BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Blue Waters

A Super-System to Explore the Expanse and Depth of 21st Century Science

Thom Dunning, William Kramer, Marc Snir,
William Gropp, Wen-mei Hwu

Cristina Beldica, Brett Bode, Robert Fiedler, Merle Giles,
Scott Lathrop, Mike Showerman

National Center for Supercomputing Applications, Department of Chemistry, Department of Computer Science, and Department of Electrical & Computer Engineering

NCSA NSF IBM GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

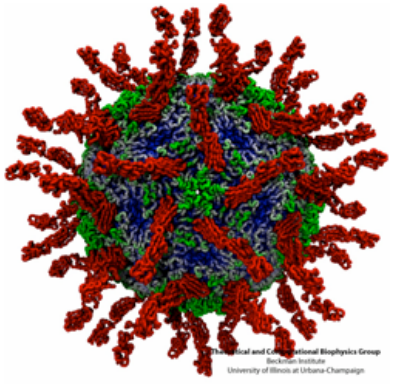
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

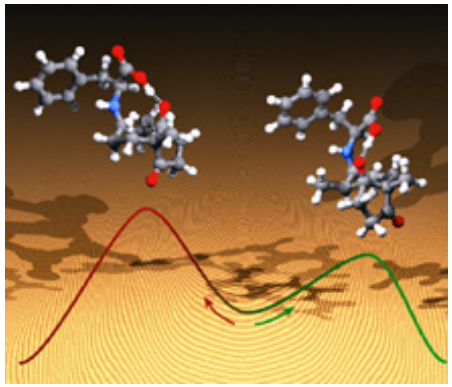
GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Science & Engineering on Blue Waters

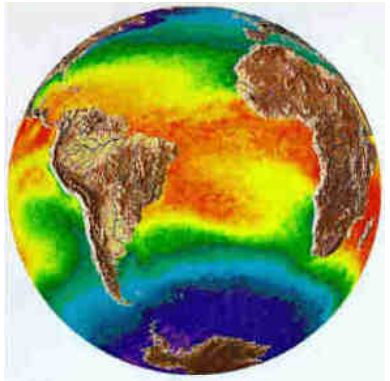
Blue Waters will enable advances in a broad range of science and engineering disciplines. Examples include:

Molecular Science



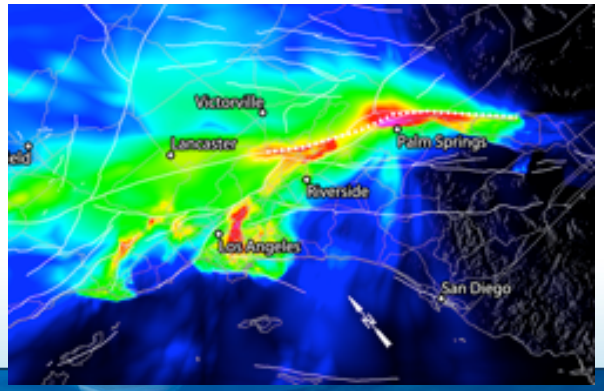


Weather & Climate Forecasting

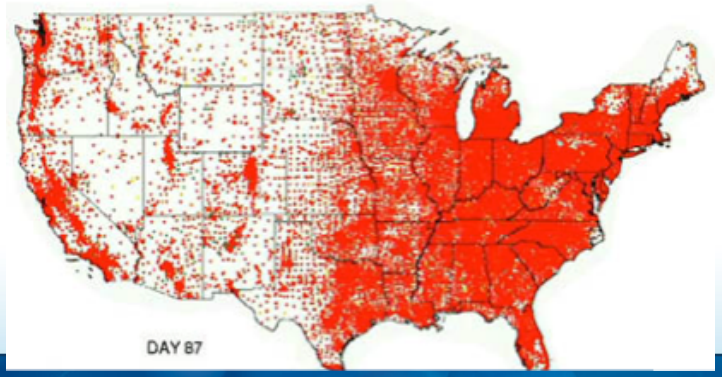


Astronomy

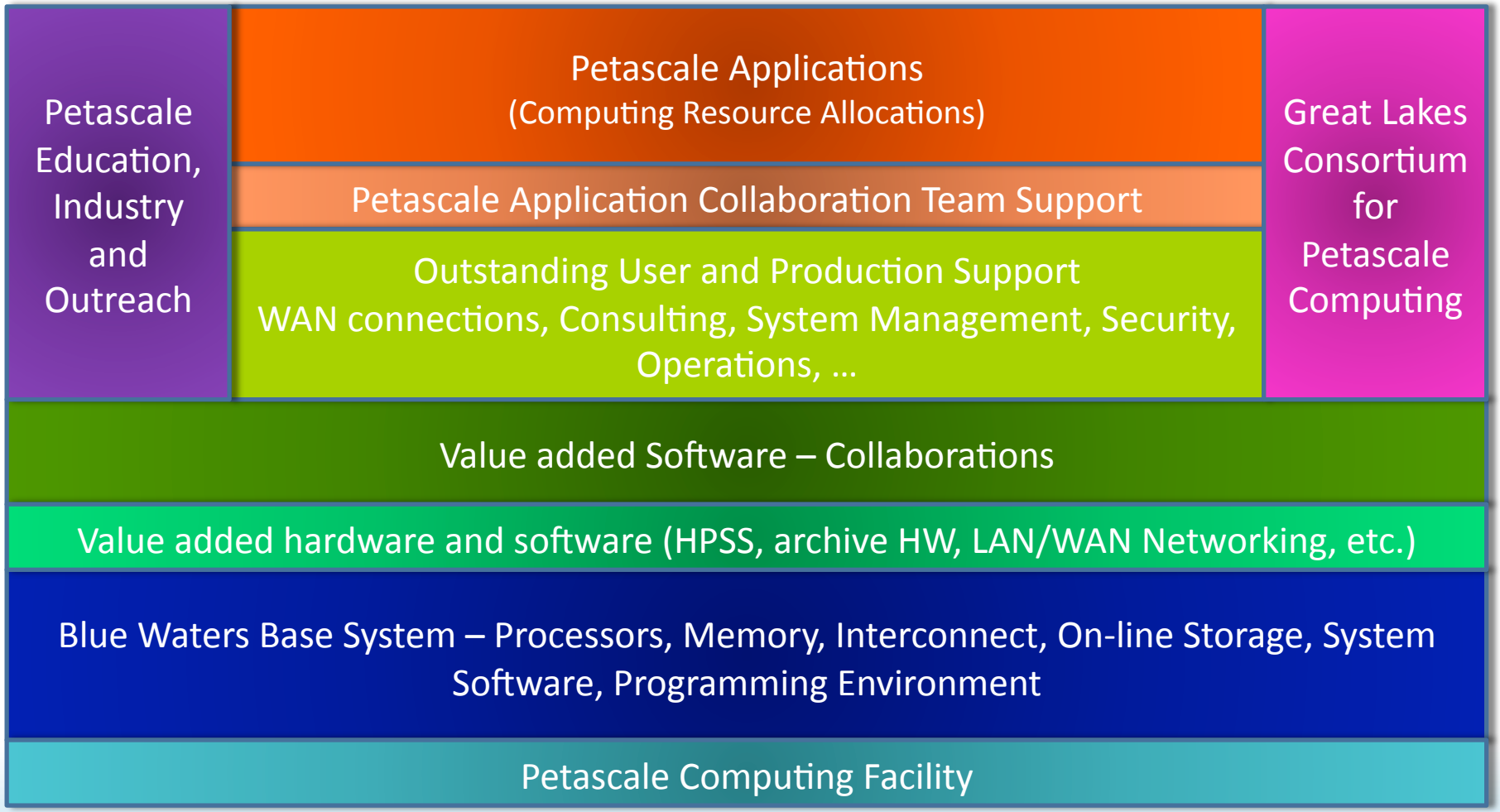
Earth Science



Health



Blue Waters Project Components



Petascale Computing Facility

Partners

EYP MCF/
Gensler
IBM
Yahoo!

- **Modern Data Center**
  - 90,000+ ft^2 total
  - 30,000 ft^2 raised floor
    20,000 ft^2 machine room
- **Energy Efficiency**
  - LEED certified Gold (goal: Platinum)
  - PUE = 1.1–1.2

Resource manager: Batch and interactive access

Performance tuning: HPC and HPCS toolkits, open source tools

Parallel debugging at full scale

Environment: Traditional (command line), Eclipse IDE (application development, debugging, performance tuning, job and workflow management)

Languages: C/C++, Fortran (77-2008 including CAF), UPC

Libraries: MASS, ESSL, PESSL, PETSc, visualization…

Programming Models: MPI/MP2, OpenMP, PGAS, Charm++, Cactus

Low-level communications API supporting active messages (PAMI/LAPI)

IO Model: Global, Parallel shared file system (>10 PB) and archival storage (GPFS/HPSS) MPI I/O

Full – featured OS Sockets, threads, shared memory, checkpoint/restart

Hardware
Multicore POWER7 processor with Simultaneous MultiThreading (SMT) and Vector MultiMedia Extensions
Private L1, L2 cache per core, shared L3 cache per chip
High-Performance, low-latency interconnect supporting RDMA

Other Project Activities

- 9 collaboration teams – for areas beyond SOW deliverables
  - Compiler Development, Advanced Tools, Communication Infrastructure for Tools, Workflow and Data staging, Advanced Programming Models, System Monitoring, Advanced Storage, Cyber-Security, MERCURY-BigSim Comparison, Network Topology and Routing
- Education Program
  - Virtual School, Fellowships, Instructor Training, Content Creation
- Private Sector Partners

APPLICATIONS USING BLUE WATERS

Summary of Science ----- Unique Features of Blue Waters

Earthquake Engineering – *Rupture to Rumble*

- **Science Goals**
- Southern California Earthquake Center is modeling Southern California's next catastrophic earthquake
- Everyone from engineers to policymakers will use the results.
- Current simulations capture seismic waves in 1Hz range. Blue Waters model a 2Hz range. Scientists want 10Hz range.
- Estimate Blue Waters simulations will be more than 256 times more computationally demanding than current simulations.

- **Blue Waters Features for Earthquake**
- Large cache memory (1.2 TB) allows aggressive cache blocking and extremely large meshes.
- High memory bandwidth (5 PB/s) and large main memory (1.2 PB).
  - Extra large memory will reduce the risk of loss of critical data from I/O failure and improve performance.
- Interconnect balanced with cores
- Overlapping communications-computation.
- Integrated online/nearline storage dramatically reduces storage management.
- High bandwidth connectivity to national networks will allow the required transfer of many terabytes of data among centers.

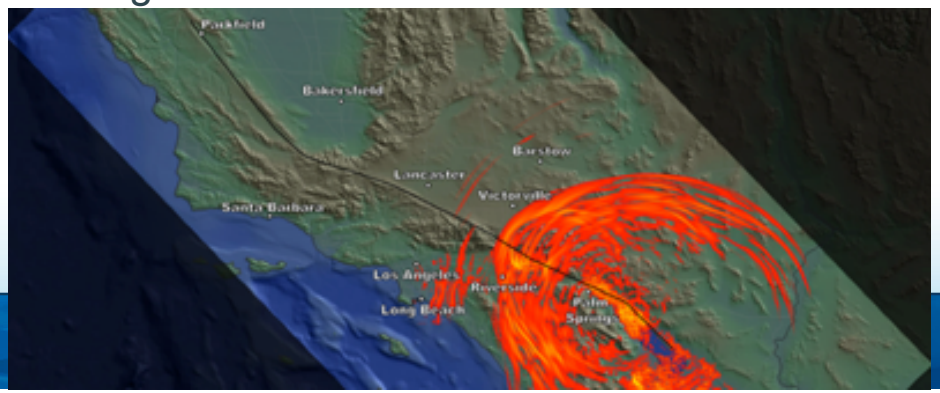


Image by: Kim Olsen, San Diego State University, Yifeng Cui and Amit Chourasia, San Diego Supercomputer Center

Cosmology

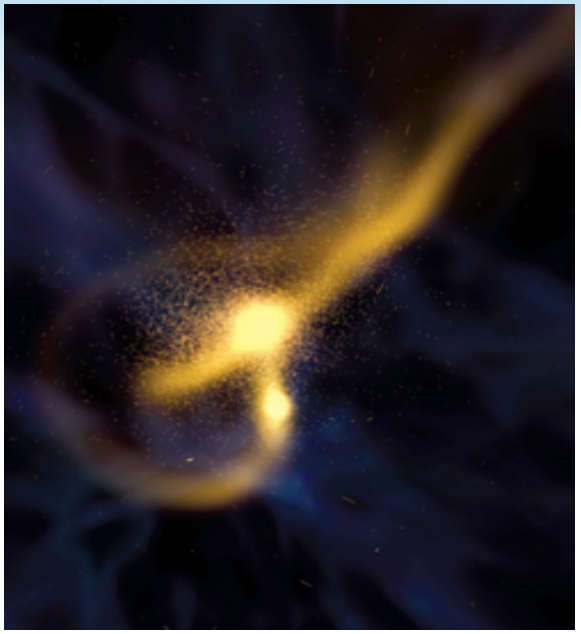


Image by: Brian O'Shea, Michigan State University

- **Science Goals**
  - Science team will model the first billion years after the Big Bang—a time in the universe that is little understood.
  - Make predictions about what upcoming instruments like the Webb telescope will detect.
  - Need to simulate hundreds of thousands of galaxies instead of a thousand galaxies.
- **Blue Waters Features for Cosmology/Astronomy**
  - With high-performance Interconnect and communication software, team will be able to scale their codes and include a broad range of physical phenomena.
  - With high-performance I/O system, team will save terabyte snapshot to disk in less than a second
  - Improve time-to-solution by working closely with Blue Waters and IBM personnel to optimize their code to take full advantage of the memory hierarchy and vector units in POWER7, as well as improve code's ability to exploit the 128 GB of shared memory per node.

