BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Power 7 for Scientific Computing

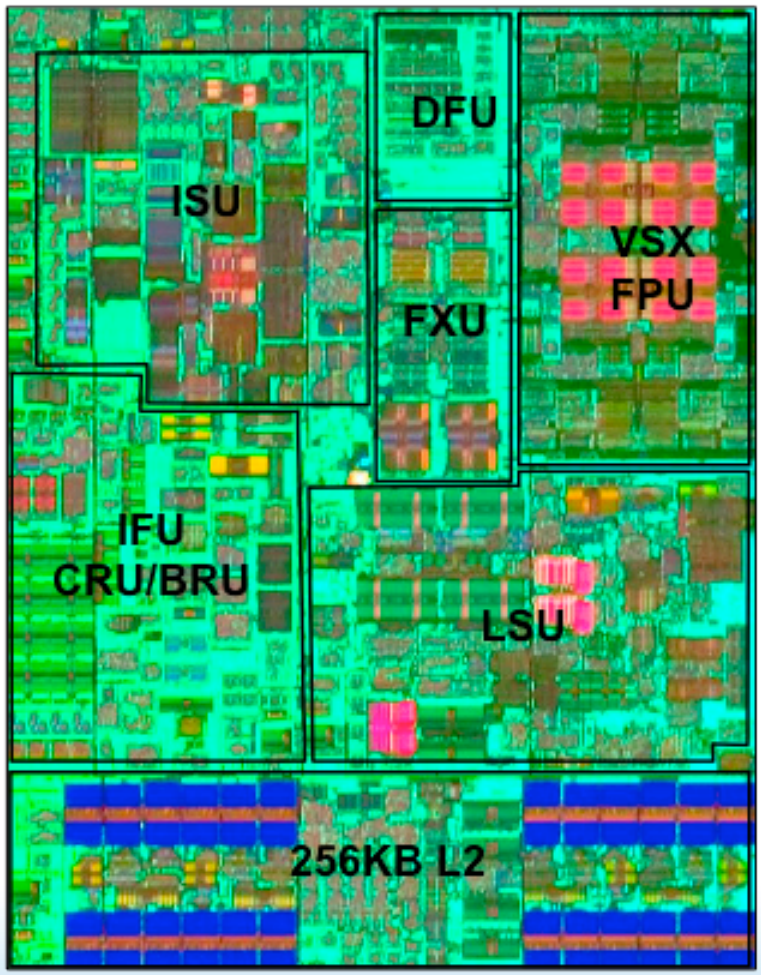
Public Information

NCSA IBM GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

ONE CORE

Power 7 Core

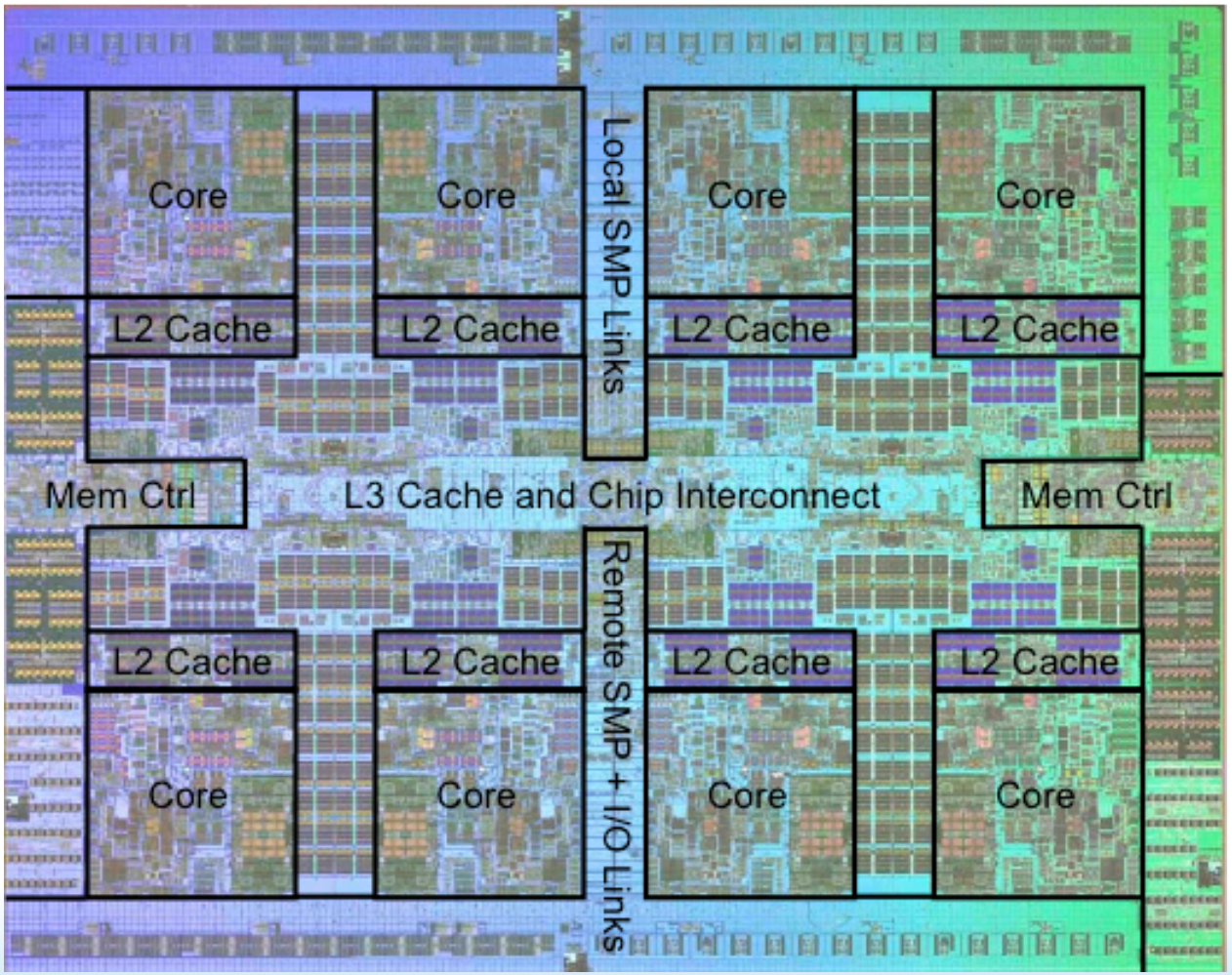
- Execution Units
  - 2 Fixed point units
  - 2 Load store units
  - 4 Double precision floating point
  - 1 Branch
  - 1 Condition register
  - 1 Vector unit
  - 1 Decimal floating point unit
  - 6 wide dispatch
- Recovery Function Distributed
- 1,2,4 Way SMT Support
- Out of Order Execution
- 32KB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

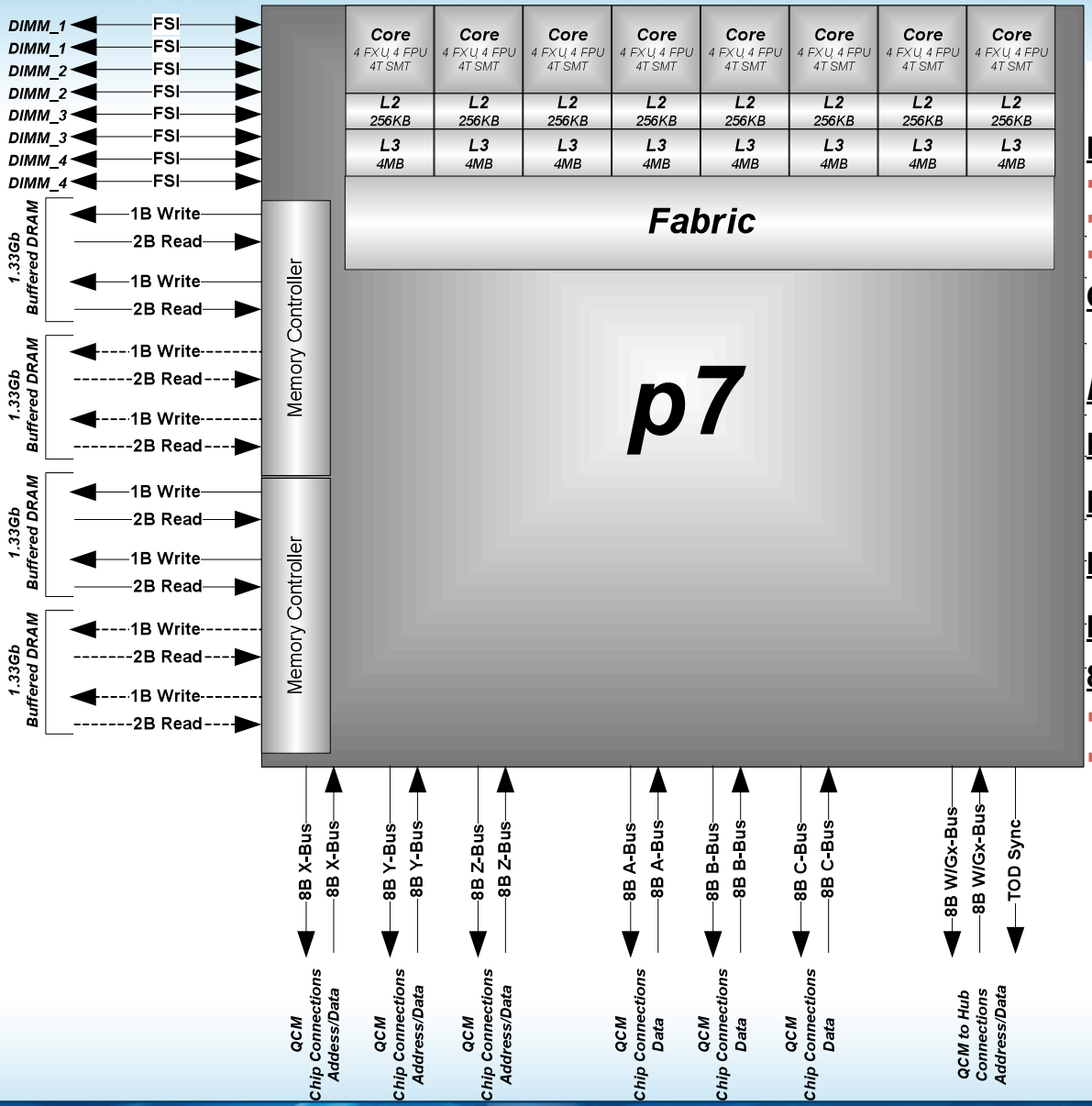


8 cores

ONE CHIP

Power 7 Chip


Core
Core
Core
Core
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Local SMP Links
Mem Ctrl
L3 Cache and Chip Interconnect
Mem Ctrl
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Remote SMP + I/O Links
Core
Core
Core
Core



P7

- 8 Cores per p7
- 4 Floating Point Units (FPU) per core
- 2 FLOPS/Cycle (Fused Operation)

Off-chip bandwidth: 336 GB/s

- (local + remote interconnect)

