IBM Deep Computing

Oklahoma Supercomputing Symposium

Kent Winchell
Deep Computing CTO

October 2010

IBM Systems
Simplify your IT.

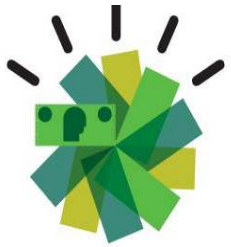
http://software-carpentry.org/

- Computers are as important to modern science as telescopes and test tubes. Unfortunately, most scientists are never taught how to use them effectively: most scientists have to figure out for themselves how to build, validate, maintain, and share complex programs. This is as fair as teaching someone arithmetic and then expecting them to figure out calculus on their own, and about as likely to succeed.

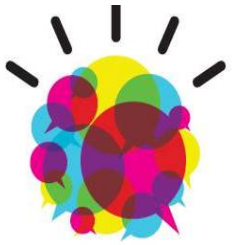
Smarter Planet Segments

smarter education

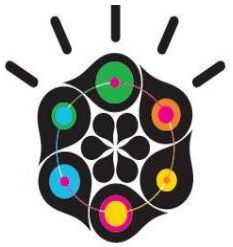
cloud computing



smarter financial systems



smarter telecommunications



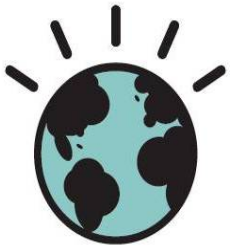
smarter oil management

public safety

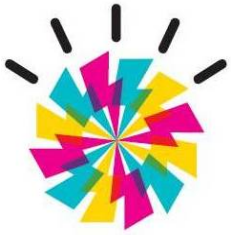
smarter buildings

smarter healthcare systems

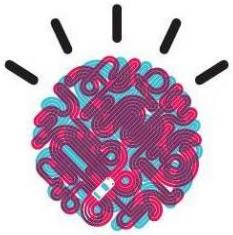
smarter cities



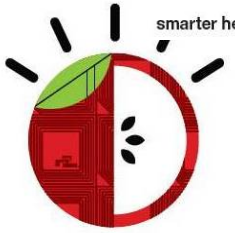
smarter planet



smarter energy grids

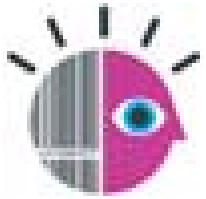


smarter traffic management

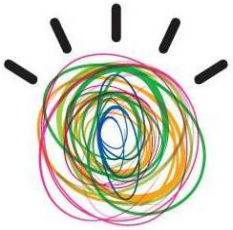


smarter food systems

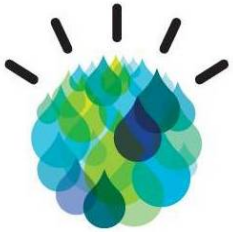
products



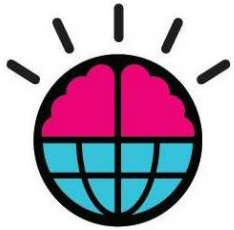
retail



smarter infrastructure



smarter water systems

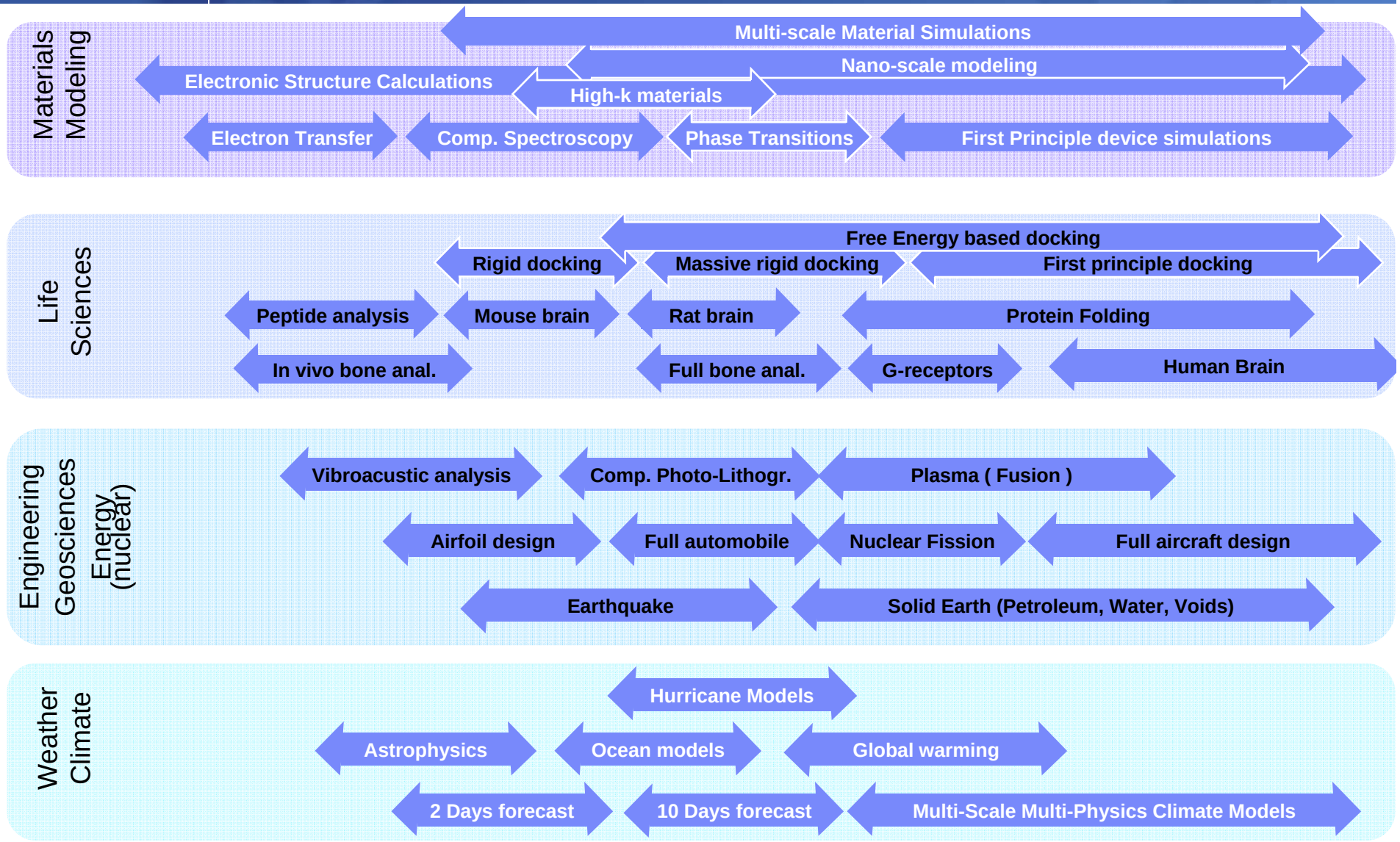


smarter information management

government services

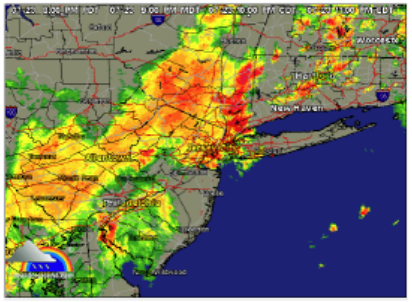
High Performance Computing for a Smarter Planet (a partial list)