Epidemiology

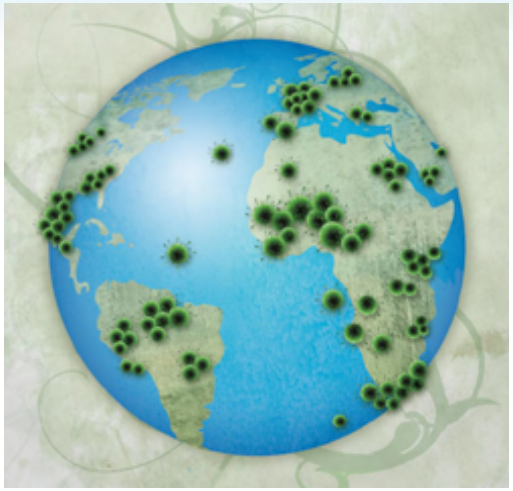


Image by: NCSA

- **Science Goals**
- Science team expects to model global disease outbreaks, as well as smaller.
- Simulate up to six billion individuals as they travel, interact, etc.
- Results used by policymakers for planning and response to outbreaks
- Running a global model on on contemporary supercomputer (~2,000 processors) would take 2 years.
- "Overall time-to-solution is the measure of effectiveness. Today, many policymakers are forced to use inaccurate tools in place of accurate, but slower, tools."
- **Blue Waters Features for Epidemiologists**
- Low latency and high bisection bandwidth of interconnect reduces load imbalances and time-to-solution—critical for irregular interaction patterns of these agent-based simulations.
- Large amounts of memory per core allows tremendous detail at very small scales.

Severe Weather

- **Science Goals**
- Understanding *tornadogenesis*—the process by which a tornado forms
- Forecasters can identify conditions that are make a tornado likely, but they want to pinpoint when and where the start, their path, and strength.
- Ultra-high resolution (~10m) needed for small-scale features that influence the evolution of the tornado
  - Thin curtains of precipitation & jets of wind just above the ground.
- **Blue Waters Features for Atmospheric Sciences**
- Low-latency interconnect minimizes the amount of time the processors spend on communication-related tasks.
- Large, high-performance memory allows overlap of computation and communication.
- Features enables movement of the data to other nodes for in-line/real-time analysis, visualization, and steering.
- Large and fast on-line/near-line storage system critical for handling 10 petabytes of data per simulation for in-depth analysis.

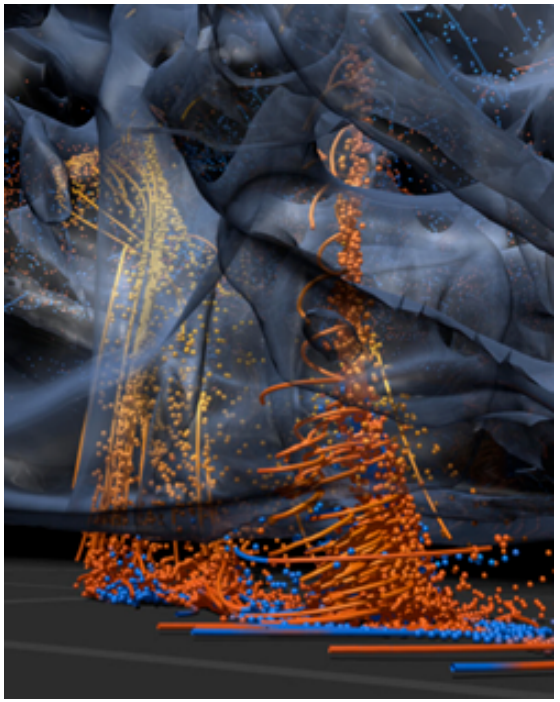


Image by: Robert Wilhelmson, University of Illinois at Urbana-Champaign; Lou Wicker, National Severe Storms Laboratory; NCSA's Advanced Visualization Lab

Biophysics

- **Science Goals**
- Science Team is investigating biomolecular processes that take a full millisecond—100 times longer than what they study today.
- Will help in the design of new antibiotics and antiviral drugs, as well as models for the solar production of electricity and fuels like hydrogen.
- **Blue Waters Features for Biophysics Challenges**
- NAMD will use a low-level communication library to avoid overhead introduced by high-level message passing libraries.
- Four independent execution threads in the POWER7 processor maximize the time spent actively performing floating-point operations.
- Large per-node memory will allow analysis using related VMD code directly on Blue Waters.

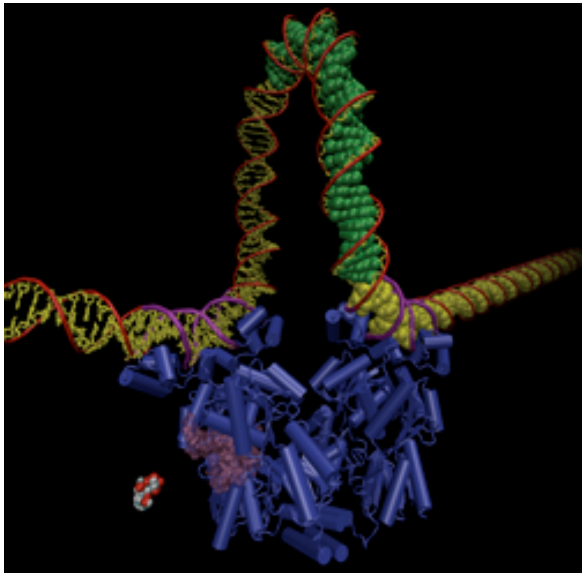


Image by: Klaus Schulten, University of Illinois at Urbana-Champaign

Chemistry

- **Science Goals**
- GAMESS and NWChem tuned to run on Blue Waters.
- Combined, they have well more than 100,000 users in about 100 countries.
- Allow the team to study ice formation, aerosol behavior, and dendrimers (polymers with medical and environmental applications)—along with many other phenomena.
- **Blue Waters Features for Chemists**
- Very powerful POWER7 processors help with electronic structure calculations, which scale exponentially with the number of atoms.
- Low-latency, high-bandwidth interconnect allows huge amounts of intermediate data to be distributed among processors quickly.
- Fast I/O system allows data to be accessed from disk rapidly.
- Large amount of memory per core enable calculations to use very efficient in-core algorithms, rather than much slower out-of-core techniques.

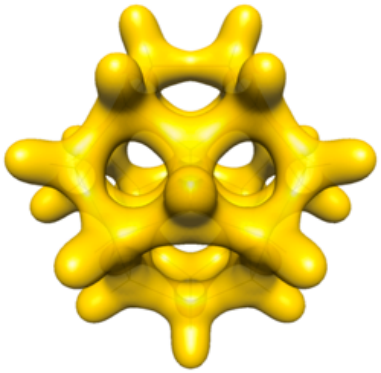


Image by: Munir Nayfeh, University of Illinois at Urbana-Champaign

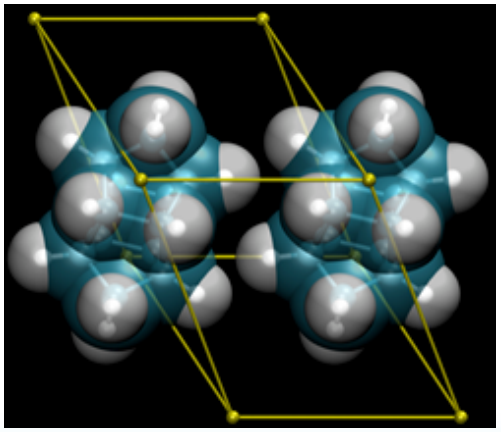
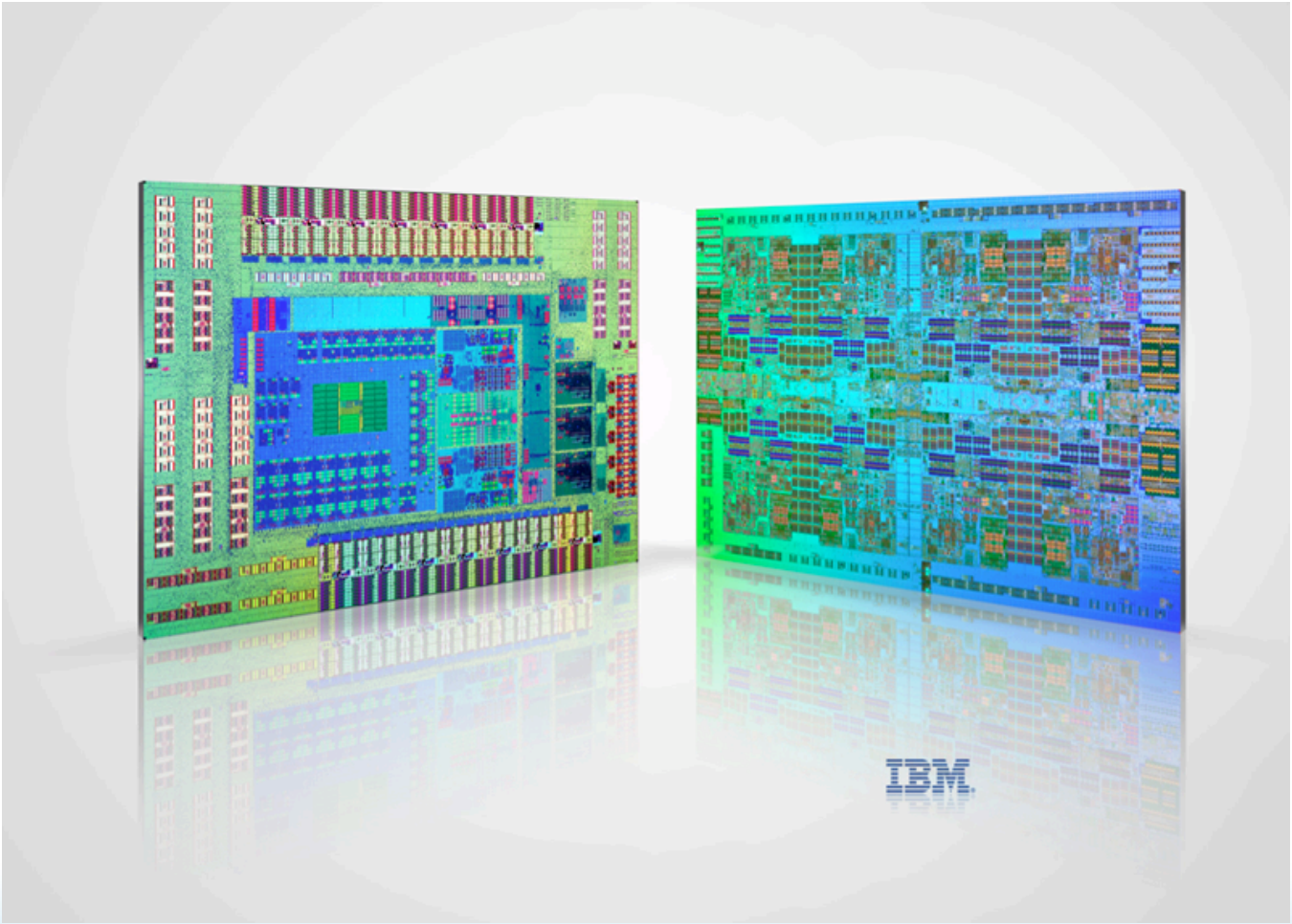
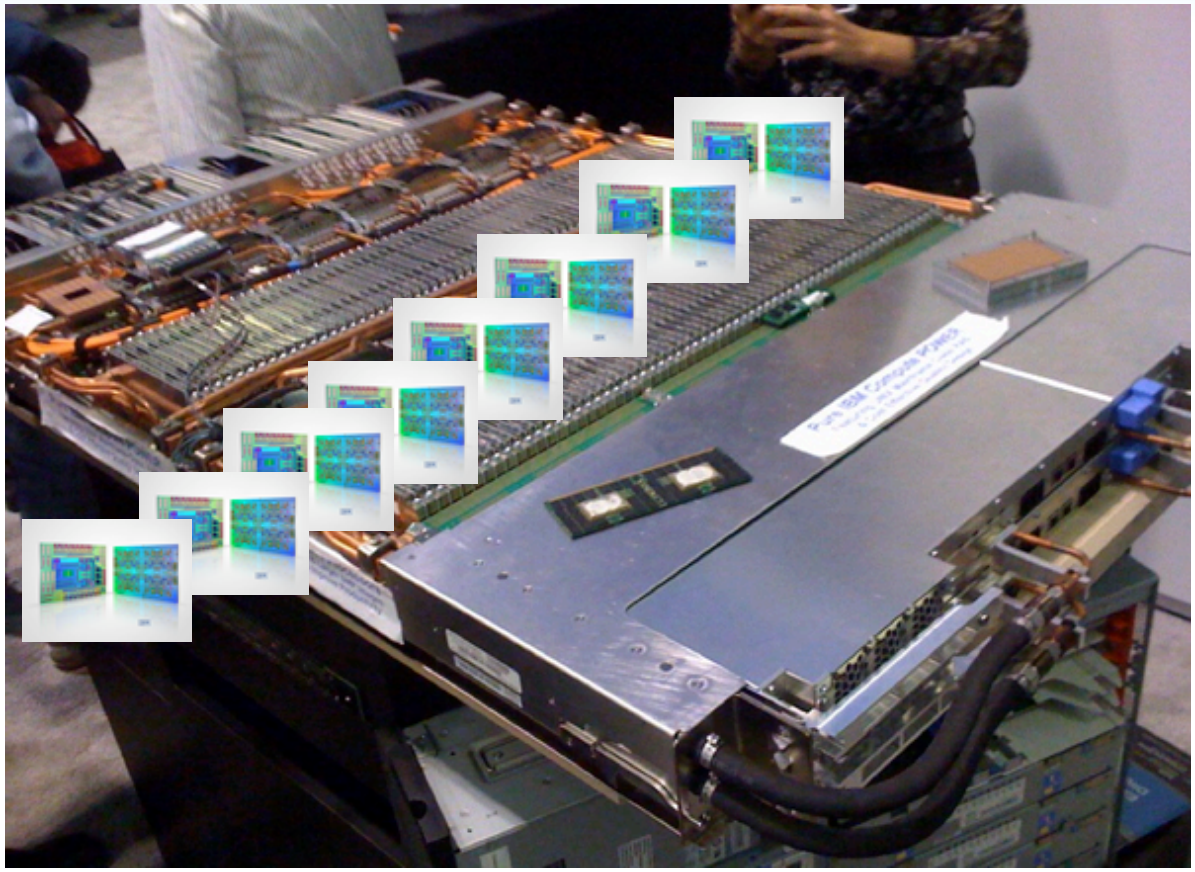