L1 32KB Instruction, 32KB Data / core

L2 = 256KB / core

L3 = 4MB eDRAM / core

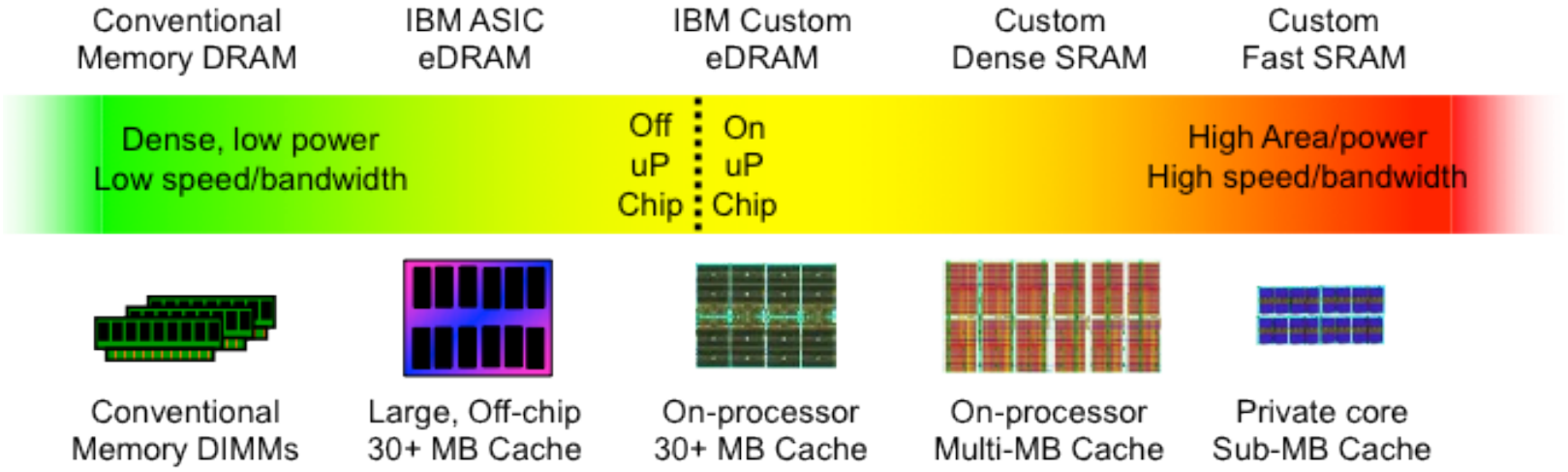
Memory Capacity: Up to 128 GB / p7

Memory bandwidth: 128 GB/s peak / p7

8 Channels of SuperNova buffered DIMMs

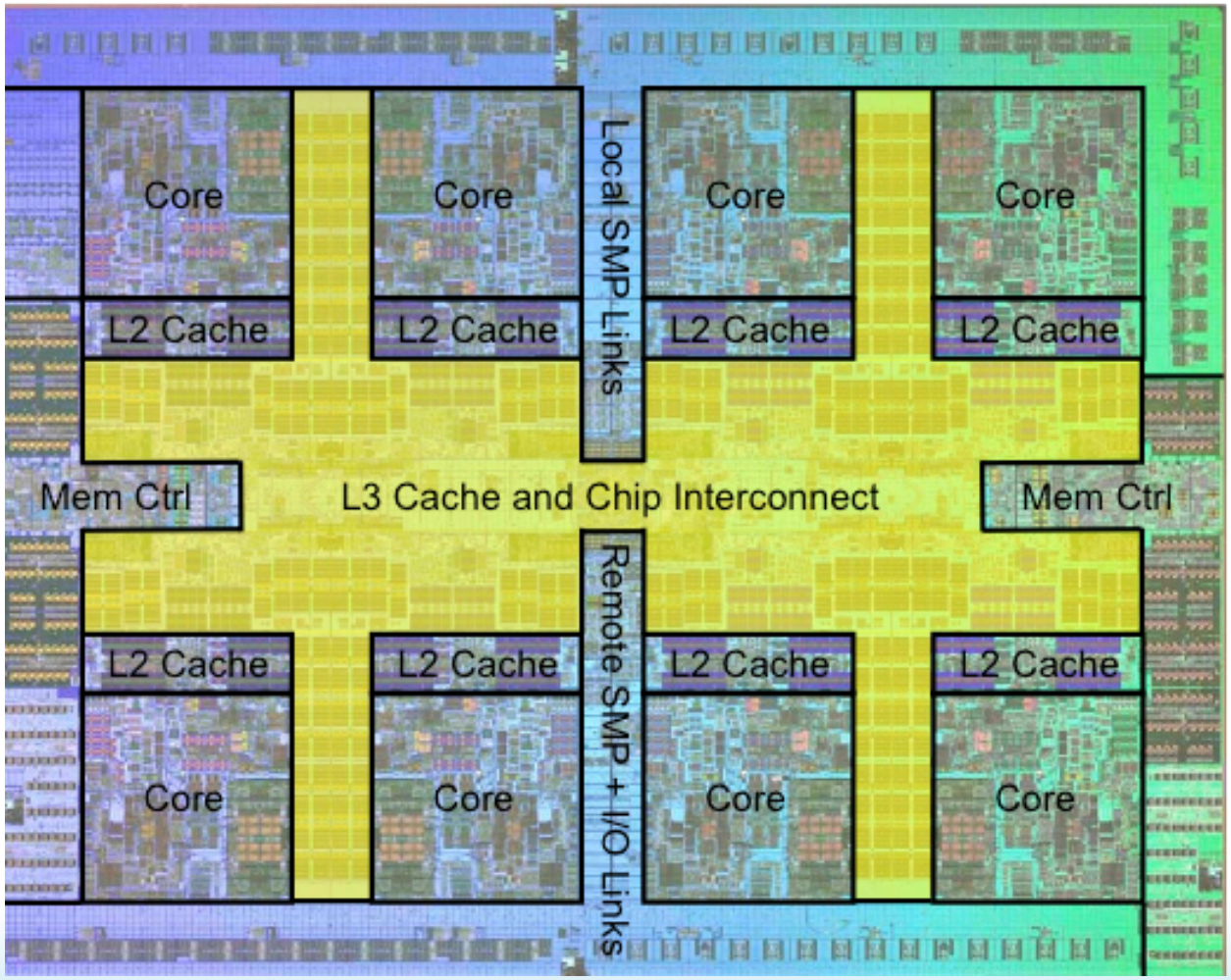
- **2 memory controllers / p7**
- **4 memory busses per memory controller**
  - Each of the 4 busses are:
  - 1B wide Write, 2B wide Read)

RAM Technologies



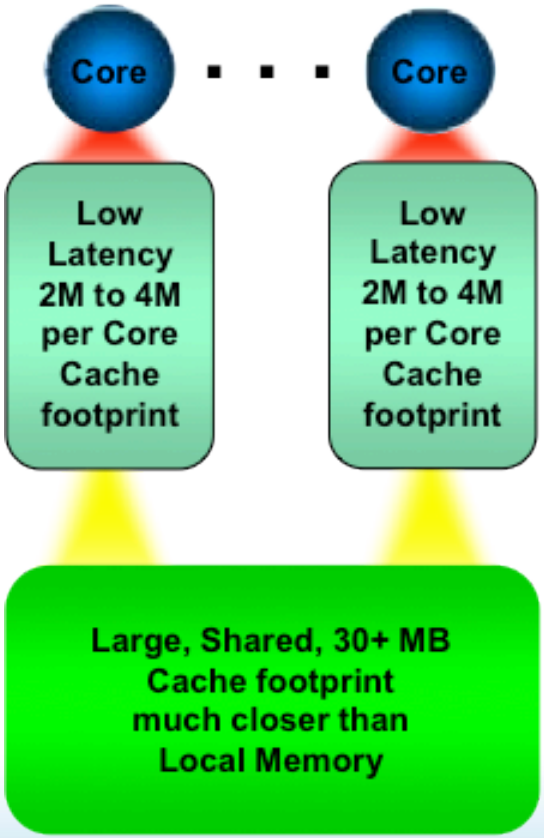
- DIMMs, Dense and Fast SRAM are industry standard; eDRAM is IBM technology
  - Used in Power 4,5,6 for off-chip L3 cache
  - Used in Power 7 for on-chip cache (to avoid pin limitations and support BW of 8 cores)

L3 Cache


Core
Core
Core
Core
Local SMP Links
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Mem Ctrl
L3 Cache and Chip Interconnect
Mem Ctrl
Remote SMP + I/O Links
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Core
Core
Core
Core

L3 Cache Requirements

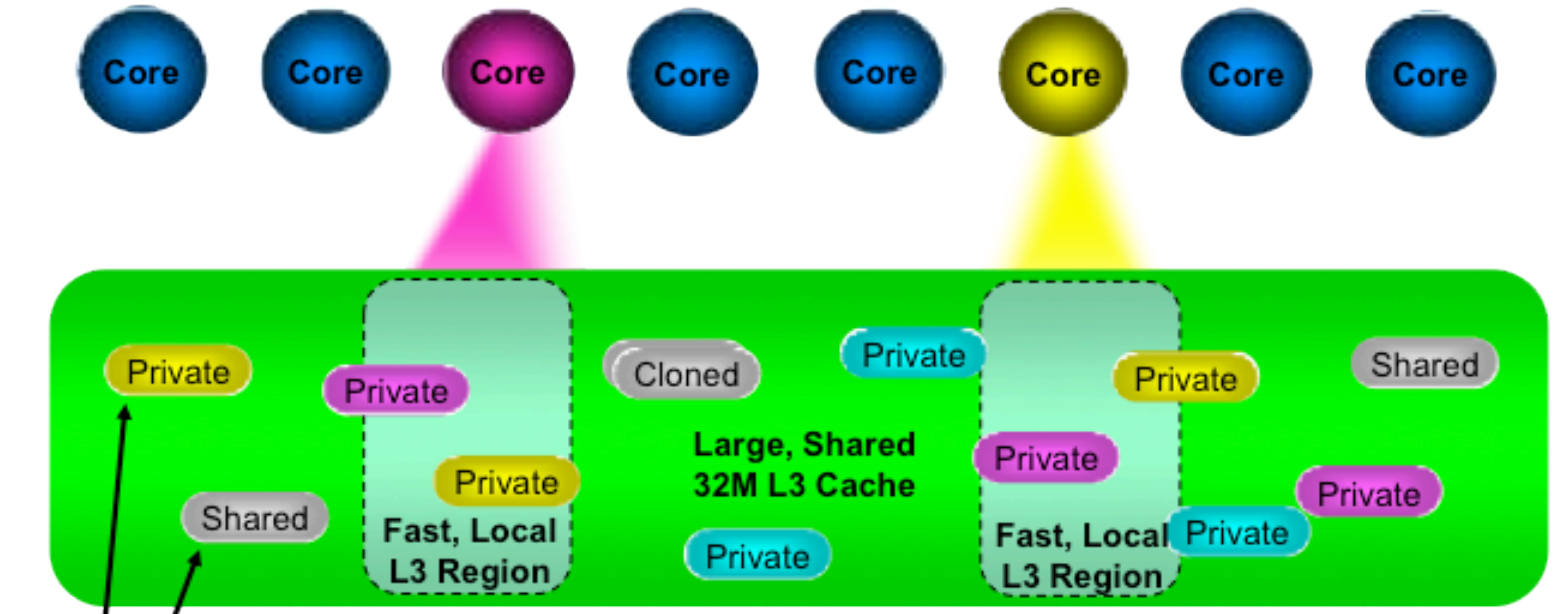
Cache Hierarchy Rqmt for POWER Servers



Challenge for Multi-core POWER7

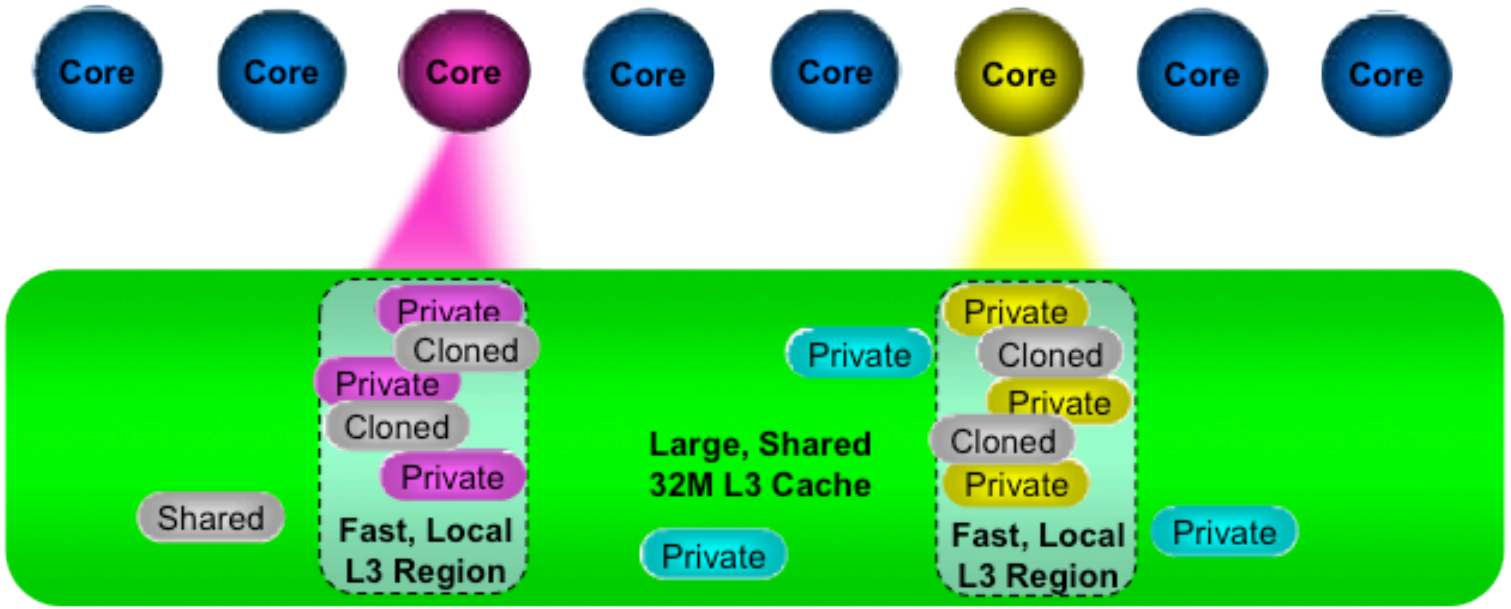
Need to satisfy both caching requirements with one cache.