Oil	Healthcare	Banking	Telecom	Infrastructure	Environment
Seismic Analysis	Drug Discovery	Actuarial Analysis	Video On Demand	Virtualization	Earthquake Modeling
Reservoir Analysis	Protein Folding	Asset Liability Management	Network Optimization	Data Management	Climate Modeling
Energy Conversion Systems	Medical Imaging	Portfolio Risk Analysis	Gaming	Cloud Computing	Remediation



Impact on Industry

Weather Forecasting

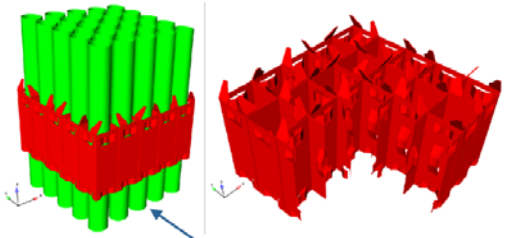


Weather Research and Forecasting Models
IBM Power Systems, smaller versions of ASC Purple, used extensively throughout the industry

Current frontier with *Code_Saturne*

Calculation under way with 100 million cells

- PWR assembly mixing grid
- calculation on 4 000 to 8 000 procs
- major lock due to mesh generation



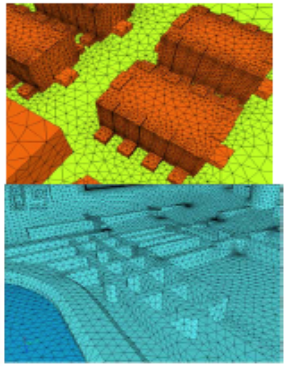
Code_Saturne: EDF's general purpose CFD software goes open source
Rencontres Mondiales du Logiciel Libre, Amiens, 11/07/2007

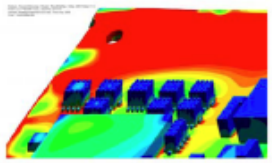
http://rd.edf.com/code_saturne
saturne-support@edf.fr

EDF

Large Scale Drop Impact Analysis of Mobile Phone

- ADVC, a commercial structural analysis code from Allied Engineering Corporation, Japan
  - An implicit based structural FEA system developed for parallel computation
- Drop Impact Analysis for a full assembly of a mobile phone is performed
  - Full model, including inner structure, virtually no simplifications
- Performance
  - 12 hours for 100 step simulation
  - 305 million degrees of freedom
  - 1.27 TF on 8192 Nodes of BGL
  - 2.8% of Peak
  - 2.4 ms of real time
  - Drop height increased from 10 cm to 1.5 m!
- 2006 Gordon Bell Finalist