Image by: Bruce Hudson, Syracuse University

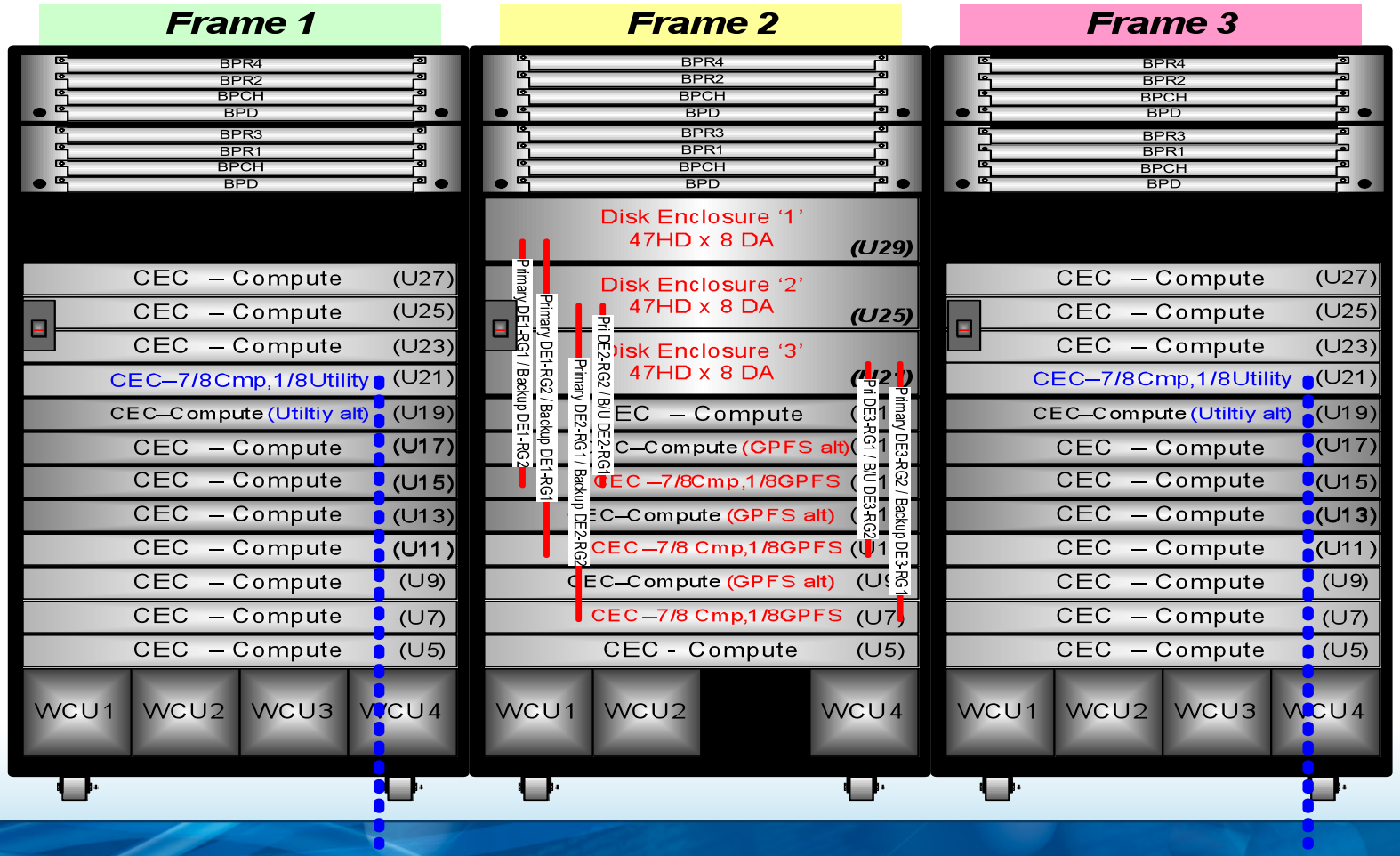
THE BLUE WATERS SYSTEM

Blue Waters *Super-System*



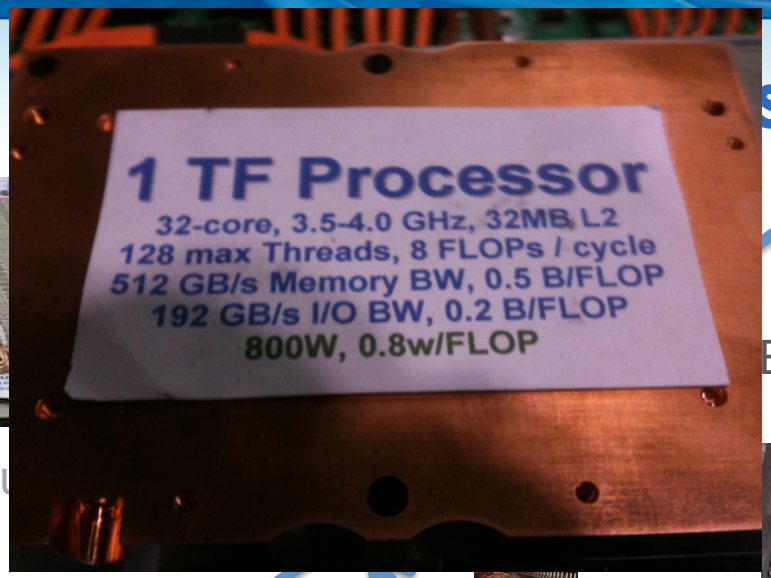


BW Building Block



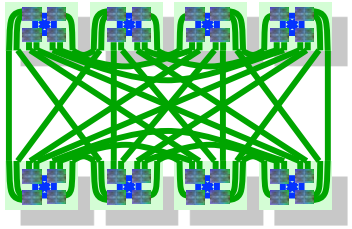
From Chip to stem

Rack/Bu



NPCF

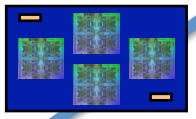
Blue Waters System

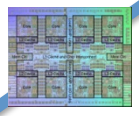


multiple MCMs

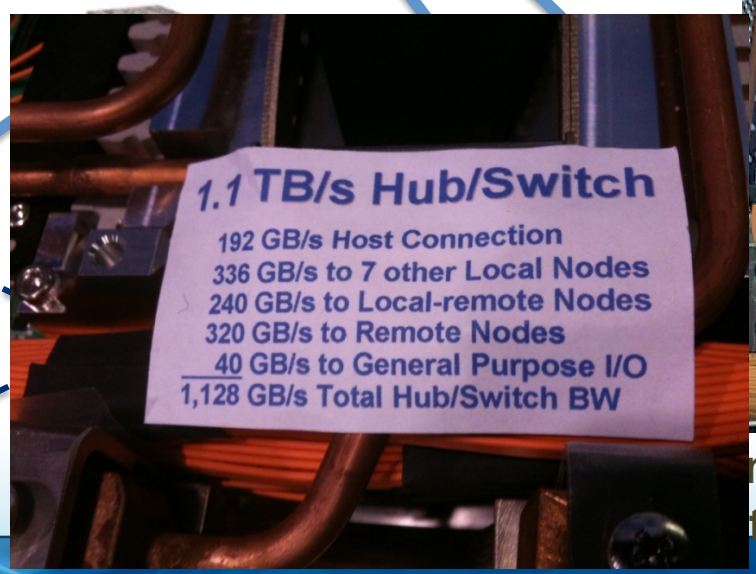
Near-li

Quad Chip MCM





Chip



ndicates relative
t of public information

Blue Waters Computing System

System Attribute	JAGUAR		Blue Waters
Vendor	CRAY XT5		IBM
Processor	AMD OPTERON		IBM Power7
Peak Performance (PF)	2.3	>4	>10
Sustained Performance (PF)			>1
Number of Cores/Chip	6	1.3	8
Number of Processor Cores	224,256	<1.2	>300,000
Amount of Memory (TB)	299	>4	1.2
Interconnect Bisection BW (TB/s)	~2	2	~1
Interconnect HW Latency (µs)		>>	
Amount of Disk Storage (PB)	5	>3	18
I/O Aggregate BW (TB/s)	.24	>6	>1.5
Amount of Archival Storage (PB)	20	>25	>500
External Bandwidth (Gbps)			100-400