Solution: "Hybrid" L3 – Fluid Cache Structure



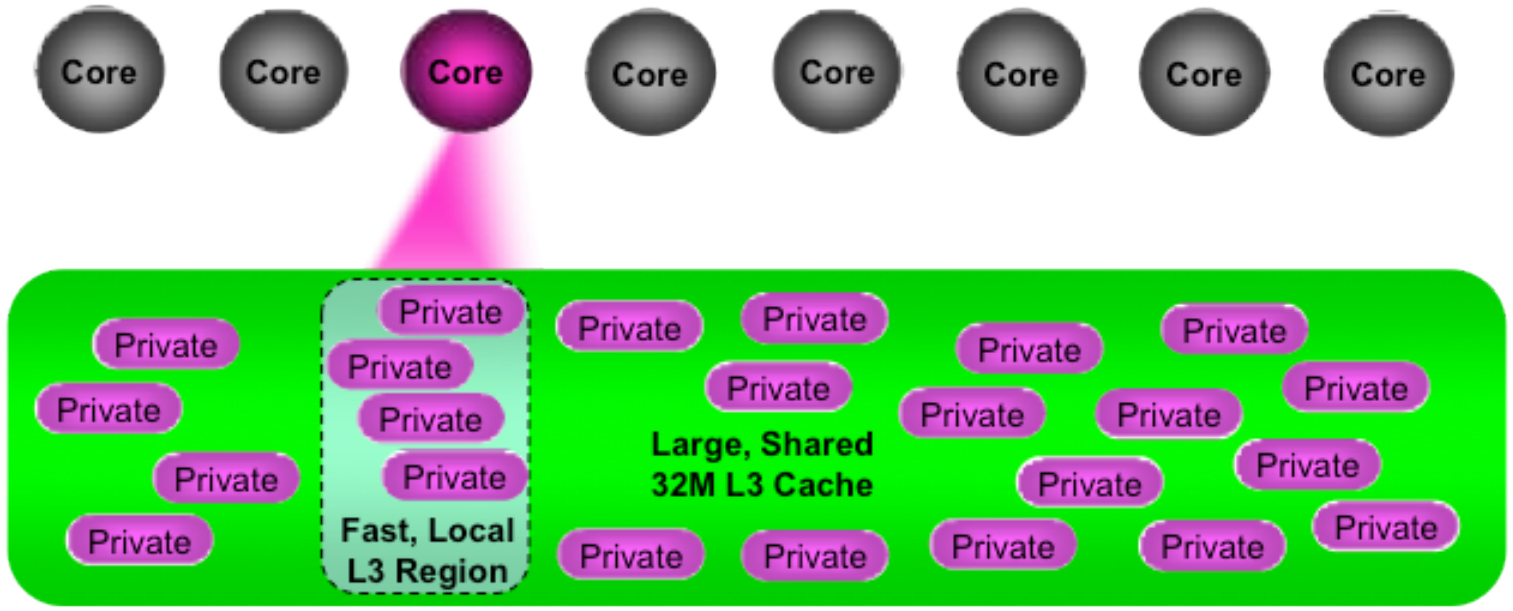
- Keeps multiple footprints at ~3X lower latency than local memory.
- Automatically migrates private footprints (up to 4M) to fast local region (per core) at ~5X lower latency than full L3 cache.
- Automatically clones shared data to multiple private regions.

Solution: "Hybrid" L3 – Fluid Cache Structure



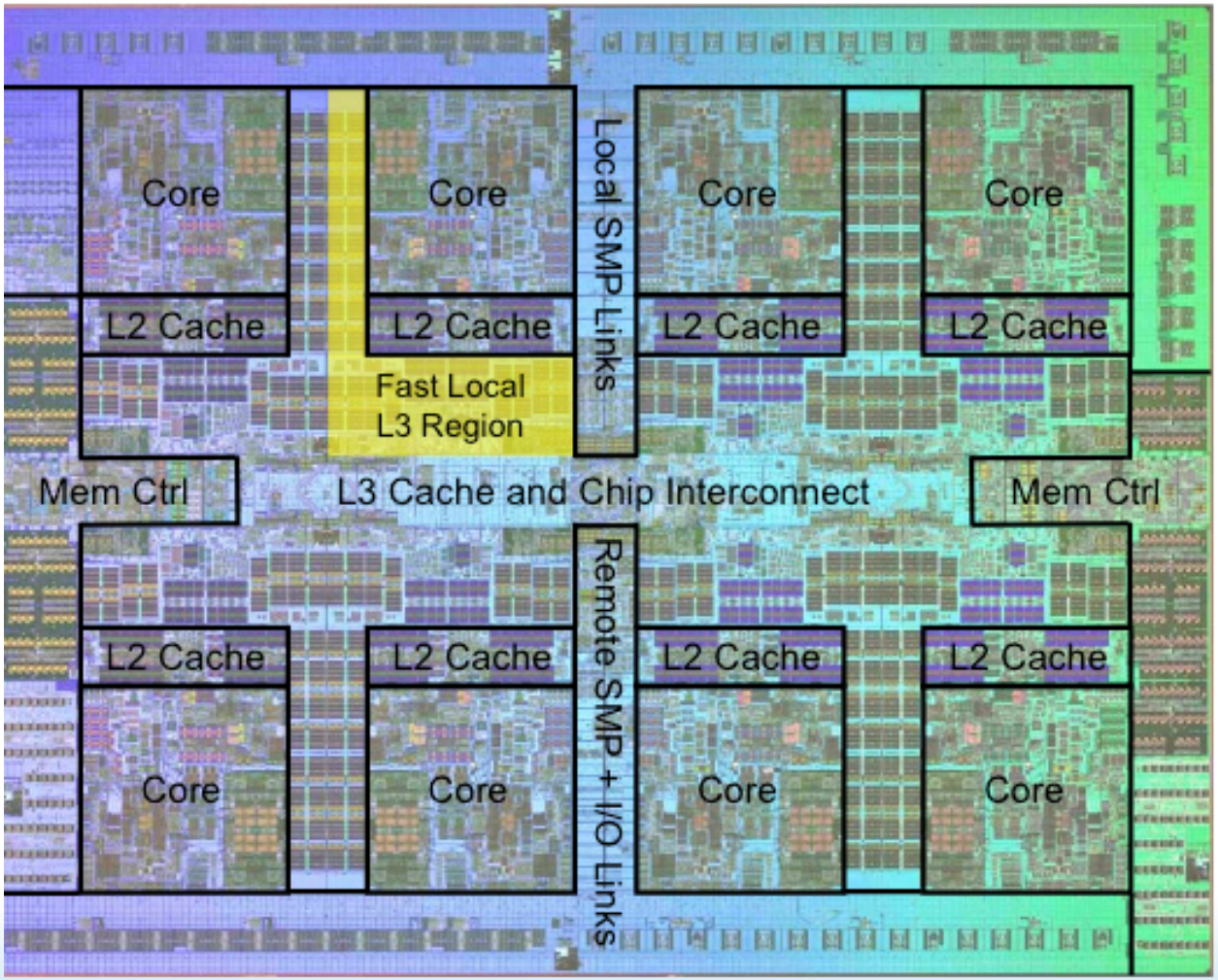
- Keeps multiple footprints at ~3X lower latency than local memory.
- Automatically migrates private footprints (up to 4M) to fast local region (per core) at ~5X lower latency than full L3 cache.
- Automatically clones shared data to multiple private regions.

Solution: "Hybrid" L3 – Fluid Cache Structure

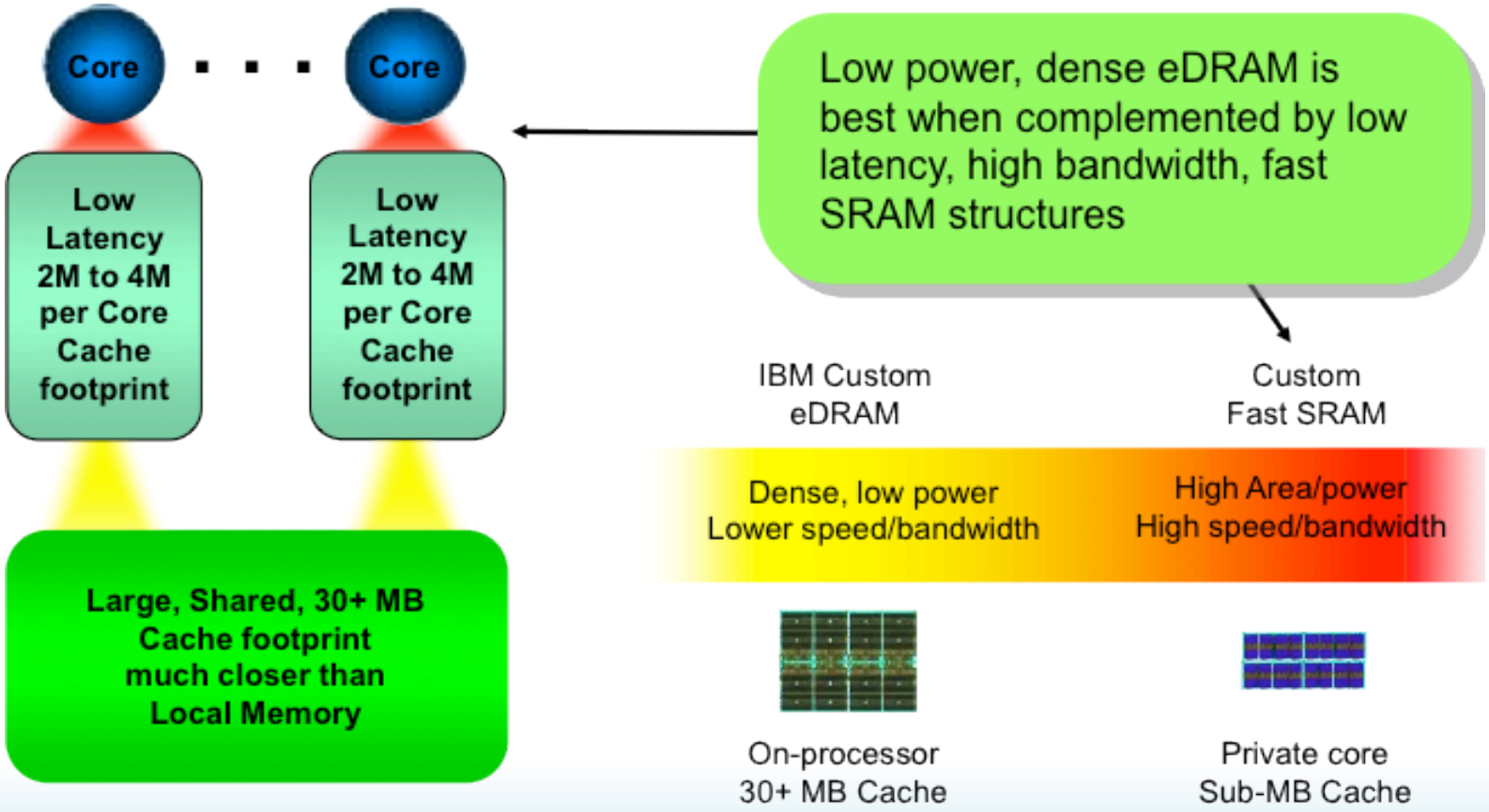


- Enables a subset of the cores to utilize the entire large shared L3 cache when the remaining cores are not using it.