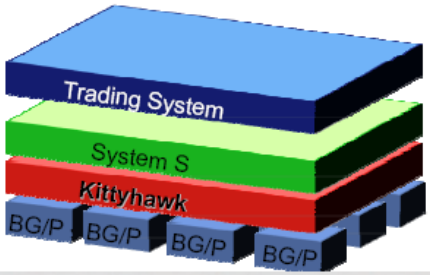
Real Time Options Trading

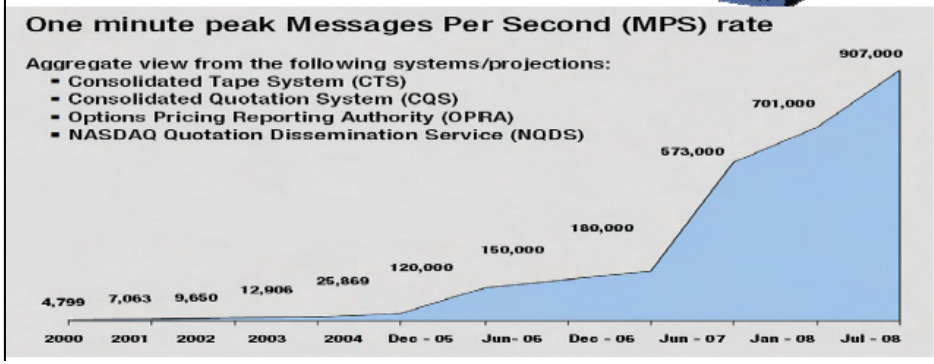
Investment Banking Application

§ Replace zoo of machines and Cisco with integrated Blue Gene/P

§ Replace non-scalable communication infrastructure with System S

§ Kittyhawk bridges between specialized Blue Gene/P hardware and legacy software





BLUE WATERS

SUSTAINED PETASCALE COMPUTING

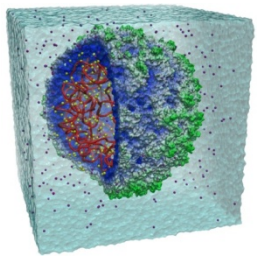
Blue Waters Update

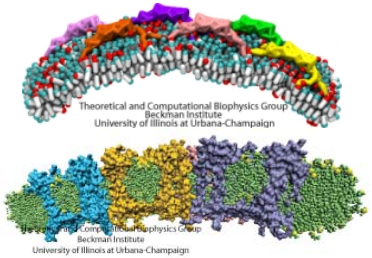
Intense Computing at the Petascale and Beyond

NCSA IBM GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

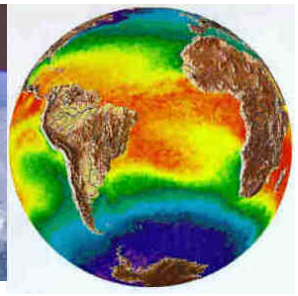
Sustained Petascale computing will enable advances in a broad range of science and engineering disciplines:

