around in rings – expected to be a tremendous asset for one-to-all communications
- **JTAG (IEEE 1149.1) for diagnostics and IPL**
 - Allows for access to the processor's registers, is connected to the 100 Mbit Ethernet port within the BlueGene/L ASIC, and is accessible using standard Ethernet I/P
- **I/O processing node is connected through the Global Combining Tree to 64 compute nodes**
 - Any compute node can read or write to disk at full gigabit speed
- **Low-latency global barrier network**

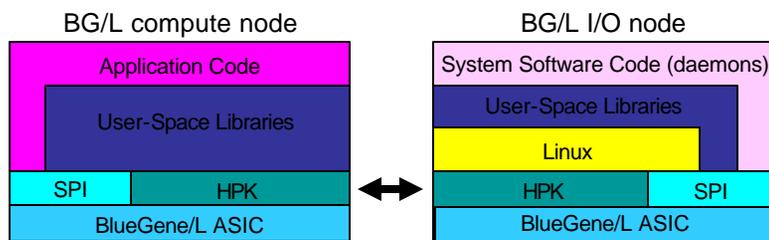
Some unique architectural features of BlueGene/L



SOC (system-on-a-chip) design



Double Hummer



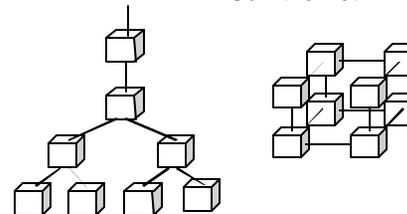
Light-weight kernel AND Linux services

JTAG (IEEE 1149.1)

100Mb Ethernet

global barriers

Gb Ethernet



Complementary networks

Our research strategy for BlueGene/L is designed to enhance its impact for computational science and simulation

□ Application performance

- ✧ **Characterize, model, simulate, and measure apps**
- ✧ **Working with other organizations to optimize FFTs, port performance analysis tools such as VGV**
(Vampir/GuideView performance-analysis tool set), **PAPI** (Performance API)
- ✧ **Porting and optimization of application-support libraries**
- ✧ **Scaling applications 65,536 nodes**
- ✧ **Identify bottlenecks and eliminate/ameliorate them**

□ Programming models

- ✧ **Default: 1 MPI task/node, offload commun. to 2nd CPU**
- ✧ **Virtual node mode, symmetric mode to exploit 2nd CPU**
- ✧ **Examining other models: UPC** (Unified Parallel C), **co-array Fortran/C**, **Stanford streams**, **CHARM++/AMPI** (Adaptive MPI), **fine-grain multithreading**

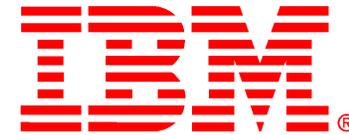
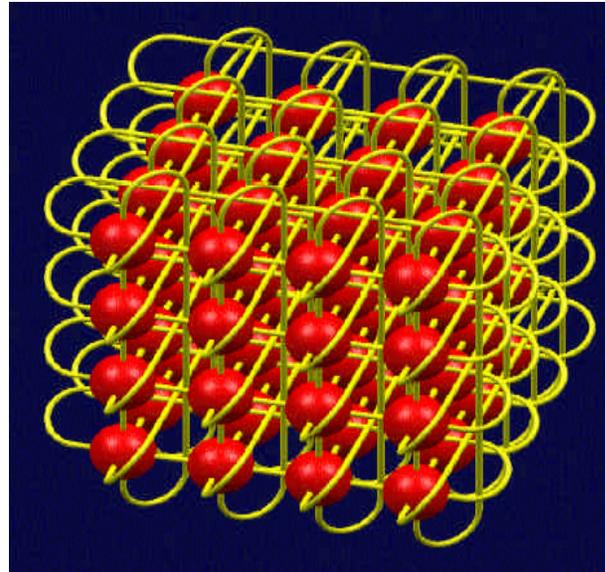
BlueGene/L collaboration involves the Tri-Lab with a growing list of industry and academia



Lawrence Livermore
National Laboratory
1952-2002



Applications
File Systems
Batch system
Kernel Evaluation
Programming Models
Debugger & Vis



Hardware design and build
Network design and build
OS and system software



Network simulator
MPI tracing
Application scaling



PAPI - performance
monitoring



MPI – message
passing interface



TECHNISCHE
UNIVERSITÄT
WIEN
VIENNA
UNIVERSITY OF
TECHNOLOGY

Optimized FFT



STAPL – standard
adaptive template
library



Debugger



Beckman Institute

for Advanced Science and Technology

Parallel Objects
CHARM++



Performance Modeling and Characterization

Applications
Application Tracing &
Performance

Our growing list partners extend beyond the Tri-Lab and IBM and bring world-class computational tools to BlueGene/L

- **Center for Advanced Computing Research (CACR), California Institute of Technology**
Network Simulation, MPI Tracing, Application Scaling
- **Performance Modeling and Characterization (PMAc) Laboratory, San Diego Supercomputer Center**
Applications, Application Tracing and Performance
- **Mathematics and Computer Science Division, Argonne National Laboratory**
Message Passing Interface (MPI)
- **Etnus, LLC**
TotalView Debugger
- **Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign**
Parallel Objects, CHARM++
- **Institute for Applied Mathematics and Numerical Analysis, Technical University of Vienna**
Optimized FFTs
- **The Innovative Computing Laboratory, University of Tennessee**
Performance Monitoring, PAPI (Performance API)
- **PARASOL Laboratory, Texas A&M University**
Standard Template Adaptive Parallel Library (STAPL), parallel extensions to C++

Summary

- **BlueGene/L** will have a scientific reach far beyond existing limits for a large class of important scientific problems
- **Cost / performance is significantly better than standard methods to get to a petaflop/s (10^{15} /s)**
 - ✧ **Embedded technology promises to be an efficient path toward building massively parallel computers optimized at the system level**
 - ✧ **Low Power is critical to achieving a dense, simple, inexpensive packaging solution**
- **A mature, sophisticated software environment needs to be developed to really determine the reach (both scientific and commercial) of this architecture**

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.