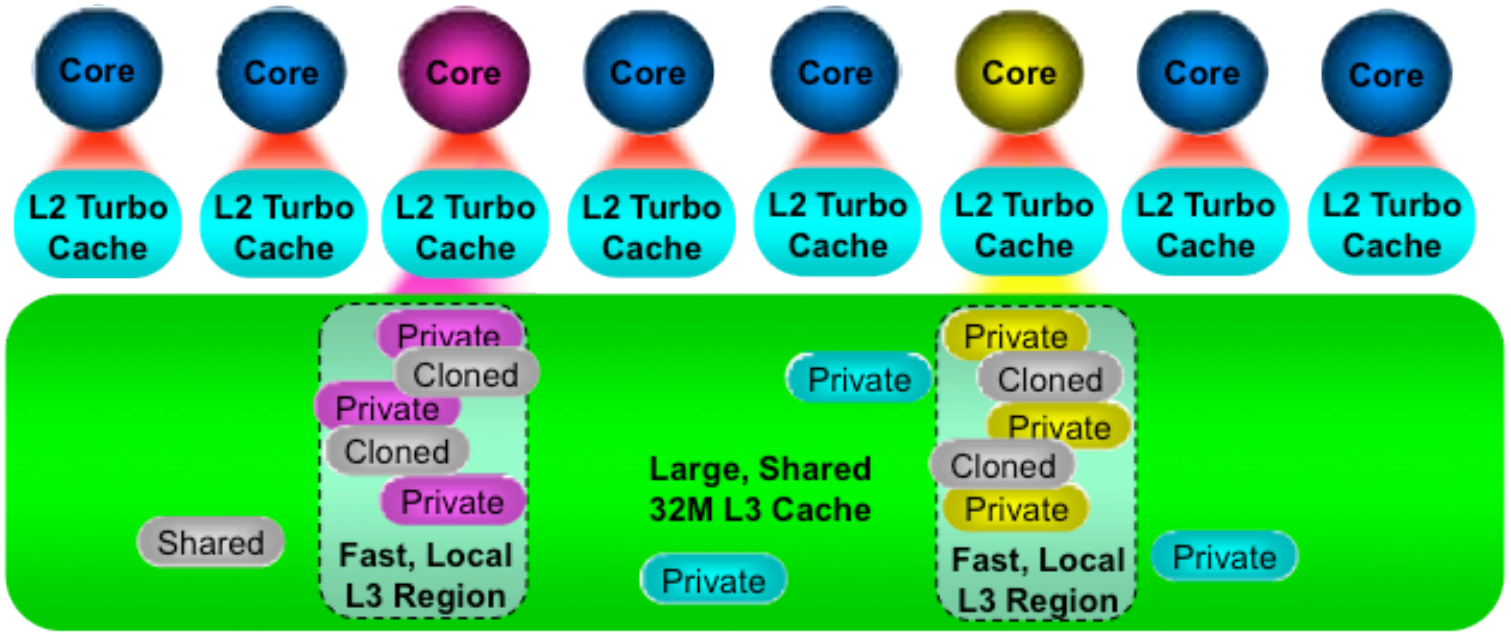
Local L3 Region



L2 Still Needed

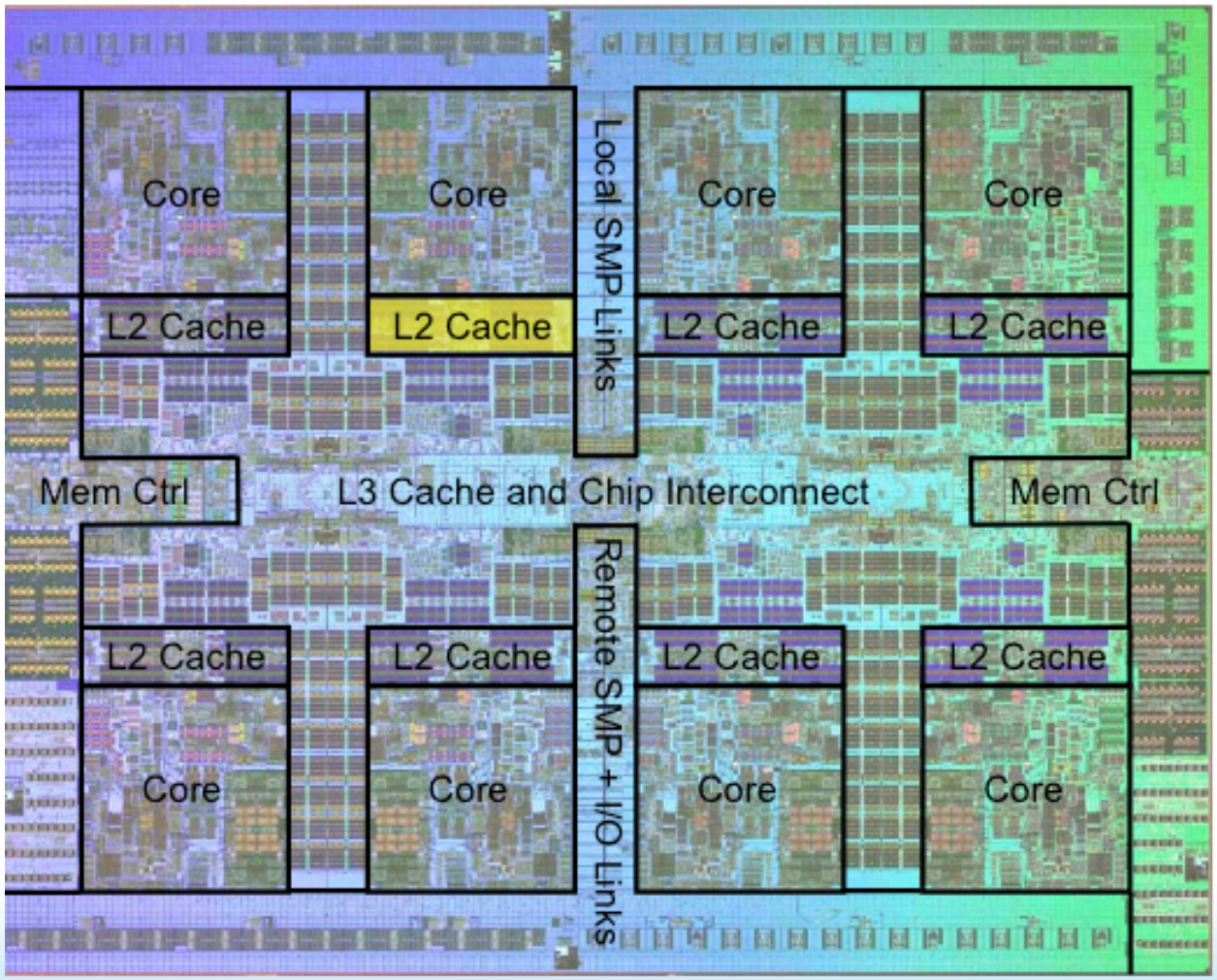


L2 Turbo Cache

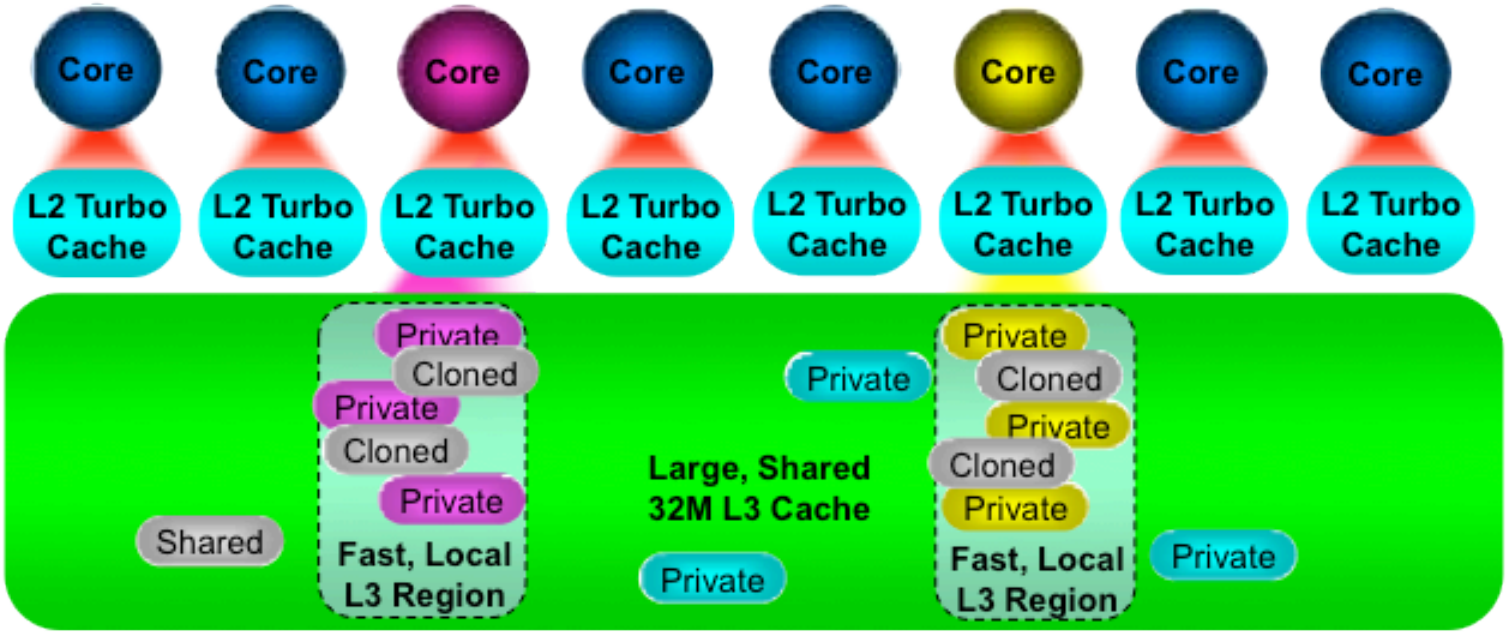


- L2 "Turbo" cache keeps a tight 256K working set with extremely low latency (~3X lower than local L3 region) and high bandwidth, reducing L3 power and boosting performance.