Molecular Science

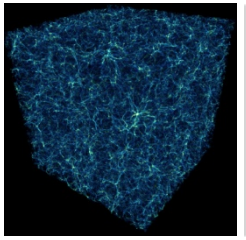


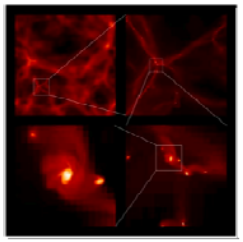


Weather & Climate Forecasting

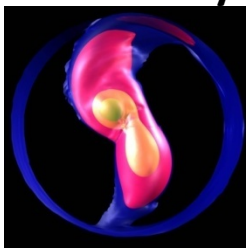


Astrophysics

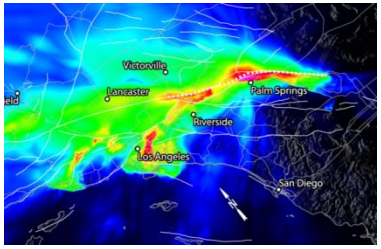




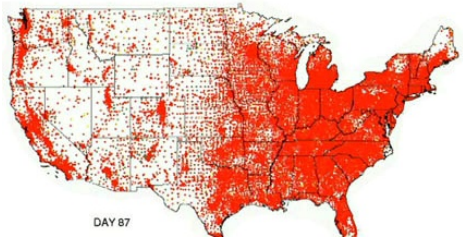
Astronomy



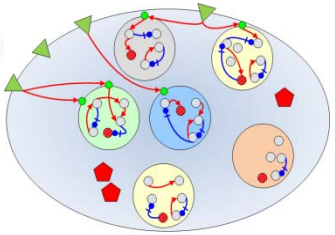
Earth Science



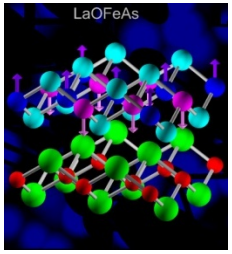
Health



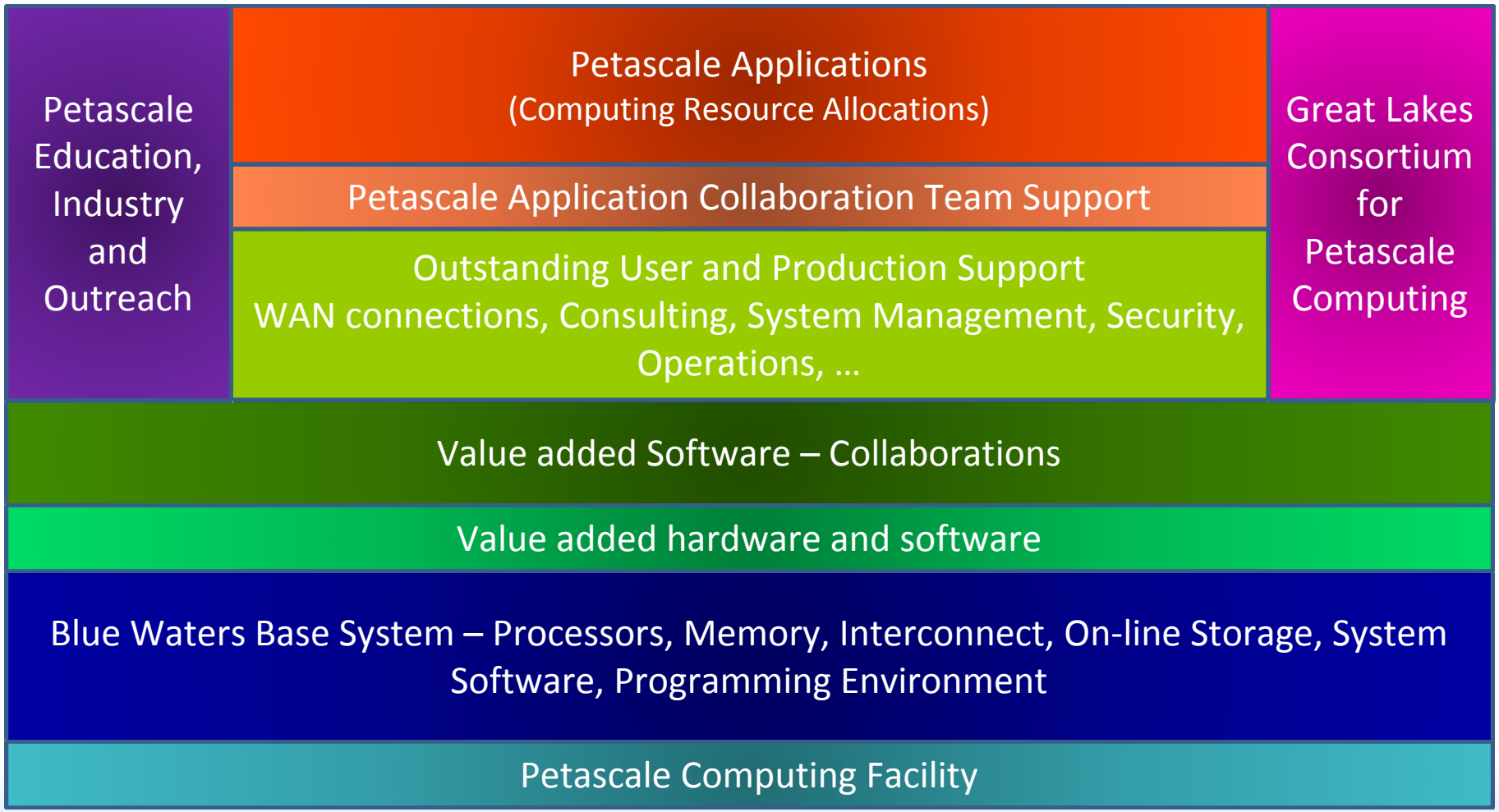
Life Science



Materials

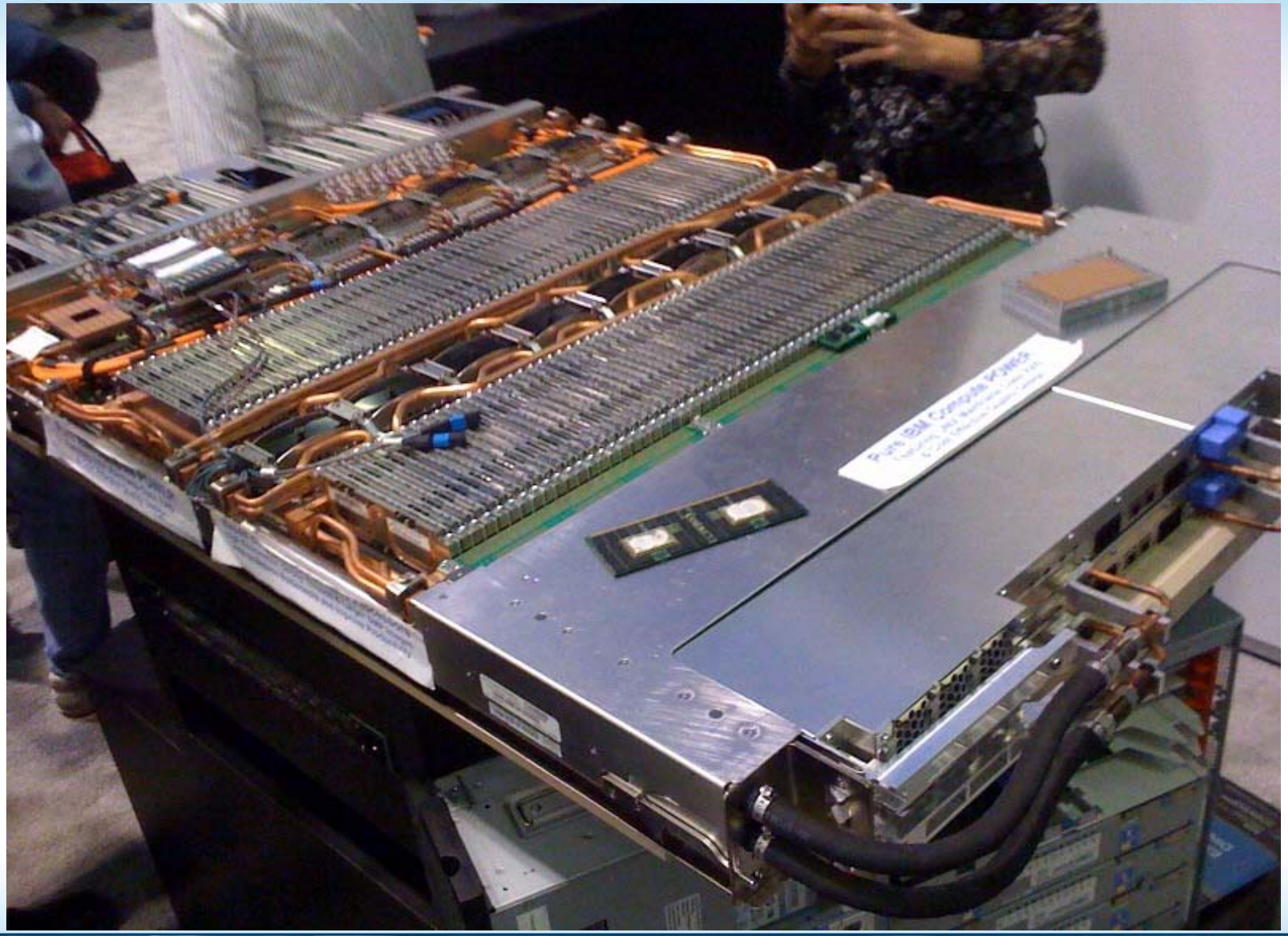


Blue Waters Project Components



Focus on Sustained Performance

- **Blue Water's and NSF are focusing on *sustained* performance in a way few have been before.**
- *Sustained* is the computer's performance on a broad range of applications that scientists and engineers use every day.
  - Time to solution is the metric – not Ops/s
  - Determined with real applications that include time to read data and write the results
- NSF's call emphasized sustained performance, demonstrated on a collection of application benchmarks (application + problem set)
  - Not just simplistic metrics (e.g. HP Linpack)
  - Applications include both Petascale applications (effectively use the full machine, solving scalability problems for both compute and I/O) and applications that use a fraction of the system
  - Metric is the time to solution
- Blue Waters project focus is on delivering sustained PetaFLOPS performance to all applications
  - Develop tools, techniques, samples, that exploit all parts of the system
  - Explore new tools, programming models, and libraries to help applications get the most from the system



PACTs = Required Benchmarks

- **Petascale Application Collaboration Team**
  - Formed around each required benchmark in NSF solicitation for sustained petascale system
- **Three petascale applications/problem sizes**
  - Lattice-Gauge QCD (MILC)
  - Molecular Dynamics (NAMD)
  - Turbulence (DNS3D)
- **Ultimate Milestone**
  - Time-to-solution target (or 1 PFLOP sustained) for specified problem (size, time, physics, method)
- **Three non-petascale applications/problem sizes**
  - Lattice-Guage QCD (MILC)
  - Materials Science (PARATEC)
  - Climate modeling (WRF)

PRAC Program