Blue Waters Computing System

System Attribute	Ranger		Blue Waters
Vendor	Sun		IBM
Processor	AMD Barcelona		IBM Power7
Peak Performance (PF)	0.579	17	>10
Sustained Performance (PF)	<0.05	>20	>1
Number of Cores/Chip	4	2	8
Number of Processor Cores	62,976	~3.5	>300,000
Amount of Memory (TB)	123	~10	>1.2
Interconnect Bisection BW (TB/s)			~1
Interconnect HW Latency (µs)	~4	>>10	
Amount of Disk Storage (PB)	1.73	>10	18
I/O Aggregate BW (TB/s)	.03	>50	>1.5
Amount of Archival Storage (PB)	2.5 (20)	>200	>500
External Bandwidth (Gbps)	10	>10	100-400

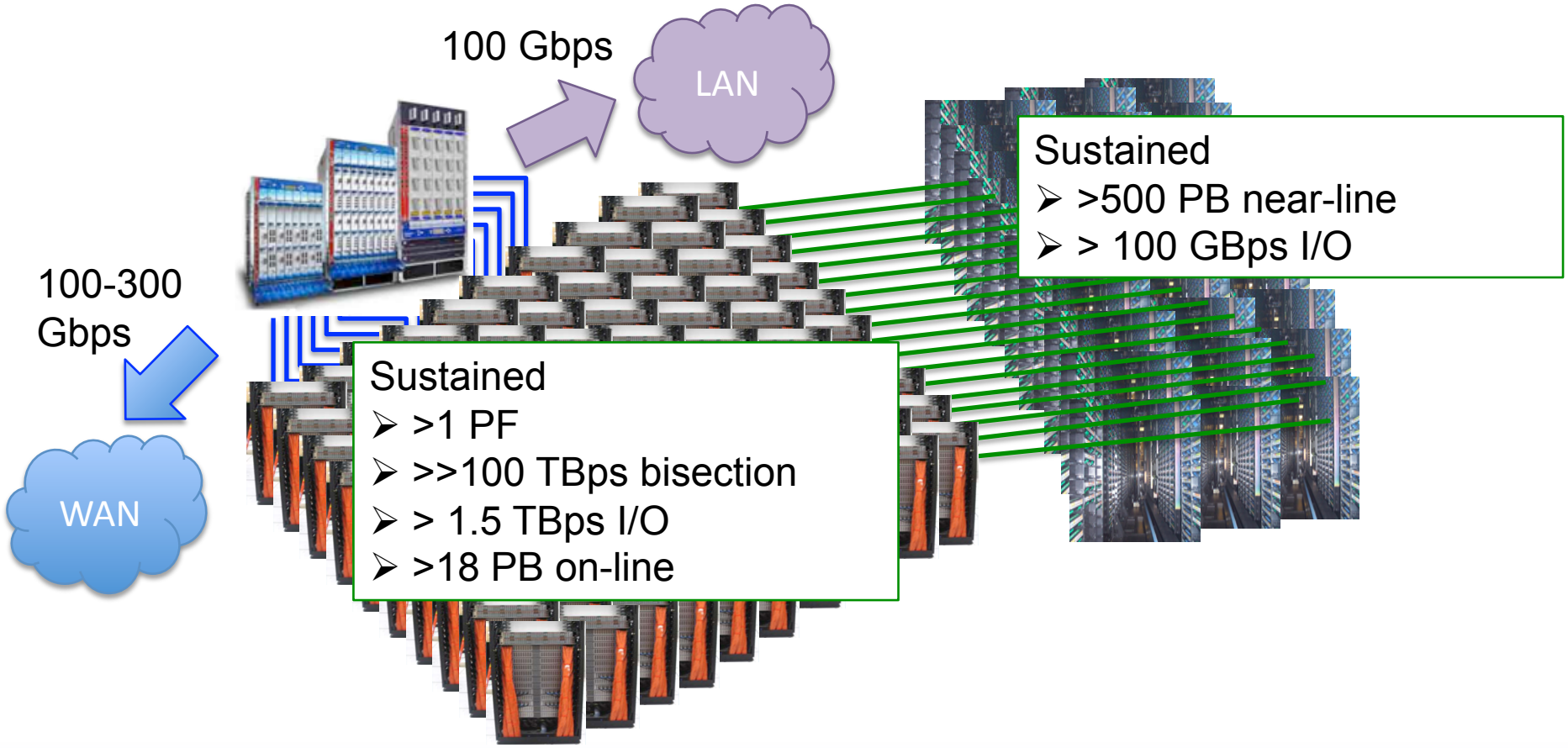
BW Innovative Storage Management

- Toughly coupled on-line and near line storage in one name space
- Declustered Raid for on-line storage
- File data is moved when one of the following occurs:
  - Migration from on-line to near-line
  - Recall on demand to on-line
  - Stage in conjunction with workflow
  - Backup – low overhead
    - No single point of failure
  - Restore
- All data transfers are distributed/parallel and multithreaded
  - Configurable
- Traditional archive use

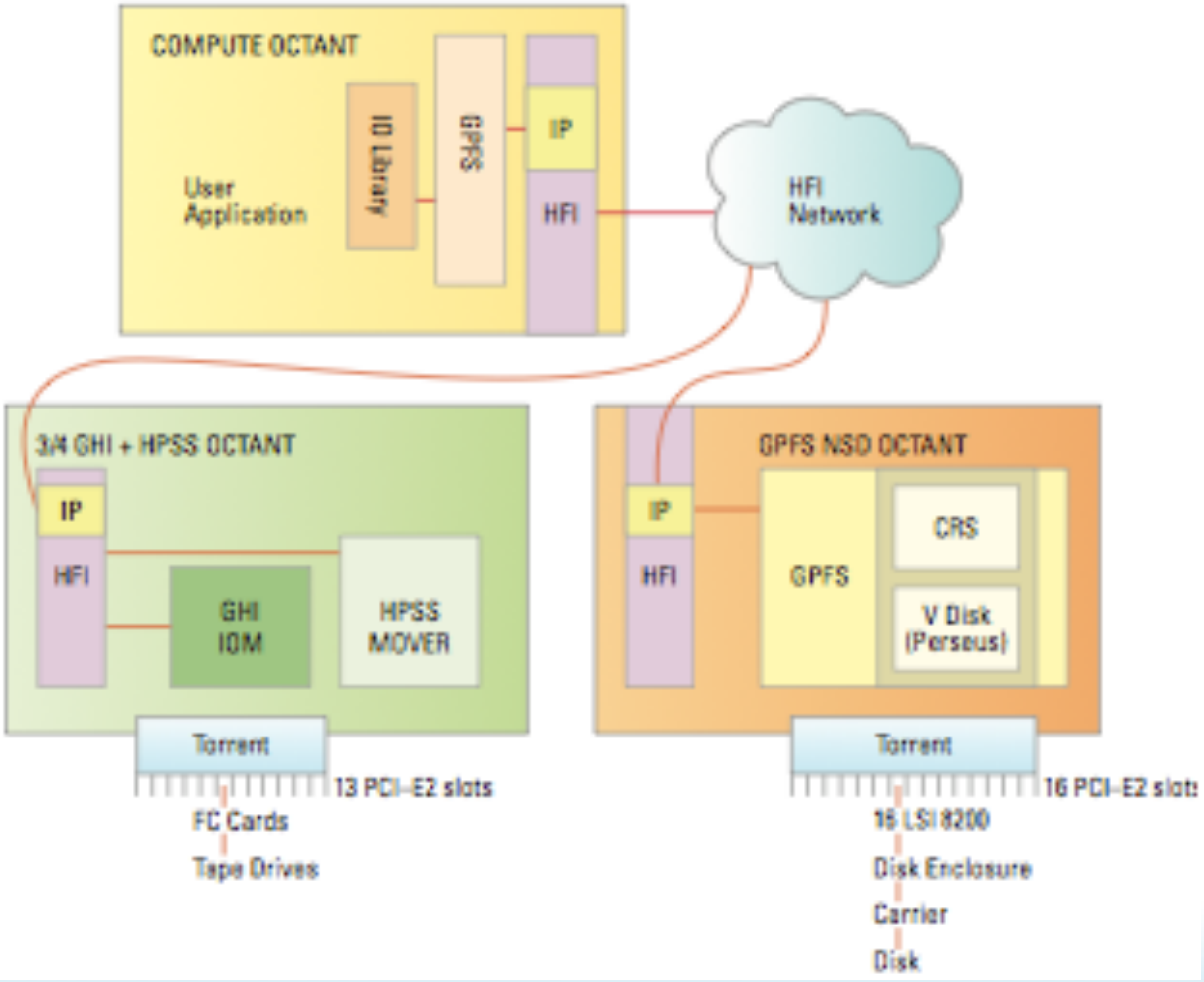
P7-IH I/O Subsystem (4 major areas)

- **HFI Network - Single network in system for all I/O traffic as well as communication**
  - Storage network (File System)
  - Archival network (HPSS Tape)
  - Networking (Gateway & Log-on Customer LAN)
- **Storage Subsystem**
  - GPFS NSD Nodes
  - SAS adapters in CEC Drawers containing Storage Nodes
  - SAS Attached Disk Enclosure
  - GPFS Cluster Manager Nodes
- **HPSS Mover Nodes**
- **Networking Nodes**