L2 Cache

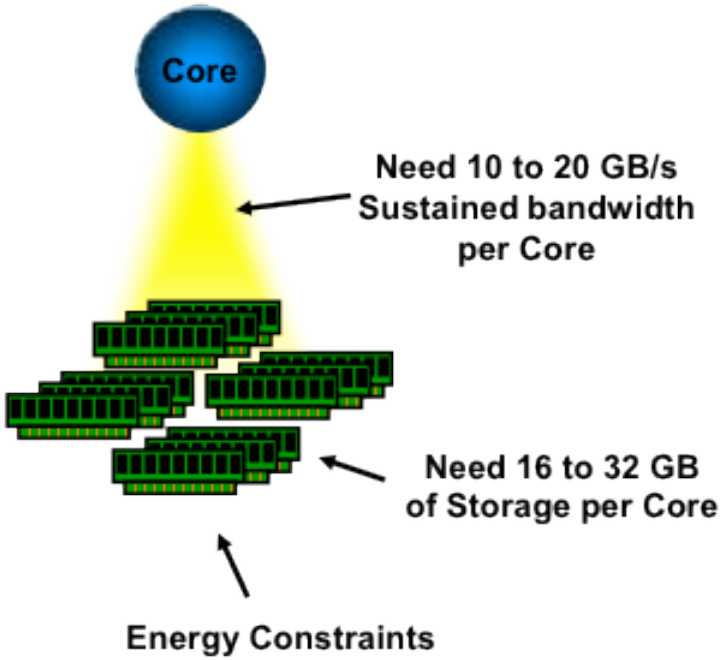

Core
Core
Core
Core
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Local SMP Links
Mem Ctrl
L3 Cache and Chip Interconnect
Mem Ctrl
L2 Cache
L2 Cache
L2 Cache
L2 Cache
Core
Core
Core
Core
Remote SMP + I/O Links

Cache Hierarchy Summary

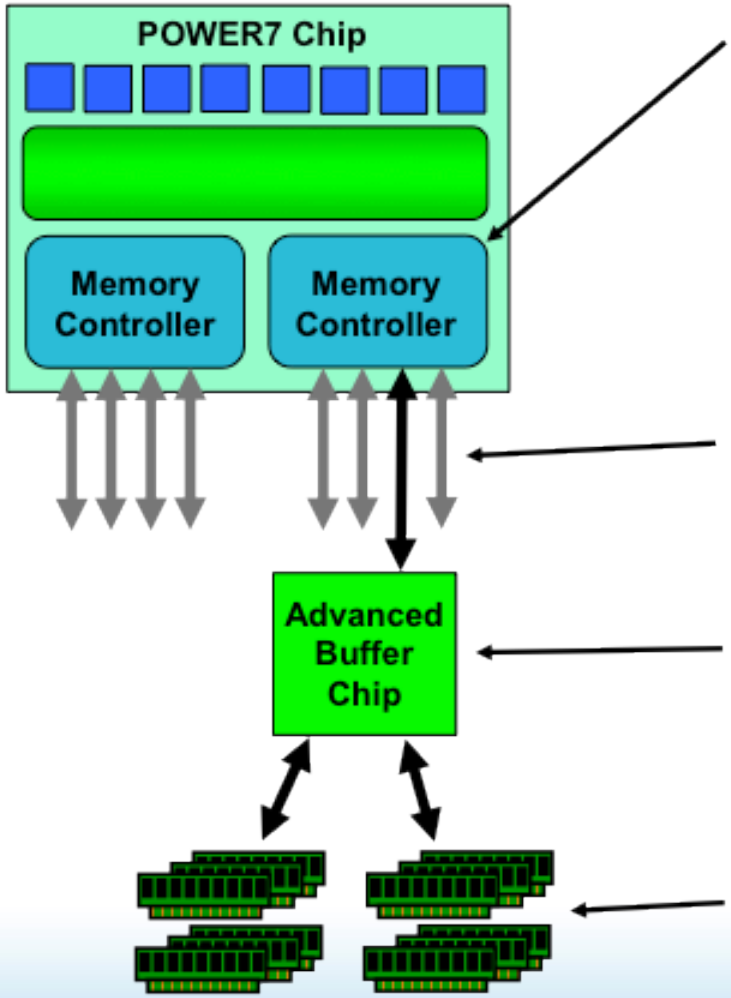


Cache Level	Capacity	Array	Policy	Comment
L1 Data	32K	Fast SRAM	Store-thru	Local thread storage update
Private L2	256K	Fast SRAM	Store-In	De-coupled global storage update
Fast L3 Region	Up to 4M	eDRAM	Partial Victim	Reduced power footprint (up to 4M)
Shared L3	32M	eDRAM	Adaptive	Large 32M shared footprint

Memory Goals



Memory Solutions



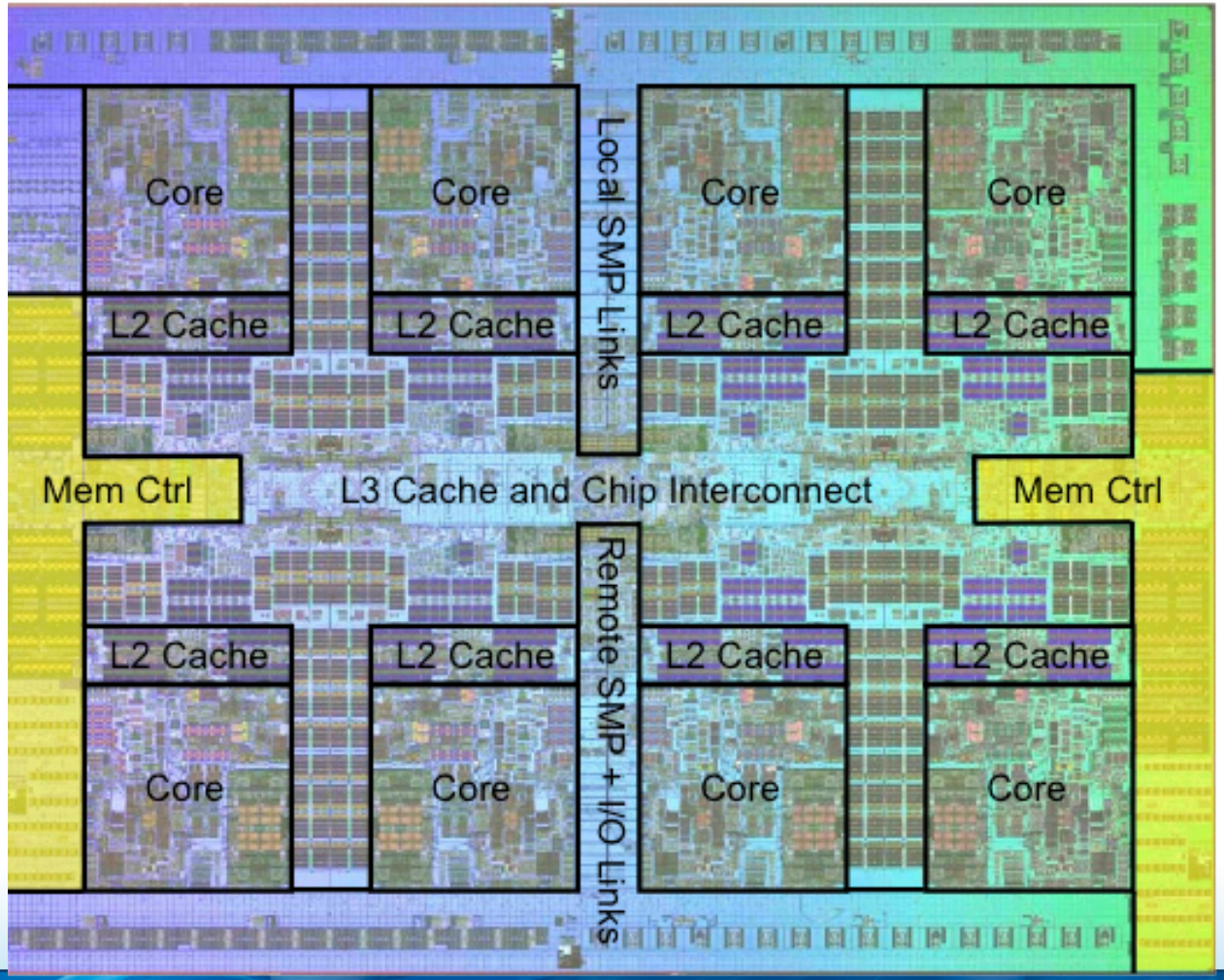
1) Dual Integrated DDR3 Controllers
- Massive 16KB scheduling window per POWER7 chip insures high channel and DIMM utilization
- Sparse access acceleration
- Advanced Energy Management
- Numerous RAS advances

2) Eight high speed 6.4 GHz channels
- New low power differential signaling
- Sustained 100+ GB/s per socket

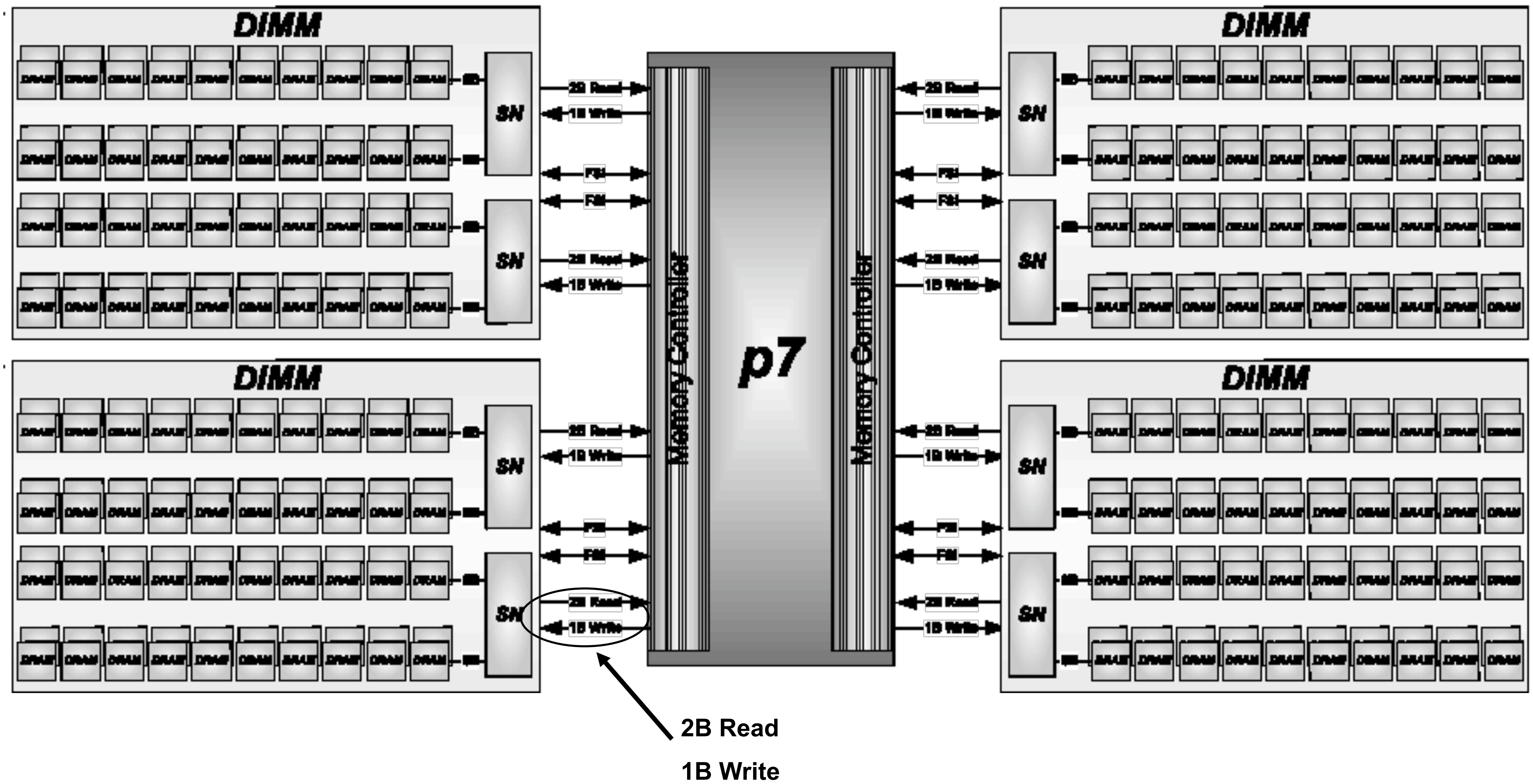
3) New DDR3 buffer chip architecture
- Larger capacity support (32 GB / core)
- Energy Management support
- RAS enablement