- Petascale Computing Resource Allocations
  - NSF to allocate Blue Waters time primarily through PRACs
  - Selected by NSF based on
    - Need for sustained petascale platform to carry out ground-breaking research
    - Likely to be ready to use Blue Waters effectively in 2011
- PRAC awardees receive travel funds and “provisional time”
- Awardees (total ~ 36 before Blue Waters operation)
  - Announcement of first round to be completed soon (~6 more expected)
  - Will accept applications on a continuing basis in future
- Blue Waters application and consulting staff will support awardees in preparing codes

Service Balance

- Job and resource scheduling to enable jobs to run for long blocks of time on large numbers of processors (as "determined by user requirements").
- It was expected that 20-50 percent of the system would be used by a single application most of the time
  - 50 percent or more of the system may be used for shorter periods.
- Scientific utilization was expected to be 95 percent of the available time used for "petascale science."

File System is GPFS

- IBM is implementing scaling changes in GPFS for the HPCS/DARPA project.
- Blue Waters will implement those changes in a persistent manner
- GPFS configured to accommodate other local systems in a single namespace
- Performance requirements are appropriately scaled to BW characteristics

Archive is HPSS

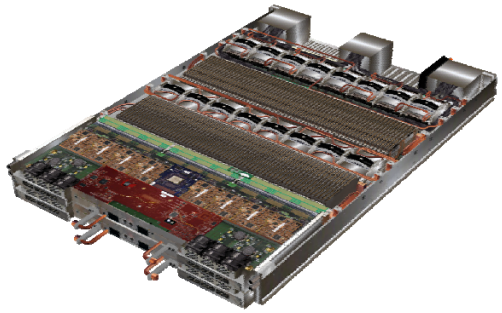
- HPSS Hardware consists of three tape robots and appropriate numbers of tape drives
  - Expect to expand this thru the lifetime of BW
- HPSS integrated with BW
  - GPFS-HPSS Interface
  - Import-Export Portal
    - Traditional HPSS commands
- NCSA is contributing RAIT implementation to the HPSS community as part of BW

- A core part of a new operational concept
  - Transparent data management for Users
  - “Virtual file system” for very large data
  - Improved productivity and schedule effectiveness
  - Lighter-weight backup

IBM Power Product Offerings

POWER Platforms

- Production ready, ultra reliable
- Market leader – sustained application performance
- Blades scaling to large memory SMP
- Rich s/w stack (from PERCS)
- Fast interconnect
- Very dense packaging