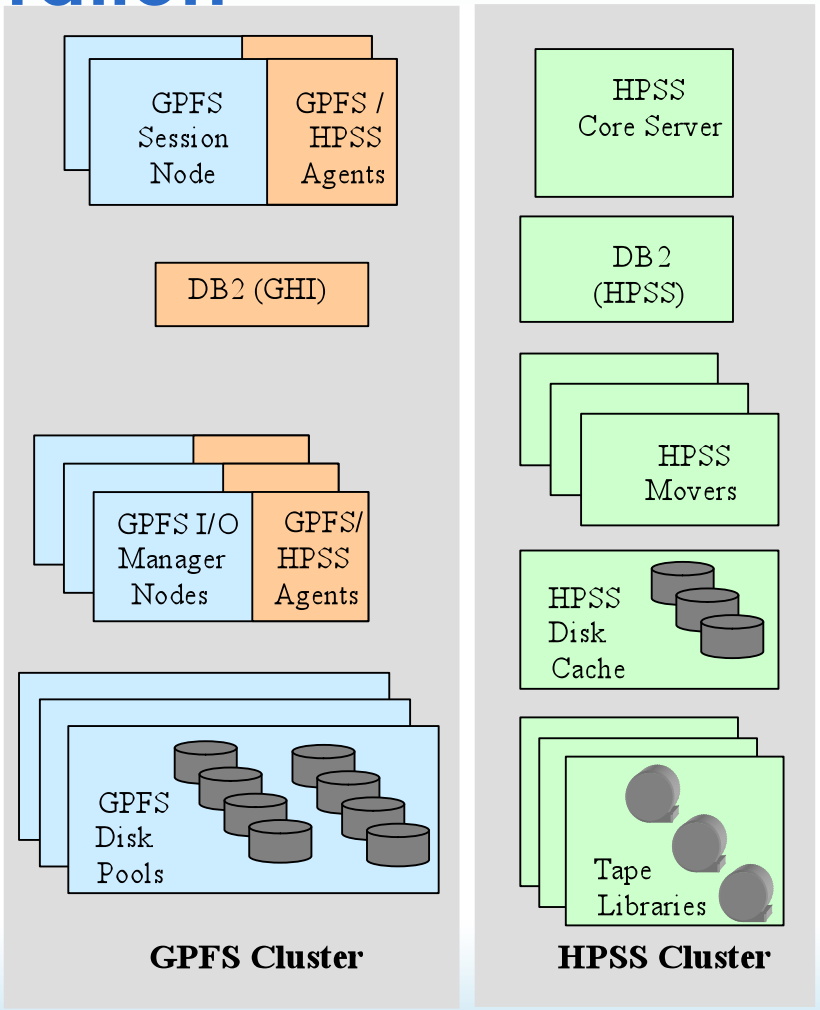
Blue Waters *Super-System*



I/O Storage Architecture



GHI Configuration



Summary

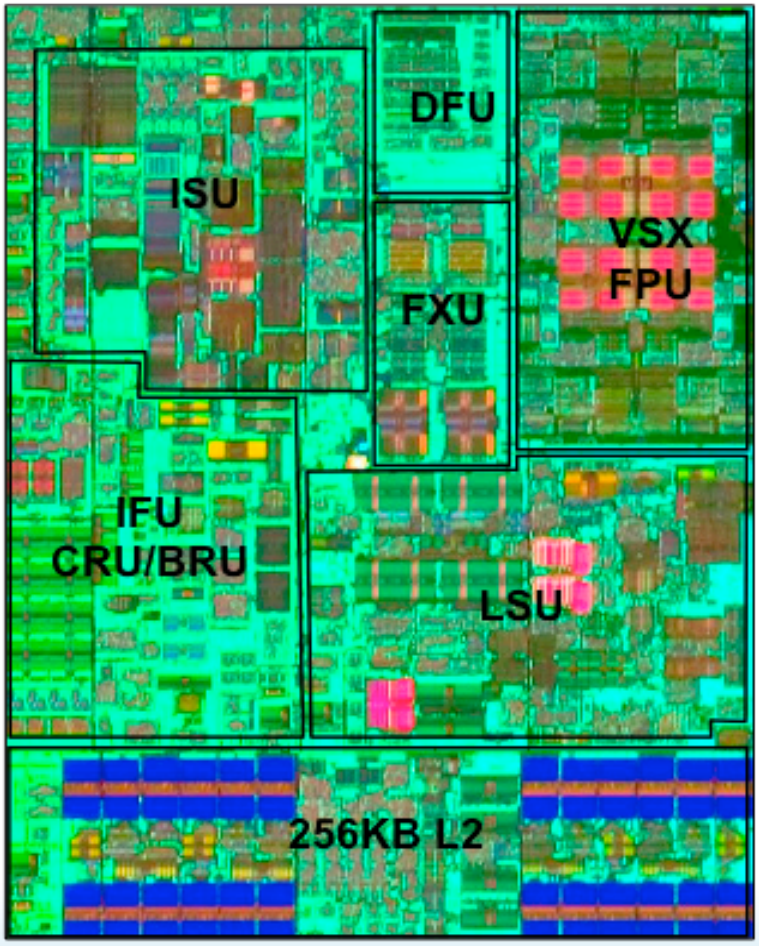
- The Blue Waters project is going very well
- National Petascale Computing Facility completed
  - Workshop on large Computational Facilities this week
- First hardware testing (processor) is underway
- 19 science teams identified and are engaged
  - Expect 10-15 more indentified by the end of the summer
  - Work for BW also benefits other systems
- Received first Power 7 hardware – a p780 – that is being used by science teams for single node (SMP) performance tuning
- Software efforts are continuing and some open SW is already delivered

2011 will be a “very interesting year”

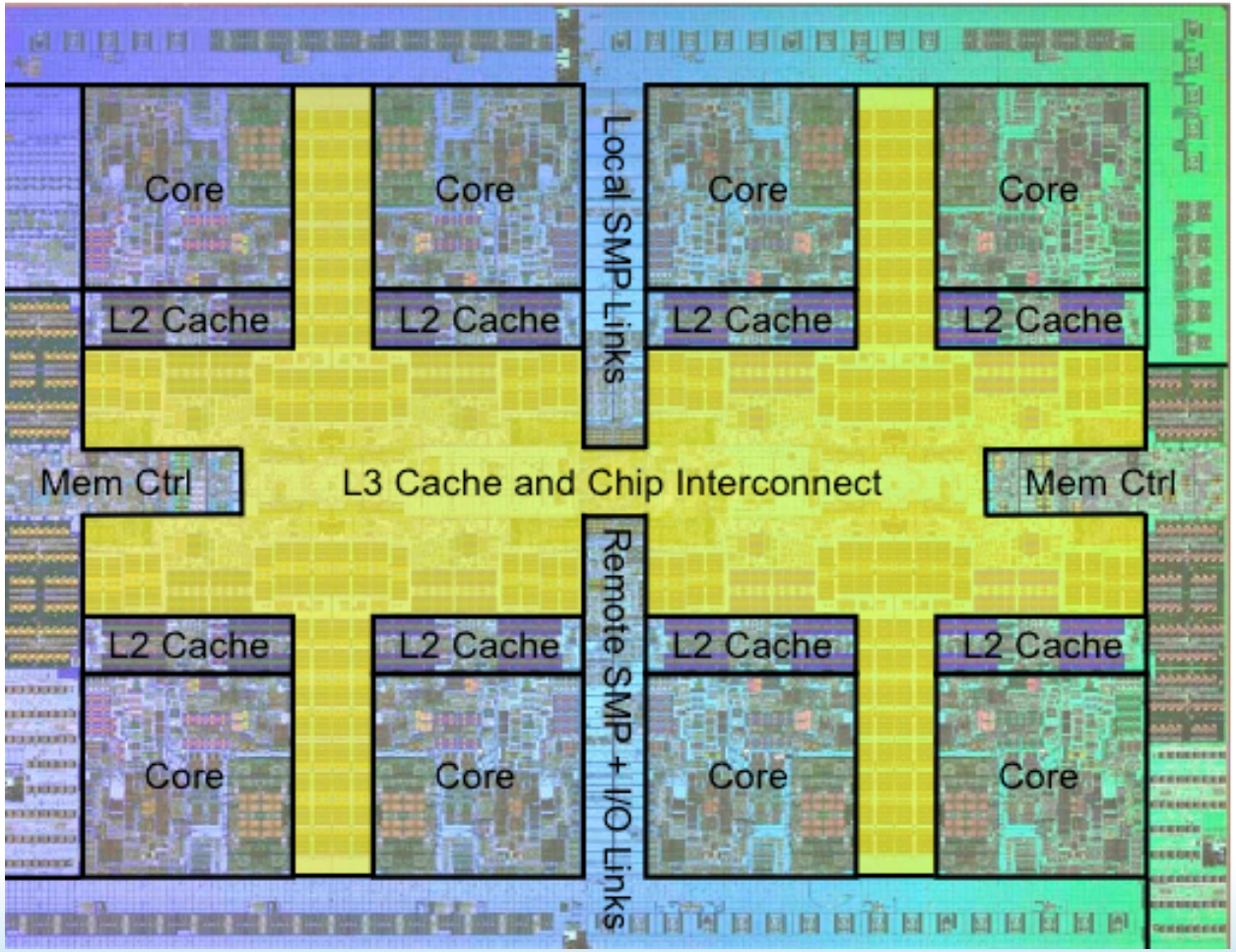
BACKUP

POWER7 Core

- Execution Units
  - 2 Fixed point units
  - 2 Load store units
  - 4 Double precision floating point
  - 1 Branch
  - 1 Condition register
  - 1 Vector unit
  - 1 Decimal floating point unit
  - 6 wide dispatch
- Recovery Function Distributed
- 1,2,4 Way SMT Support
- Out of Order Execution
- 32KB I-Cache
- 32KB D-Cache
- 256KB L2
  - Tightly coupled to core



One Processor Chip – 8 Cores



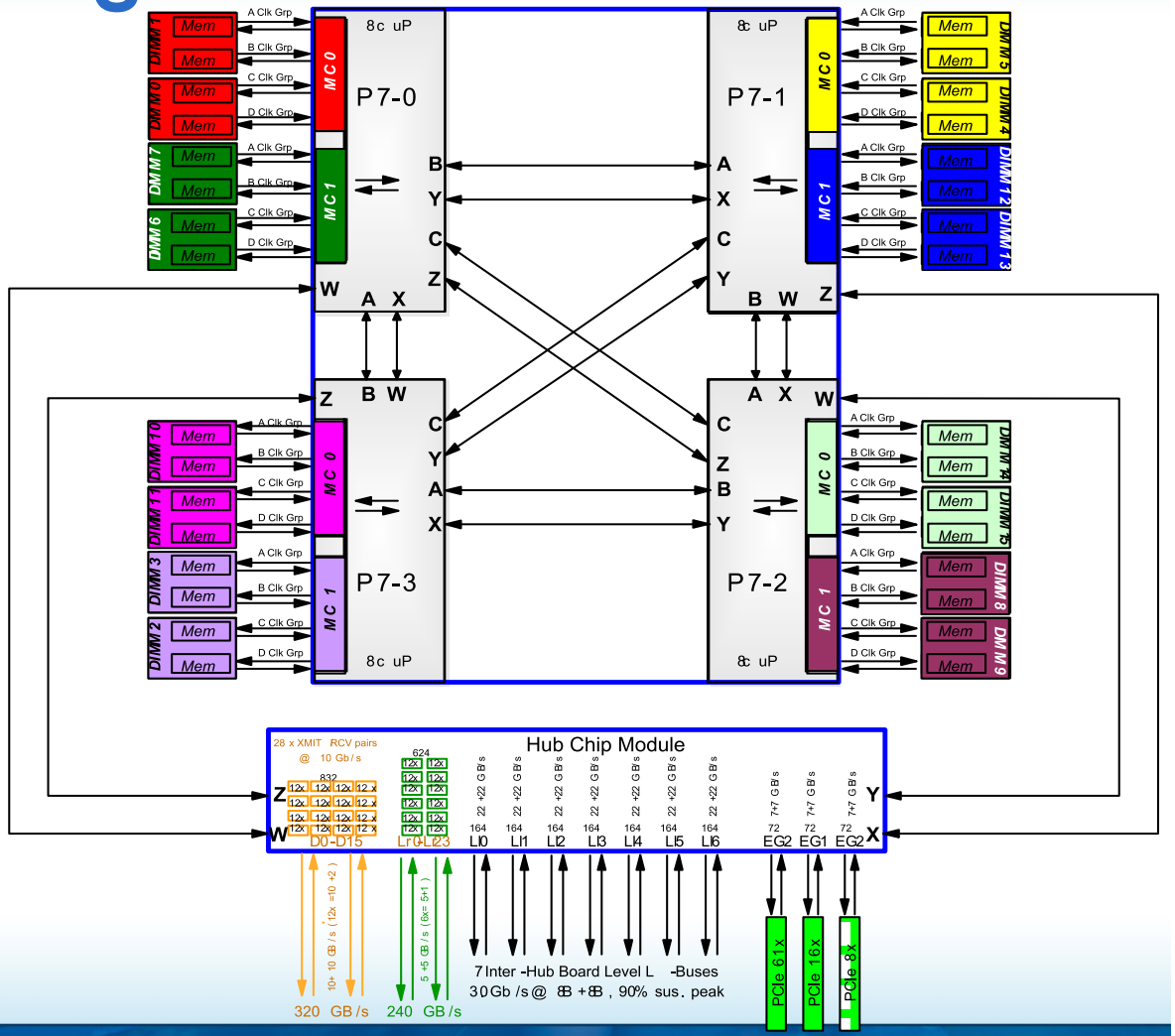
Caches

- Low latency L1 (32KB) and L2 (256KB) dedicated caches per core
  - ~45x lower latency than memory
- 32MB shared L3 cache
  - ~3x lower latency than memory
  - Automatically migrates per-core private working set footprints (up to 4MB) to fast local region per core at ~15x lower latency than memory
  - Automatically clones shared data to multiple per core private regions
  - Enables subset of cores to utilize entire L3 when remaining cores are not using it

Cache Level	Capacity	Type	Policy	Comment
L1 Data	32 KB	Fast SRAM	Store-thru	Local thread storage update
Private L2	256KB	Fast SRAM	Store-In	Coherency maintained throughout system
Fast L3 "Private"	Up to 4 MB	eDRAM	Partial Victim	Reduced latency & power consumption
Shared L3	32MB	eDRAM	Adaptive	Coherency maintained throughout system