4) DDR3 DRAMs
- Supports 800, 1066, 1333, and 1600

On-Chip Memory Controllers



Memory


DIMM
DIMM
DIMM
DIMM
SN
p7
Memory Controller
Memory Controller
2B Read
1B Write

Memory Technology

L1 32KB Instruction, 32KB Data / core

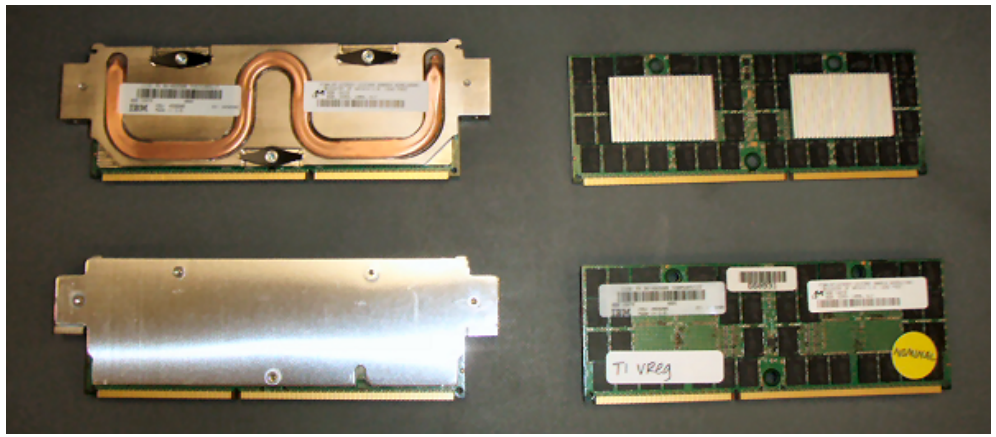
L2 = 256KB / core

L3 = 4MB eDRAM / core

Memory Capacity: Up to 128 GB / p7

8 Channels of SuperNova buffered DIMMs / p7

- **2 memory controllers / p7**
- **4 memory busses per memory controller**
  - Each of the 4 busses are: 1B wide Write, 2B wide Read)



Dual SuperNova DIMM

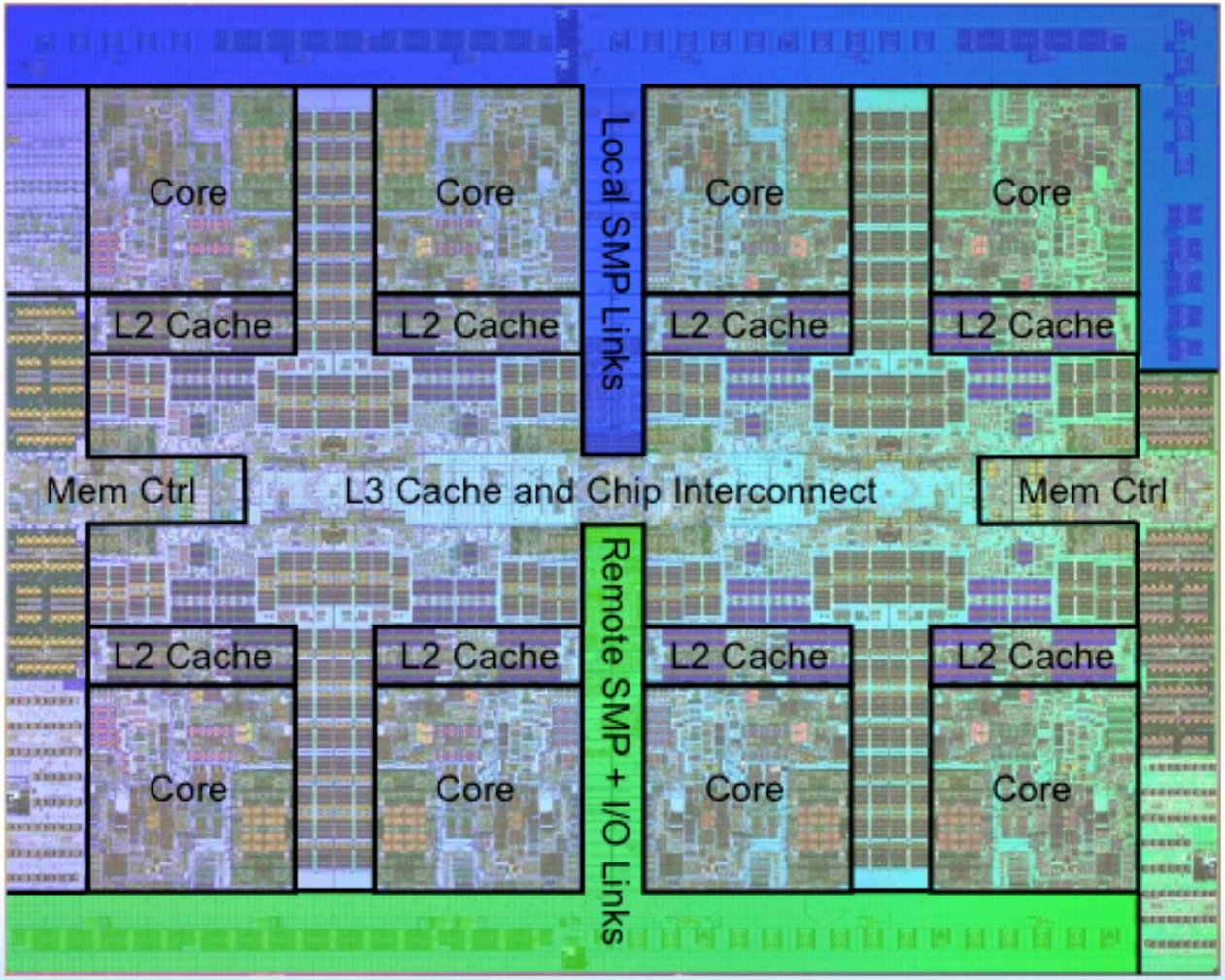
Off-Chip Connectivity

Interface	Signal Type	Info Width	Frequency	Bandwidth
Off-chip Cache	none	none	none	none
Memory Channels	Differential	28 bytes	6.4 Ghz	180 GB/s
I/O Bridge	Single-ended	20 bytes	2.5 Ghz	50 GB/s
SMP Interconnect	Single-ended	120 bytes	3.0 Ghz	360 GB/s
Total Bandwidth				590 GB/s

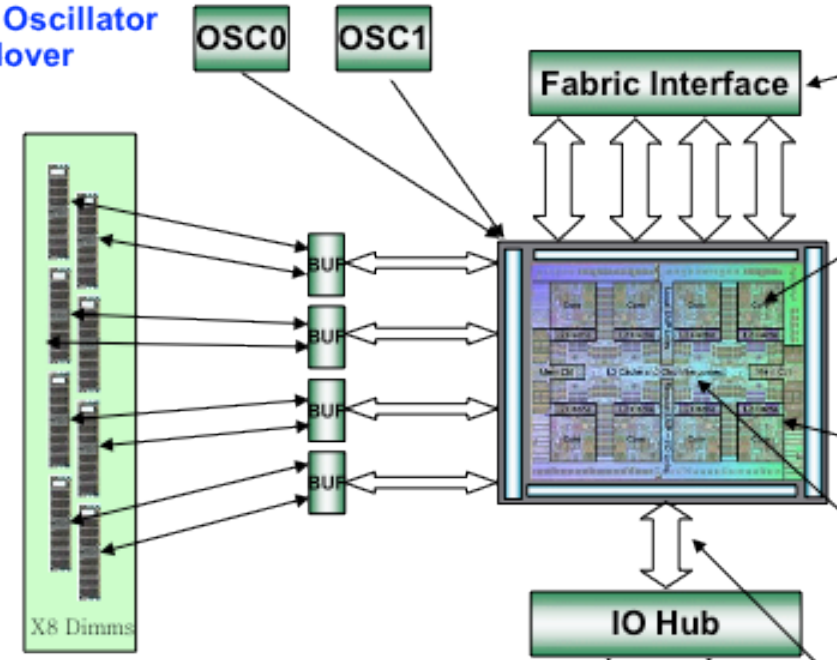
(Note that bandwidths shown are raw, peak signal bandwidths)

- Better signaling technology
- On-chip L3 reduces BW requirements
- Advanced scheduling improves BW utilization

Shared Memory and I/O



High-End Server Resilience for Good MTBF



- 64 Byte ECC on Memory
  - Corrects full chip kill on X8 dimms
  - Spare X8 devices implemented
- Dual memory chip failures do not cause outage
- Selective memory mirror capability to recover partition from dimm failures
- HW assisted scrubbing
- SUE handling
- Dynamic sparing on channel interface
- PowerVM Hypervisor protected from full dimm failures

Fabric Bus Interface to other Chips and Nodes
- ECC protected
- Node hot add /repair

Core Recovery
- Leverage speculative execution resources to enable recovery
- Error detected in GPRs FPRs VSR, flushed and retried
- Stacked latches to improve SER

Alternate Processor Recovery
- Partition isolation for core checkstops

L3 eDRAM
- ECC protected
- SUE handling
- Line delete
- Spare rows and columns

GX IO Bus
- ECC protected
- Hot add

InfiniBand® Interface
- Redundant paths

* Statements regarding SMP servers do not imply that IBM will introduce a system with this capability.

Feeds and Speeds per QCM

- 32 cores
- 8 Flop/cycle per core
- 4 threads per core max
- 3.5 – 4 GHz
- 1 TF/s
- 32 MB L3
- 512 GB/s memory BW (0.5 Byte/flop)
- 800 W (0.8 W/flop)