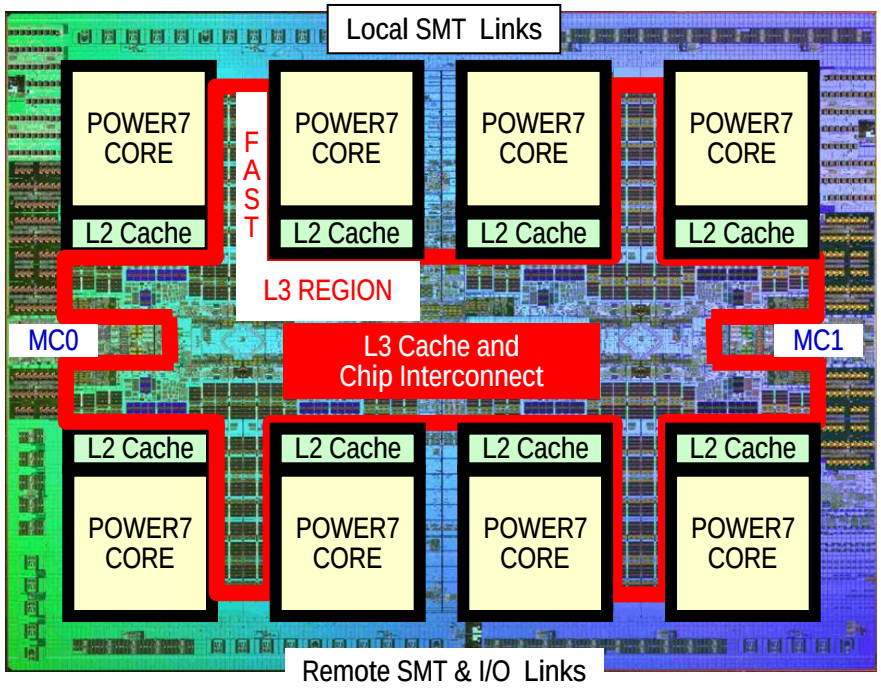
256 Core Nodes
Water Cooled

4U 32 Core
Air Cooled

Blades

POWER7

POWER7 Processor Chip



Binary Compatibility with POWER6

Core options: 8 (For HPC)

567mm^2 Technology:
- 45nm lithography, Cu, SOI, eDRAM

Transistors: 1.2 B
- Equivalent function of 2.7B
- eDRAM efficiency

Eight processor cores
- 12 execution units per core
- 4 Way SMT per core
- 32 Threads per chip
- 256 KB L2 per core

32MB on chip eDRAM shared L3

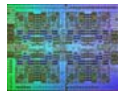
Dual DDR3 Memory Controllers
- 100 GB/s Memory bandwidth per chip

Scalability up to 32 Sockets
- 360 GB/s SMP bandwidth/chip
- 20,000 coherent operations in flight

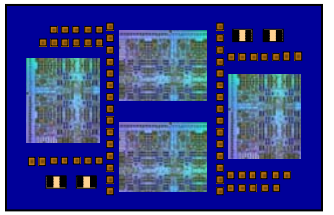
Advanced pre-fetching Data and Instruction

PERCS POWER7 Hierarchical Structure

- **POWER7 Chip**
  - 8 Cores

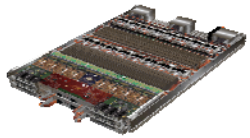


- **POWER7 QCM & Hub Chips**
  - QCM: 4 POWER7 Chips
    - 32 Core SMP Image
  - Hub Chip: One per QCM
    - Interconnect QCM, Nodes, and Super Nodes