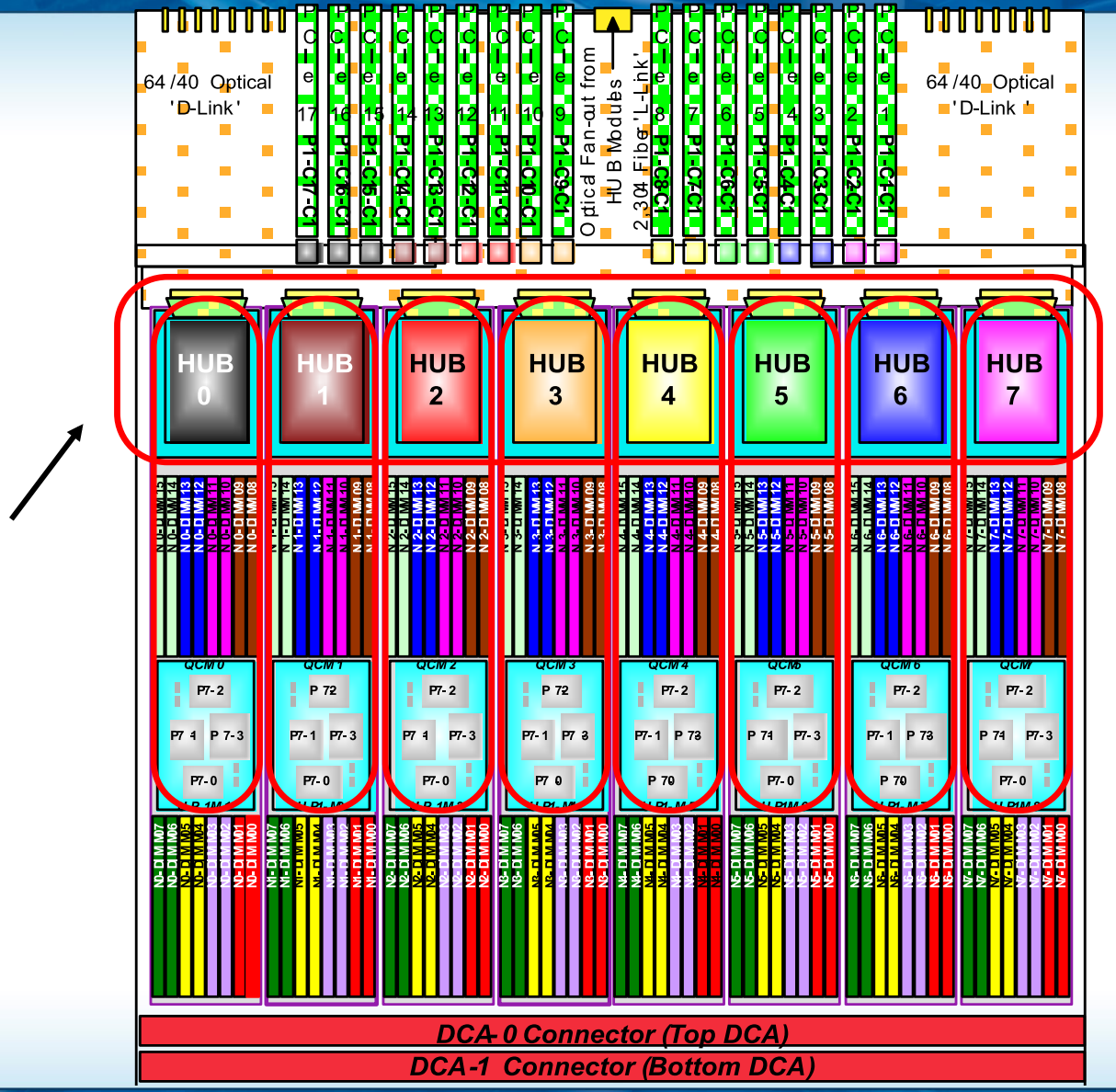
Node (Octant) - Logical

- P7-1H Quad Block Diagram
- 32 w SMP / 2 Tier SMP Fabric
- 4 chip Processor MCM
- Hub SCM
- On-Board Optics



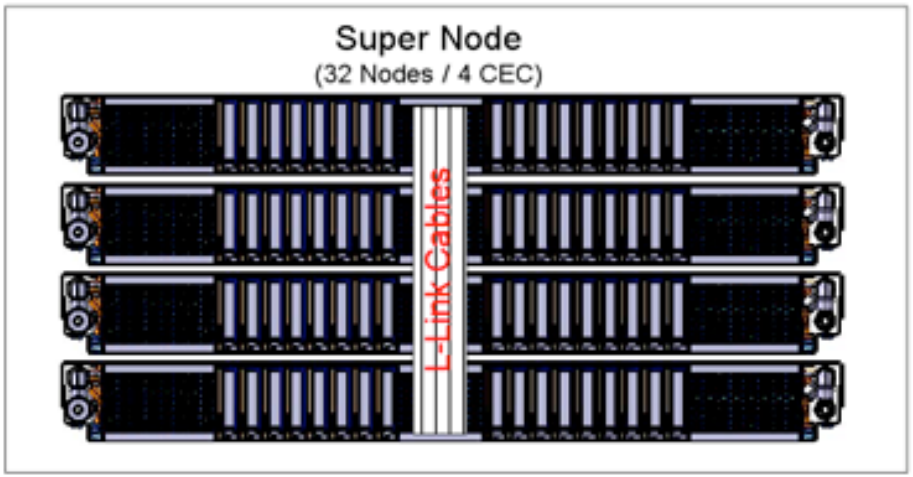
First Level Interconnect

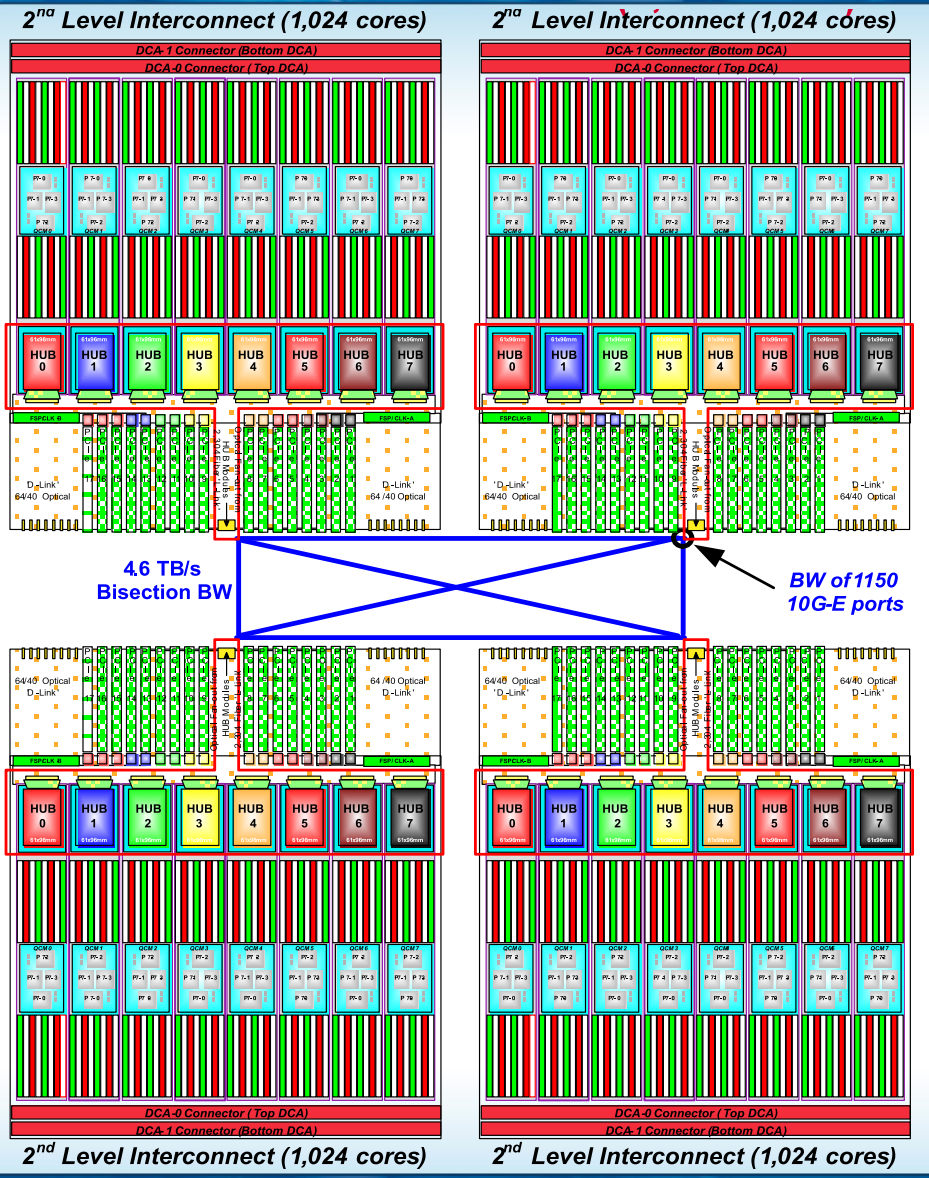
- L-Local
- HUB to HUB Copper Wiring
- 256 Cores



Second Level Interconnect

- Optical 'L-Remote' Links from HUB
- Construct Super Node (4 CECs)
- 1,024 Cores
- Super Node



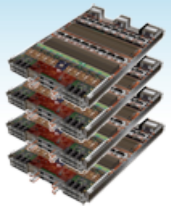


I/O HUB AND INTERCONNECTION NETWORK

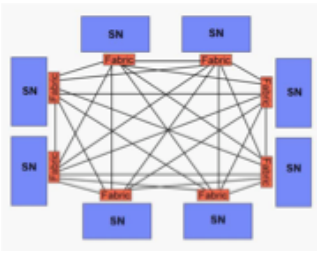
Hub Chip Module (Torrent)

- Connects QCM to PCI-e (two 16x and one 8x PCI-e slot)
- Connects 8 QCM's Together via low latency, high bandwidth, copper fabric.
  - Enables a single hypervisor to run across 8 QCM's
  - Allows I/O slots attached to the 8 hubs to be directed to the compute power of any of the 8 QCM's
  - Provides a message passing mechanism with very high bandwidth
  - Provides the lowest possible latency between 8 QCM's (7.6TF) of compute power
- Connects four P7-IH planers Together via the L Remote Optical connections (Super Node)
- Connects up to 512 Super Nodes Together via the D Optical Buses

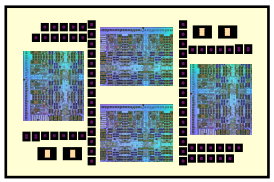
POWER7 Hub Chip Overview

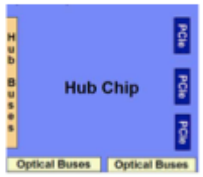


- Replaces external switching and routing functions in prior networks

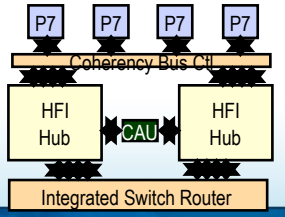


- Highly Integrated Design
  - Integrated Switch/Router (ISR)
  - Integrated Host Channel Adapter (HFI)
  - Integrated Memory Management Unit (NMMU)
  - Integrated PCIe
  - Distributed function across the POWER7 and Hub chipset
  - On-module optical interconnect
  - Enables maximum packaging density



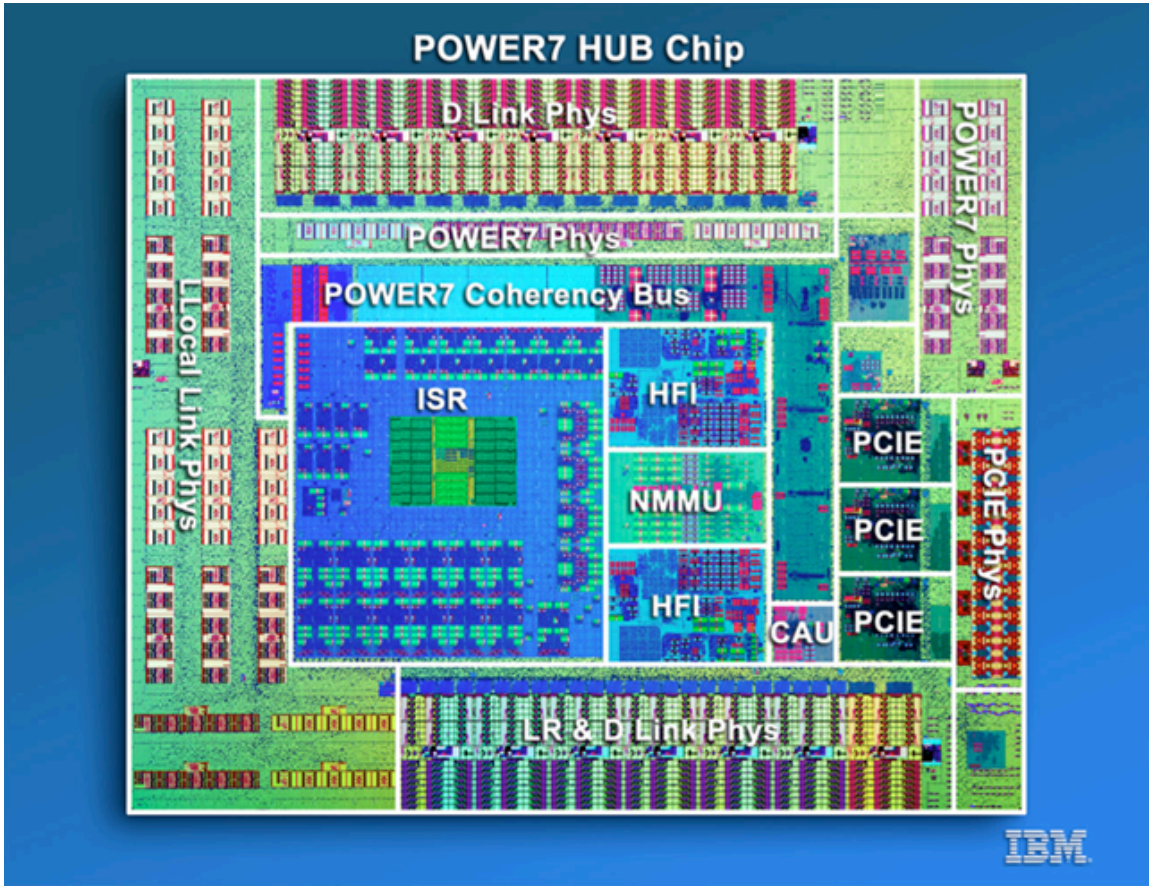


- Hardware Acceleration of key functions
  - Collective Acceleration
  - Global Shared Memory
    - No CPU overhead for remote atomic updates
  - Virtual RDMA
    - No CPU overhead for address translation