4 chips, 32 cores

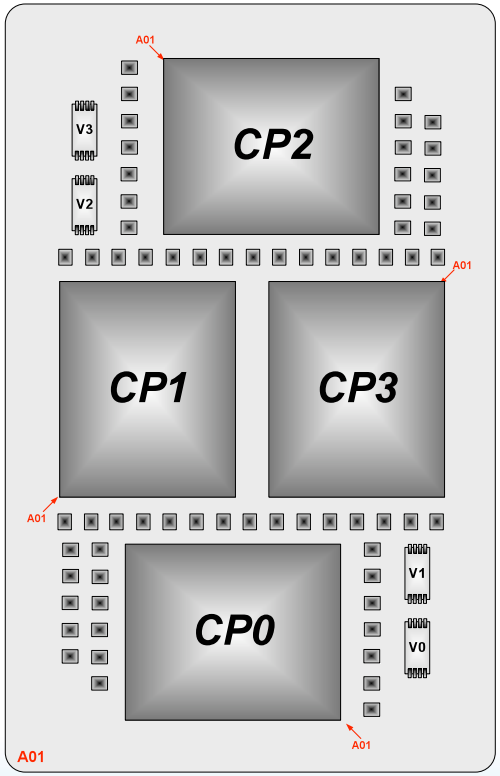
ONE MULTI-CHIP MODULE

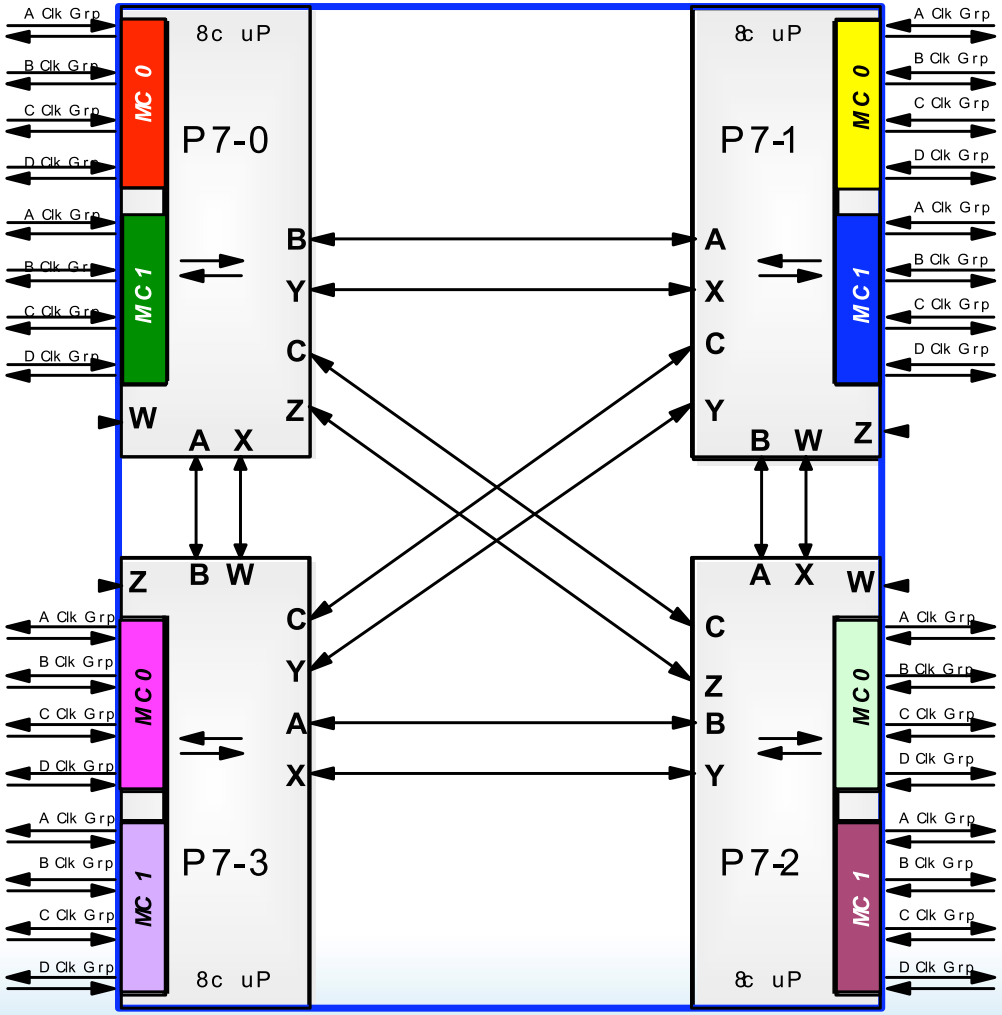
(aka QCM, Node, Octant)

Compute Intensive Quad-chip MCM

QCM

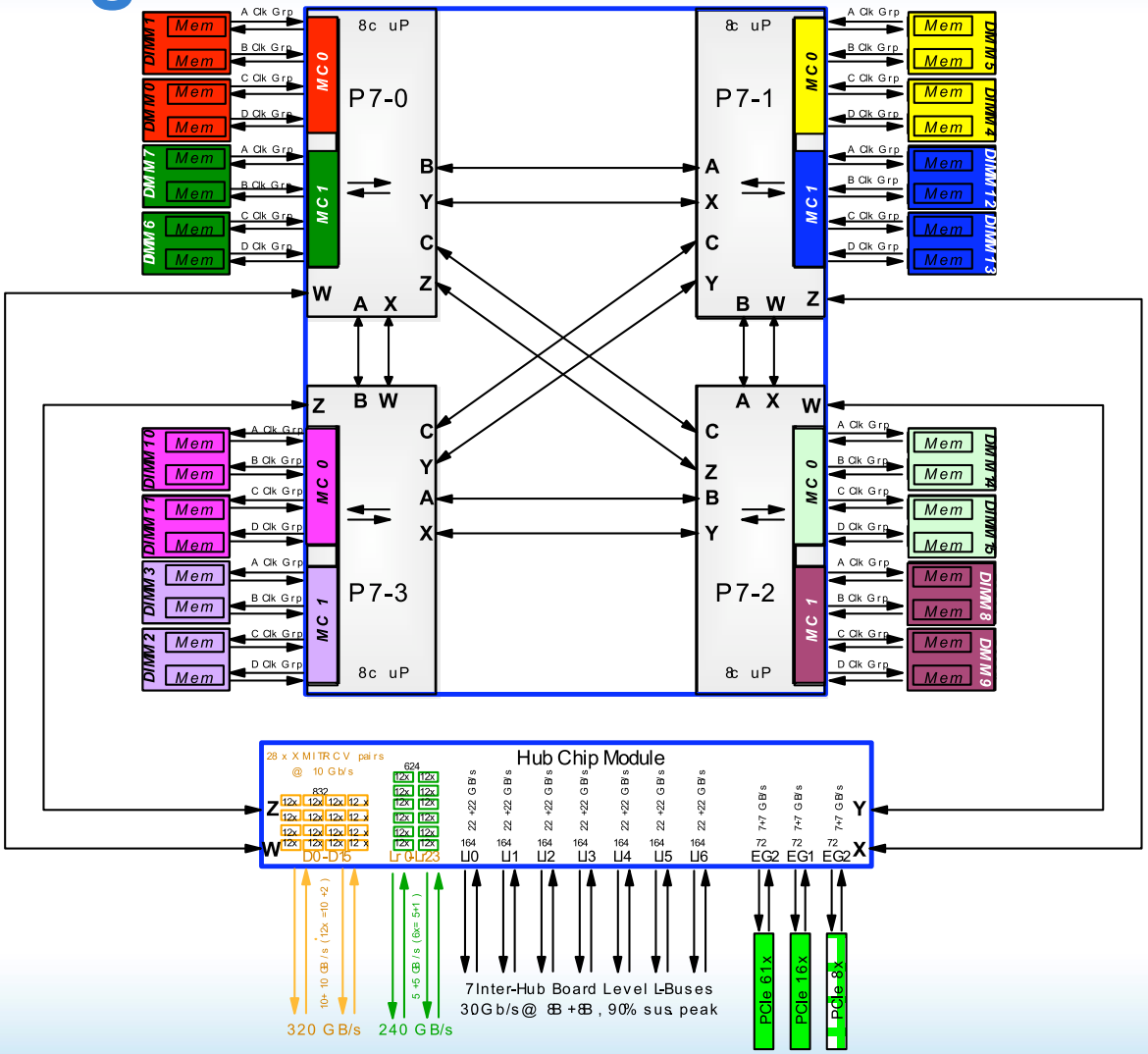
- 4 x p7
- 32 Cores (4 x 8 =32)
- Up to 512 GB memory





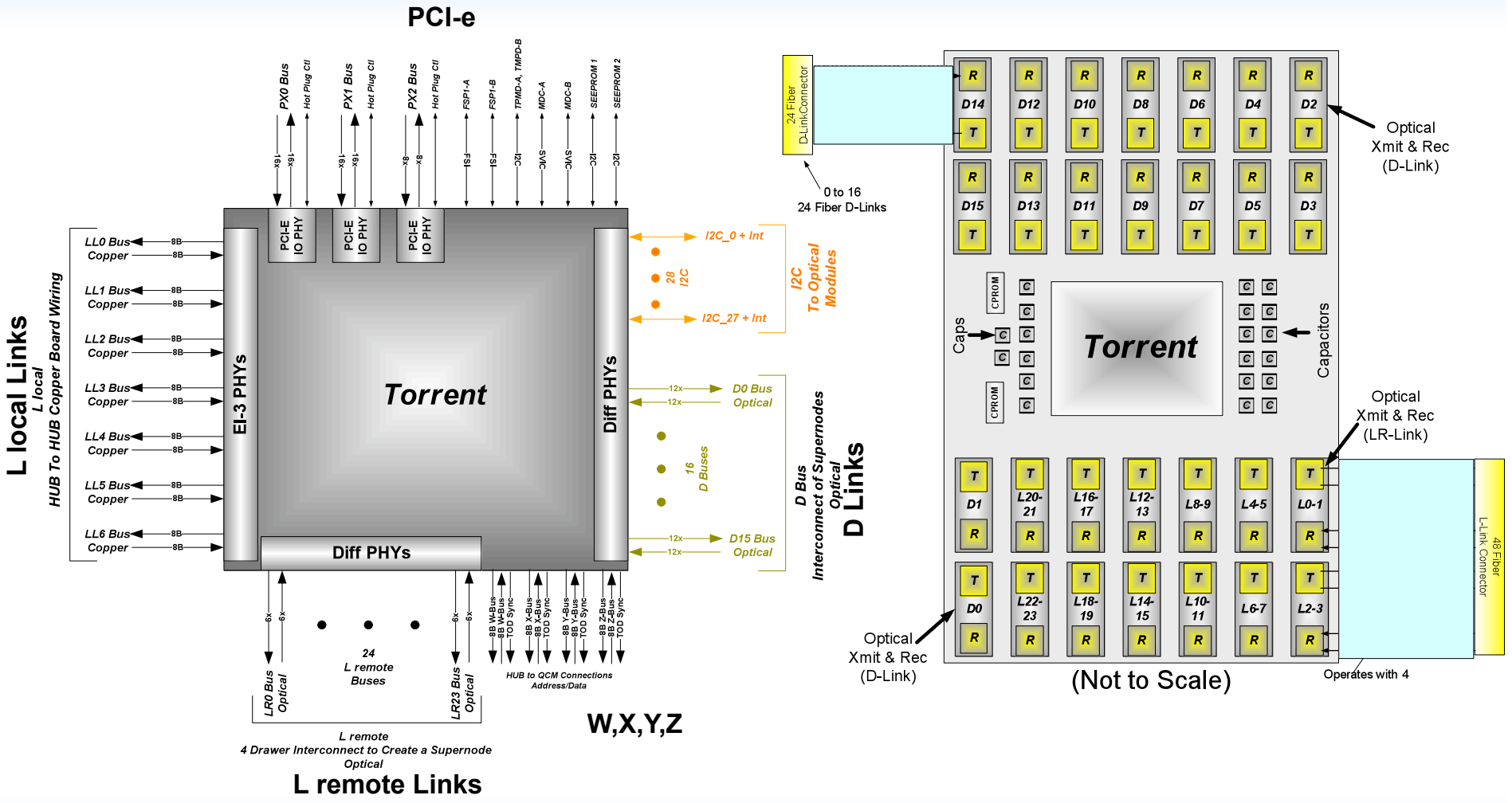
Node (Octant) - Logical

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



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


PCI-e
PX0 Bus
PX1 Bus
PX2 Bus
Hot Plug Ctl
FSP1-A
FSP1-B
TPMD-A, TMPD-B
MDC-A
MDC-B
SEEPROM 1
SEEPROM 2
PCI-E IO PHY
Torrent
EI-3 PHYs
Diff PHYs
LL0 Bus Copper
LL1 Bus Copper
LL2 Bus Copper
LL3 Bus Copper
LL4 Bus Copper
LL5 Bus Copper
LL6 Bus Copper
L local Links
L local
HUB To HUB Copper Board Wiring
I2C_0 + Int
I2C_27 + Int
28 I2C
I2C To Optical Modules
D0 Bus Optical
D15 Bus Optical
16 D Buses
D Bus Interconnect of Supernodes Optical
D Links
24 L remote Buses
LR0 Bus Optical
LR23 Bus Optical
HUB to QCM Connections Address/Data
L remote
4 Drawer Interconnect to Create a Supernode
Optical
L remote Links
W,X,Y,Z
24 Fiber D-LinkConnector
0 to 16
24 Fiber D-Links
D14 D12 D10 D8 D6 D4 D2
D15 D13 D11 D9 D7 D5 D3
Optical Xmit & Rec (D-Link)
CPROM
Caps
Capacitors
Optical Xmit & Rec (LR-Link)
D1 L20-21 L16-17 L12-13 L8-9 L4-5 L0-1
D0 L22-23 L18-19 L14-15 L10-11 L6-7 L2-3
Optical Xmit & Rec (D-Link)
48 Fiber L-Link Connector
Operates with 4
(Not to Scale)