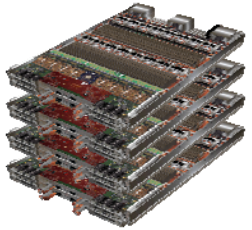


Hub Chip

- **POWER7 IH Node**
  - 2U Node
  - 8 QCMs
    - 256 Cores

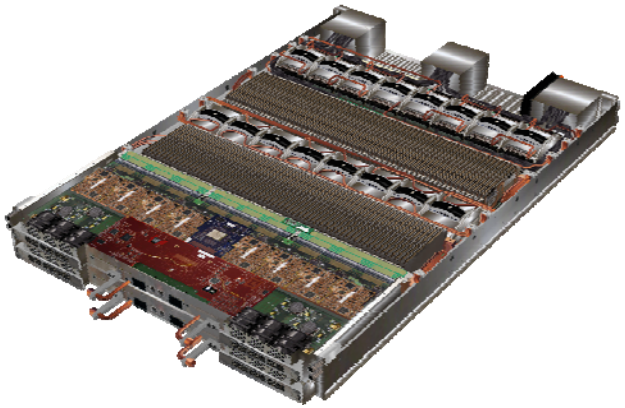


- **POWER7 'Super Node'**
  - 4 Drawers / Nodes
    - 1024 Cores



- **Full System**
  - Up to 512 'Super Nodes'
  - 512K Cores

Planned POWER7 Compute Node



Chip Performance: ≥224 GFLOPS

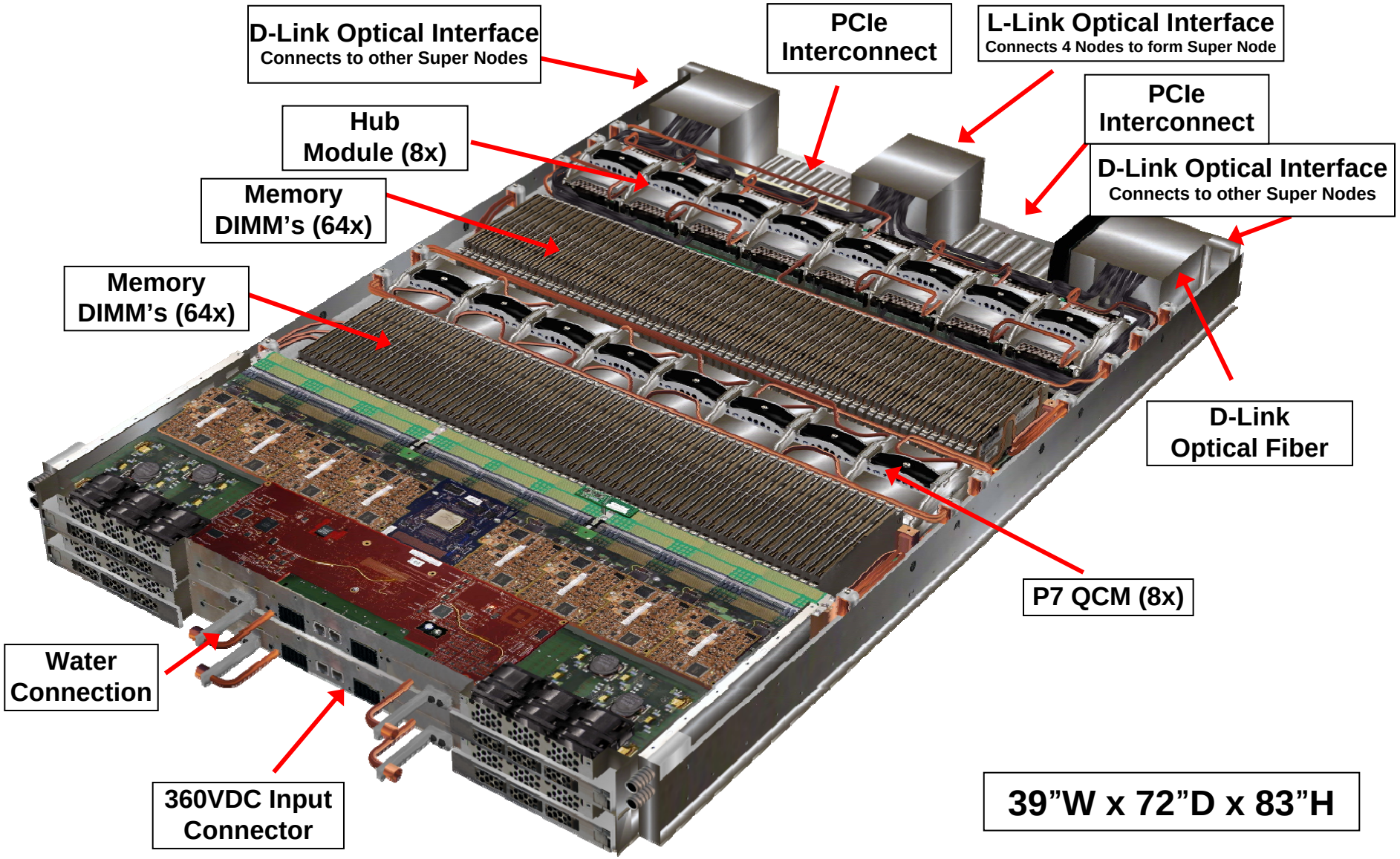
- 8 Cores per Chip
- Core Freq: 3.7GHz+
- 4 Floating Point Units (FPU) per core
- 2 FLOPS/Cycle
- 8 cores x 3.7 GHz x 4 FPU/Core x 2 Flops/Cycle

Node Performance: ≥7.6 TF w/ Integrated SMP Fabric

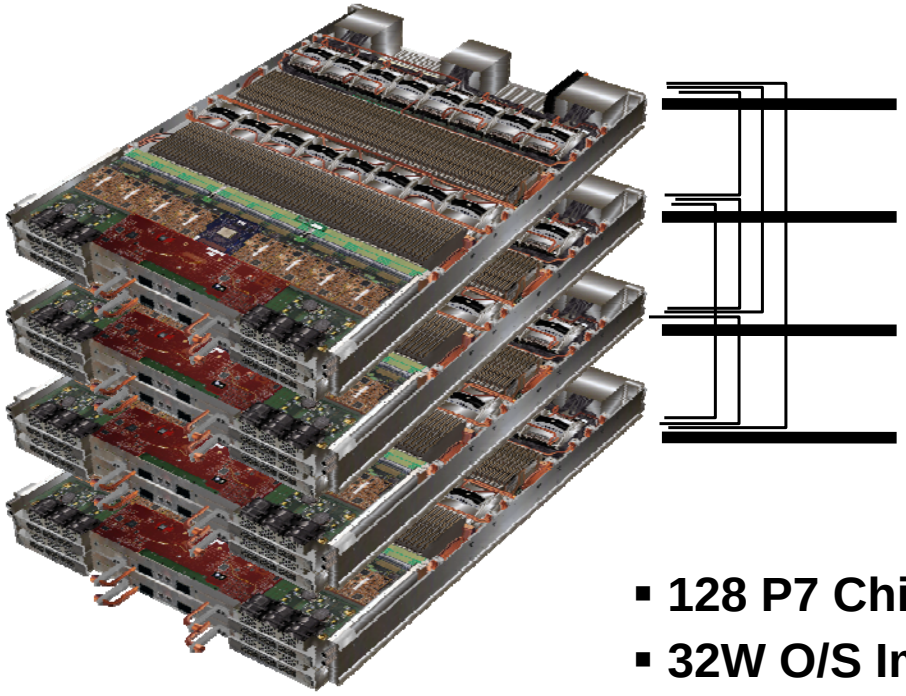
- *256 Cores*
- *32 Chips x ≥237 GFLOPs*

POWER7 Compute Node	
Nodes	Up to 12 per rack
Architecture	POWER7 256 Cores / Node
Cache	On Chip L2 & L3
DDR3 Memory	128 DIMM Slots / Node Up to 2 TB / Node Up to 24 TB / Rack
PCI Expansion / Node	16 – 16X PCIe Gen 2, 1 - 8X PCIe Gen 2
Storage Enclosure	Up to 6 per rack Up to 384 SFF Drives / Drawer
Ethernet / Node	Up to 16 Quad Port 1 Gb Up to 16 Dual Port 10/100
Cluster Attach	PERCS Interconnect Fabric
Power	N+1 Line Cords
Cooling	Water (100% Heat capture)