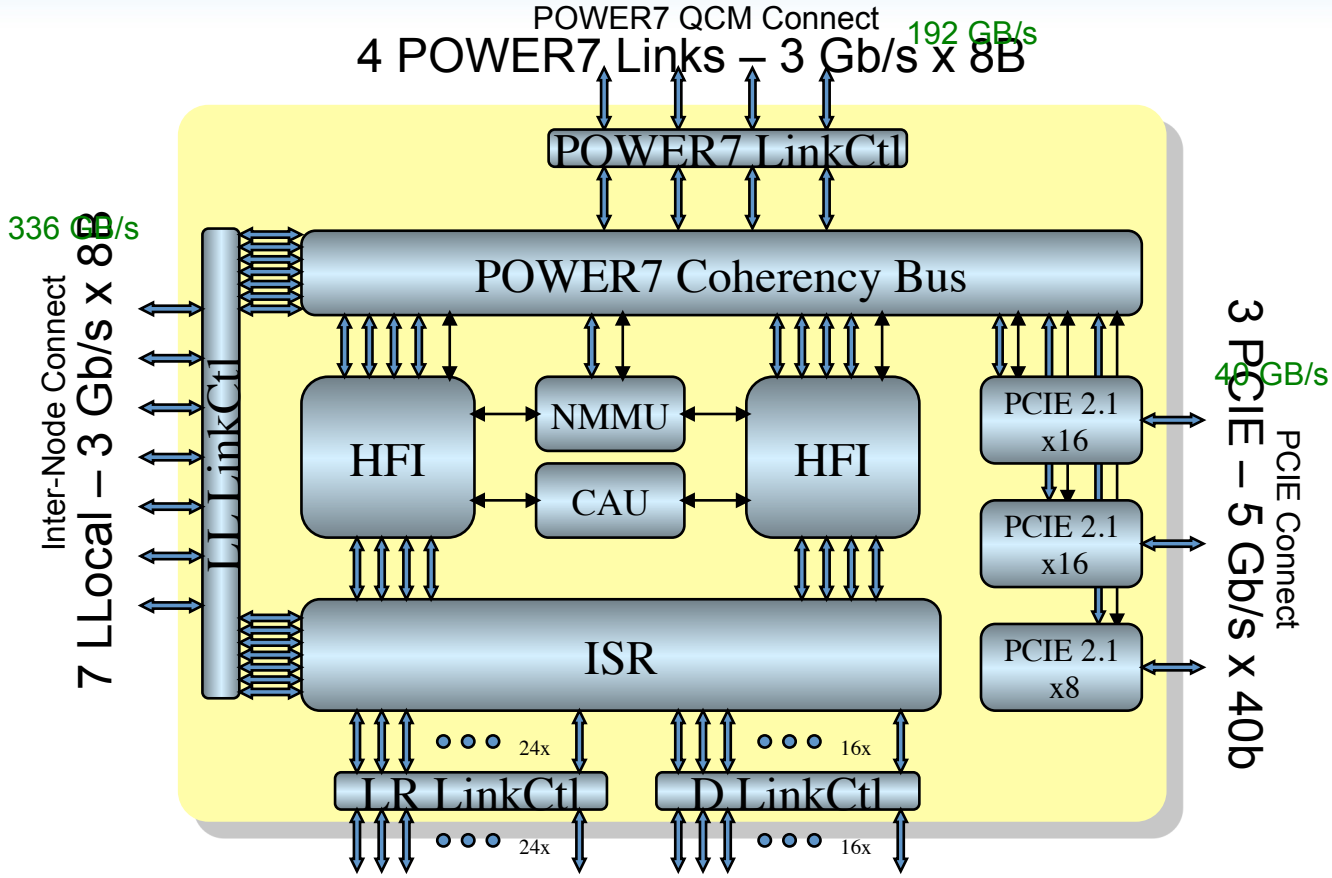


POWER7 Hub Chip

- 45 nm lithography, Cu, SOI
  - 13 levels metal
  - 440M transistors
- 582 mm^2
  - 26.7 mm x 21.8 mm
  - 3707 signal I/O
  - 11,328 total I/O
- 61 mm x 96 mm Glass Ceramic LGA module
  - 56 – 12X optical modules
    - LGA attach onto substrate
- 1.128 TB/s interconnect bandwidth



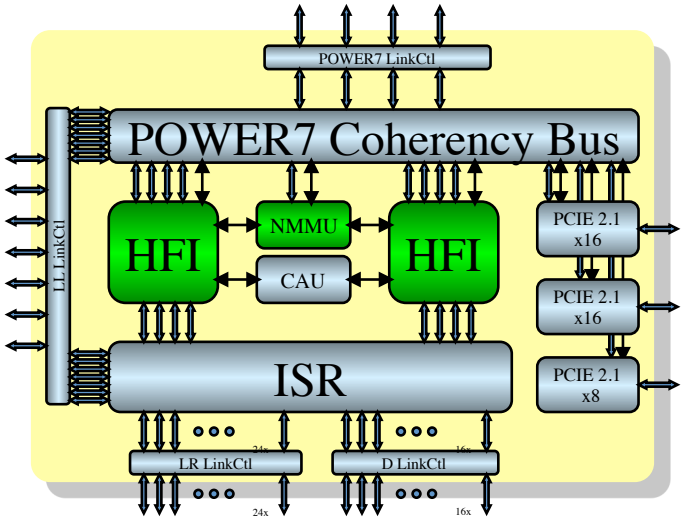
POWER7 Hub Chip Block Diagram



1.128 TB/s of off-chip interconnect bandwidth

Host Fabric Interface (HFI) Features

- Communication controlled through “windows”
  - Multiple supported per HFI
- Address Translation services provided by NMMU
  - HFI provides EA, LPID, Key, Protection Domain
  - Multiple page sizes supported
- Cache-based sourcing to HFI, injection from HFI
  - HFI can extract produced data directly from processor cache
  - HFI can inject incoming data directly into processor L3 cache



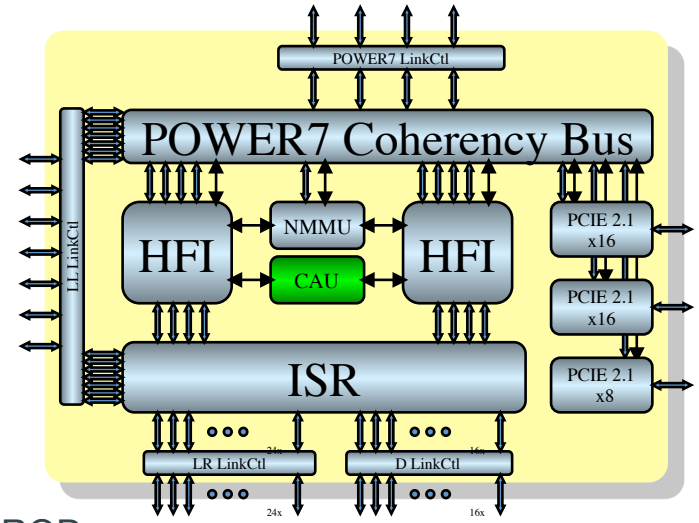
Host Fabric Interface (HFI) Features (cont'd)

- Supports three APIs
  - Message Passing Interface (MPI)
  - Global Shared Memory (GSM)
    - Support for active messaging in HFI (and POWER7 Memory Controller)
  - Internet Protocol (IP)
- Supports five primary packet formats
  - Immediate Send
    - ICSWX instruction for low latency
  - FIFO Send/Receive
  - IP
    - IP to/from FIFO
    - IP with Scatter/Gather Descriptors
  - RDMA
    - Hardware and software reliability modes
  - Collective: Reduce, Multi-cast, Acknowledge, Retransmit

Host Fabric Interface (HFI) Features (cont'd)

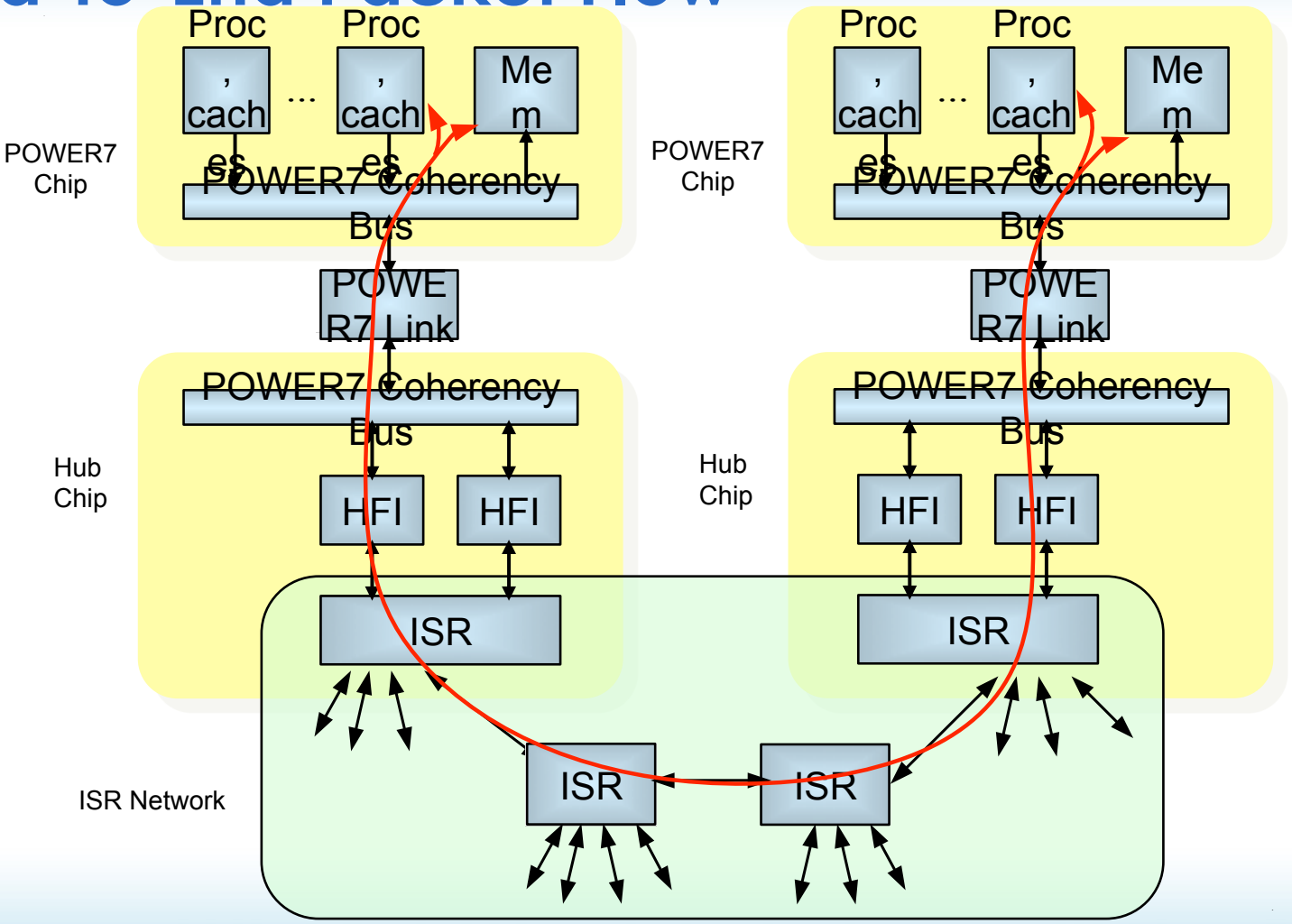
- RDMA Packet Formats
  - Full RDMA (memory to memory)
    - Write, Read, Fence, Completion
    - Large message sizes with multiple packets per message
  - Half-RDMA (memory to/from FIFO)
    - Write, Read, Completion
    - Single packet per message
  - Small-RDMA (FIFO to memory)
    - Remote atomic updates
    - ADD, AND, OR, XOR, and Cmp & Swap with and without Fetch
  - GUPS-RDMA (FIFO to memory)
    - Multiple independent remote atomic updates
    - ADD, AND, OR, XOR
    - Hardware guaranteed reliability mode