P7 IH System Hardware – Node Front View



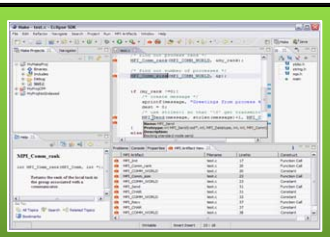
PERCS POWER7 Super Node Description

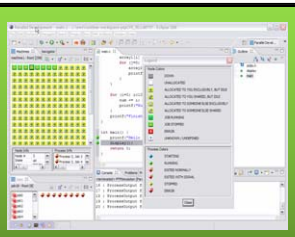


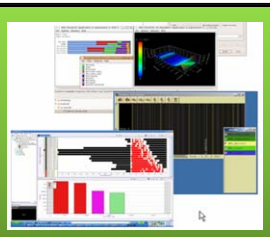
Optical D-Links

- 128 P7 Chips + 32 Hub Chips
- 32W O/S Image
- 8 X 32Ws Coherent SMP & I/O Size
- 4 Compute Node (256Cores / Node)
- 8U in 30” Rack
- 32 SCMs + 32 QCMs
- 1024 Cores
- Up to 8 TB Memory

HPC Software Stack







Eclipse PTP Workbench

APPLICATIONS

C, C++ OpenMP	Fortran (77, 95) OpenMP, CAF	ESSL/PESSL Math Libs	UPC	Global Arrays	MPI	Open Shmem

xCAT Plus

On-demand Provisioning

Rolling Updates

Energy Optimization

LoadLeveler
Local, Multi-cluster & Scale-across Scheduling
Resource Monitoring & Management

Task Migration

Job Placement and Execution

GPFS

Storage mgt

Node config & mgt
VM / Bare-metal deployment

POE Runtime

Metacluster
Checkpoint / Restart

MPI

PAMI

Topology awareness

Network Management

PowerVM

POWER Systems

POWER Systems

BlueGene Q

KVM ESX Xen

X86 Systems

X86, Systems

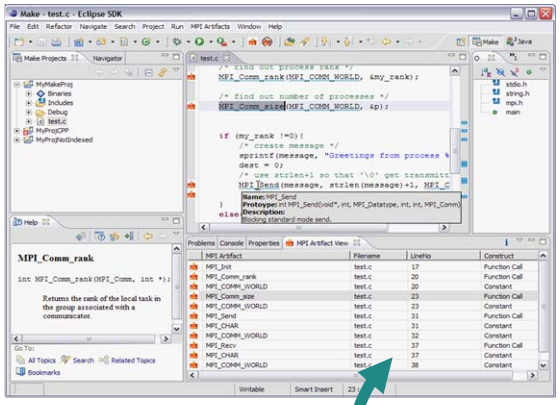
Storage Servers, Disks

Network Switch

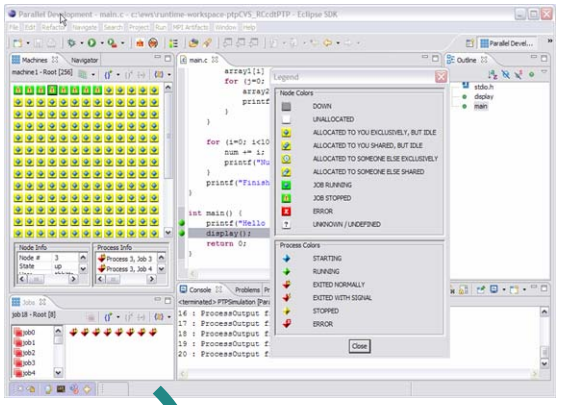
Tape Storage

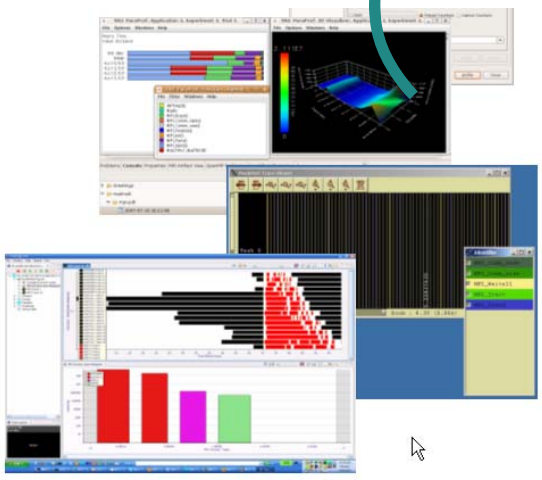
Advanced Application Development Workbench

Coding & Analysis Tools

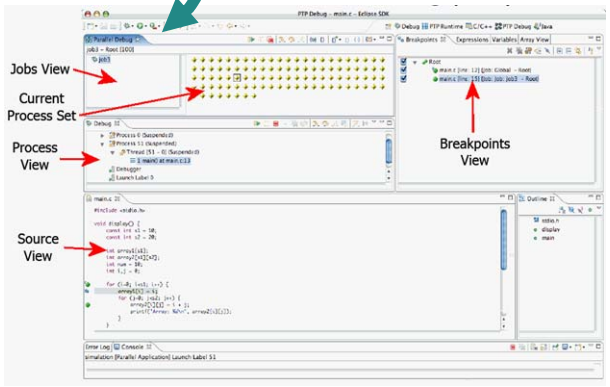


Launching & Monitoring Tools





Performance Tuning Tools



Debugging Tools

Portfolio

- **POWER Platforms**
  - Production ready, ultra reliable
  - Market leader – sustained application performance
  - Large memory SMP
  - Rich s/w stack (from PERCS)
  - Fast interconnect
  - Very dense packaging
- **Blue Gene**
  - Production ready, ultra reliable
  - Ultra high scaling capability
  - Fast interconnect
  - Highly energy efficient
  - Very dense packaging
  - Strong PEAK $/Mflp price/performance
- **X86 Clusters**
  - Focused on “capacity”, scalability
  - High ISV coverage
  - Strong PEAK $/Mflp price/performance