Collectives Acceleration Unit (CAU) Features



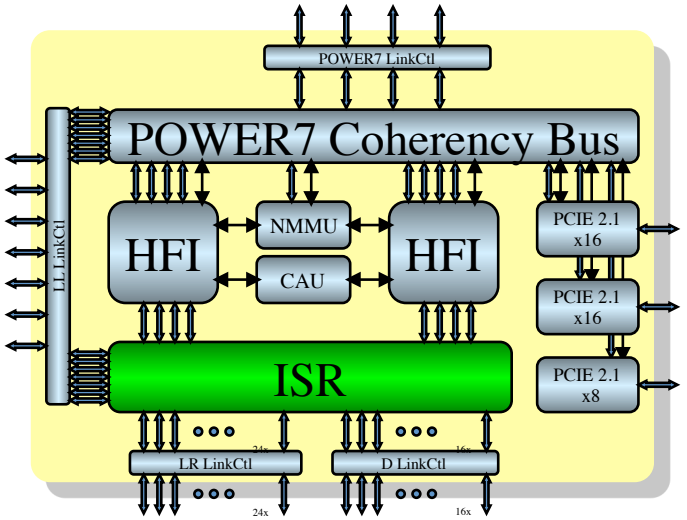
- Operations
  - Reduce: NOP, SUM, MIN, MAX, OR, AND, XOR
  - Multicast
- Operand Sizes and Formats
  - Single Precision and Double Precision
  - Signed and Unsigned
  - Fixed Point and Floating Point
- Extended Coverage with Software Aid
  - Types: barrier, all-reduce
  - Reduce ops: MIN_LOC, MAX_LOC, (floating point) PROD
- Tree Topology
  - Multiple entry CAM per CAU: supports multiple independent trees
  - Multiple neighbors per CAU: each neighbor can be either a local or remote CAU /HFI
  - Reliability/Pipelining using Sequence Numbers and Retransmission protocol
  - Each tree has one and only one participating HFI window on any involved node
  - It's up to the software to setup the topology

End-to-End Packet Flow



Integrated Switch Router (ISR) Features

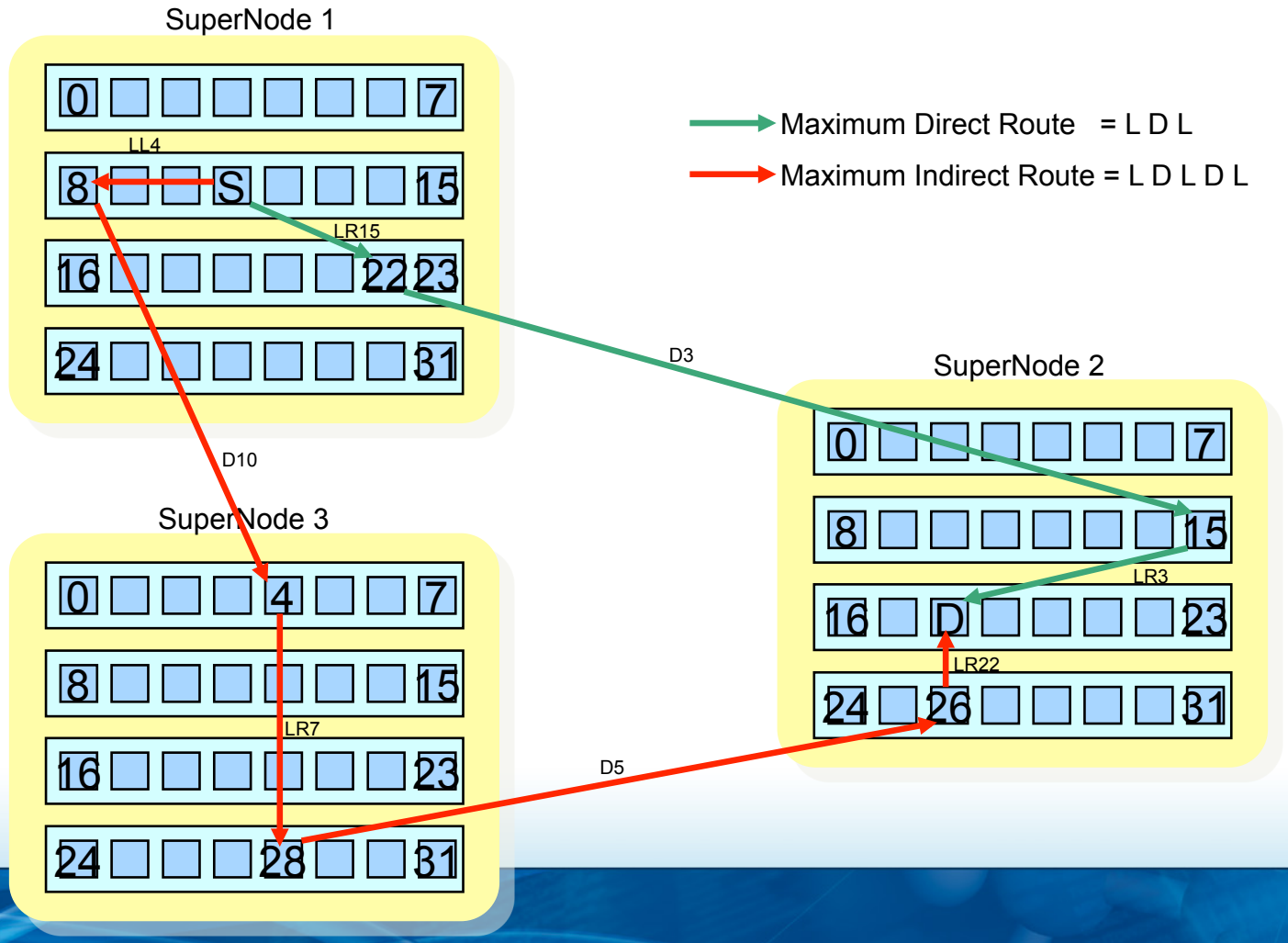
- Two tier, full graph network
- 3.0 GHz internal 56x56 crossbar switch
  - 8 HFI, 7 LL, 24 LR, 16 D, and SRV ports
- Virtual channels for deadlock prevention
- Input/Output Buffering
- 2 KB maximum packet size
  - 128B FLIT size
- IP Multicast Support
  - Multicast route tables per ISR for replicating and forwarding multicast packets
- Global Counter Support
  - HW synchronization with Network Management setup and maintenance



Integrated Switch Router (ISR) Features

- Routing
  - L-D-L longest direct route
  - L-D-L-D-L longest indirect route
  - Cut-through Wormhole routing
  - Full hardware routing using distributed route tables
    - Source route tables for packets injected by the HFI
    - Port route tables for packets at each hop in the network
    - Separate tables for inter-supernode and intra-supernode routes
  - HW Direct Routing
    - Multiple direct routes for less than full-up system
  - HW Indirect Routing for data striping and failover
    - Round-Robin, Random
  - Software controlled indirect routing through hardware route tables
  - FLITs of a packet arrive in order, packets of a message can arrive out of order

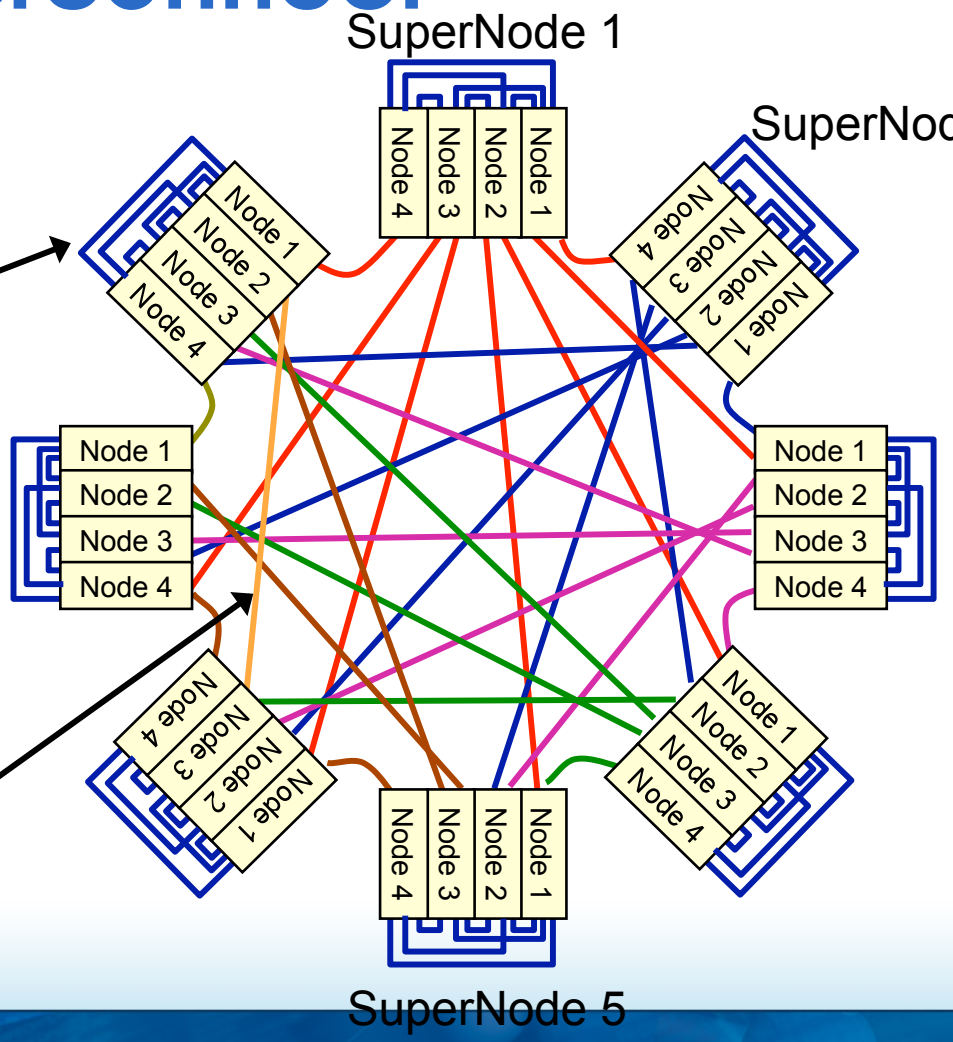
Sample Direct and Indirect Routes through the Interconnect



Two Level Interconnect

1st level: L-Links, 24 & 6 GB/s
- Connect 4 drawers together to form a SuperNode
- Copper &Optical Cable

2nd level: D-Links, 10 GB/s
- Optical Cable
- Connects SuperNodes to all other SuperNodes
- Up to 512 SuperNodes fully connected



POWER7 Hub Chip

- 45 nm lithography, Cu, SOI
  - 13 levels metal
  - 440M transistors
- 582 mm^2
  - 26.7 mm x 21.8 mm
  - 3707 signal I/O
  - 11,328 total I/O
- 61 mm x 96 mm Glass Ceramic LGA module
  - 56 – 12X optical modules
    - LGA attach onto substrate
- 1.128 TB/s interconnect bandwidth