1.1 TB/s HUB

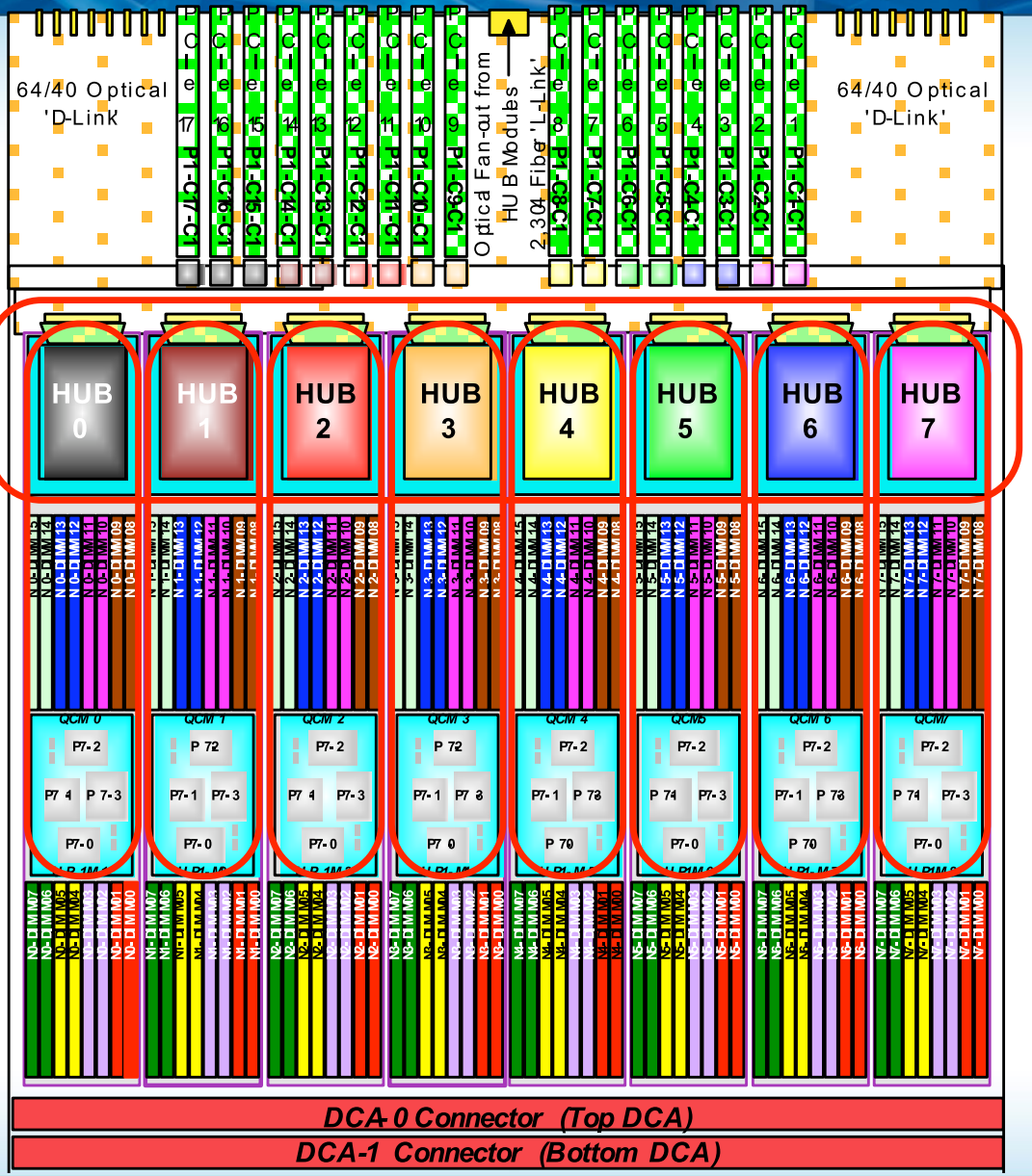
- 192 GB/s Host Connection
- 336 GB/s to 7 other local nodes
- 240 GB/s to local-remote nodes
- 320 GB/s to remote nodes
- 40 GB/s to general purpose I/O

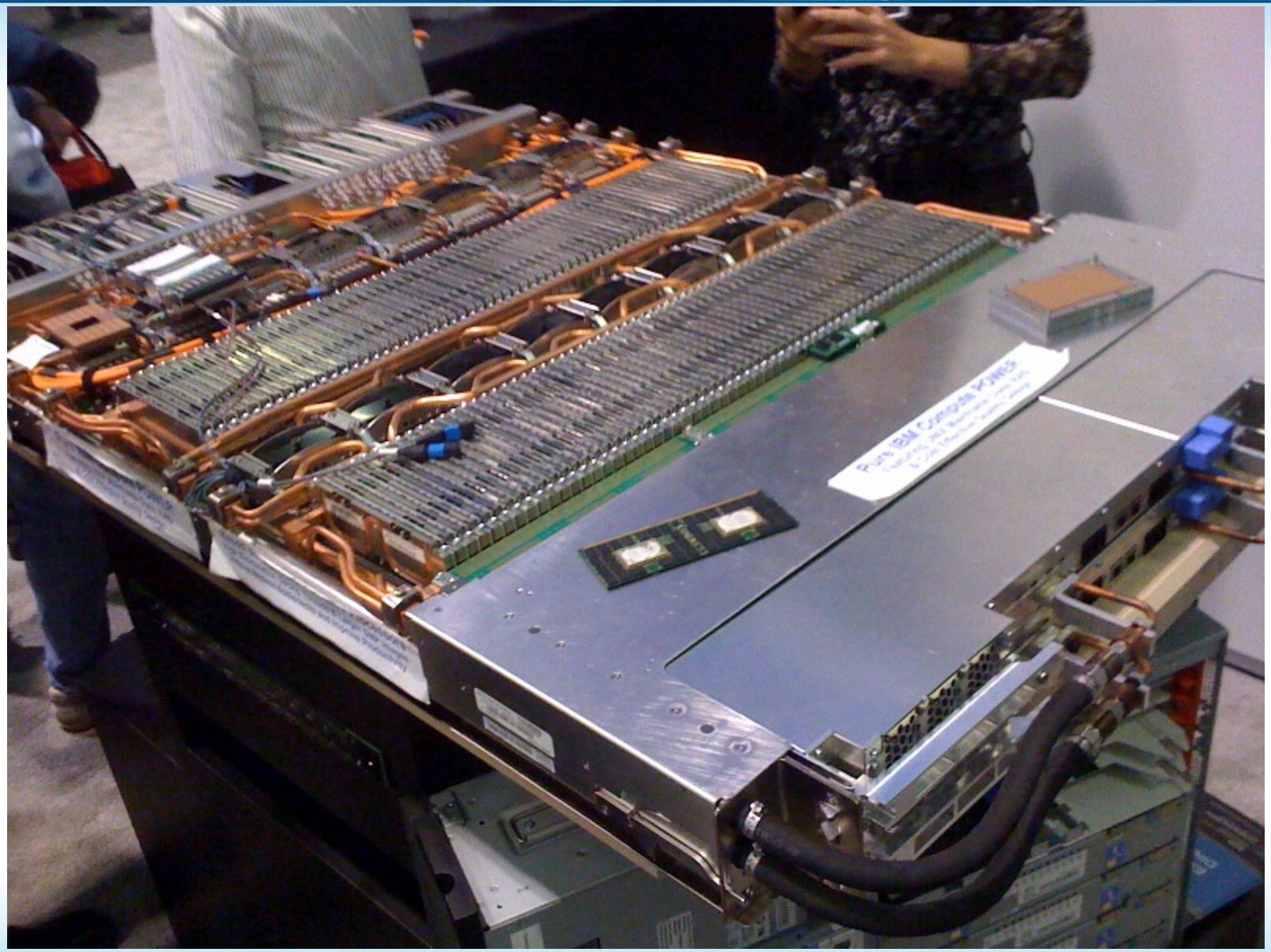
8 MCMs, 32 chips, 256 cores

ONE DRAWER

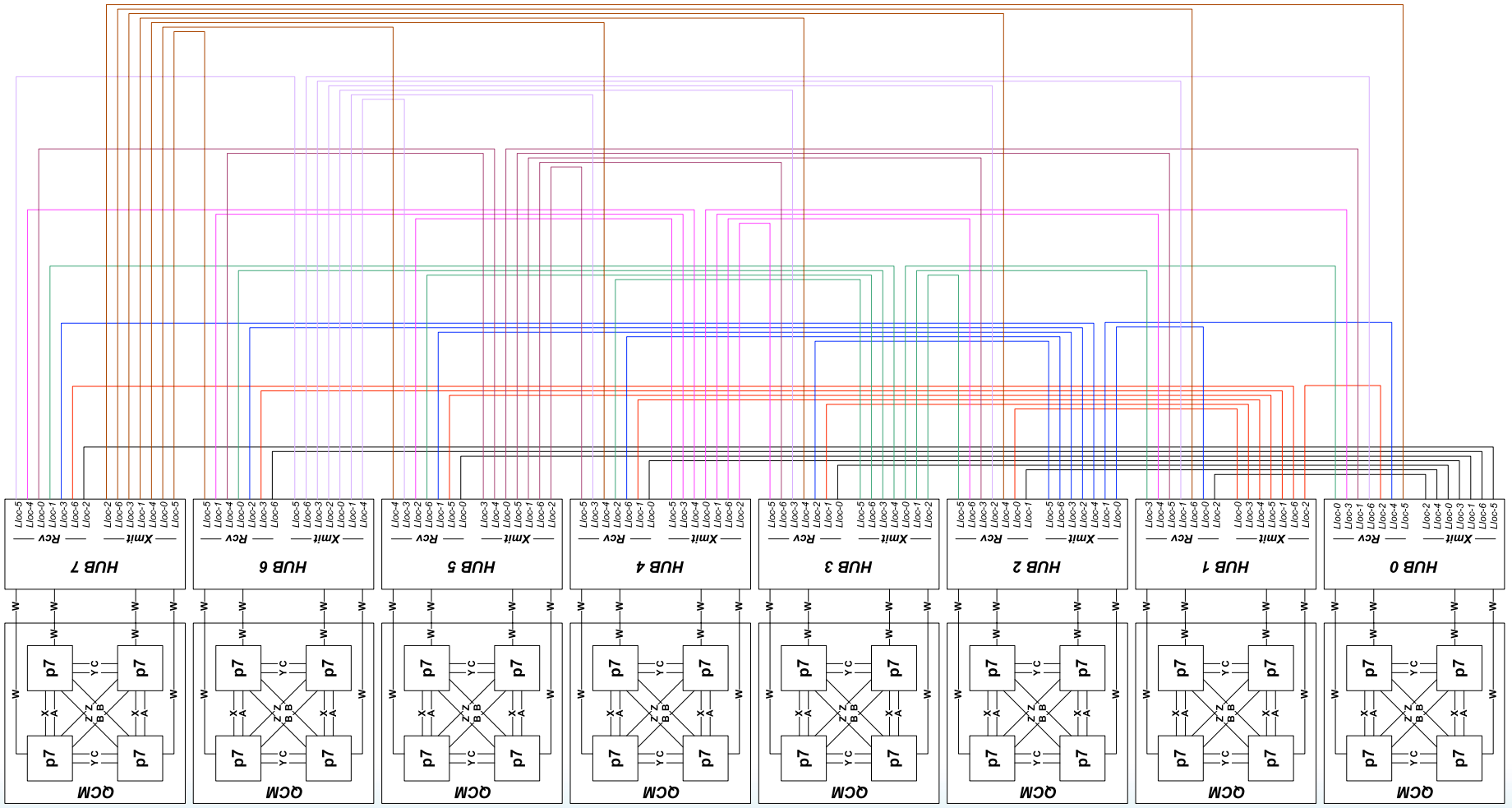
First Level Interconnect

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





Every HUB Connected to Every Other HUB in CEC



(CEC = Central Electronic Complex: IBM Jargon for drawer or packaging unit)

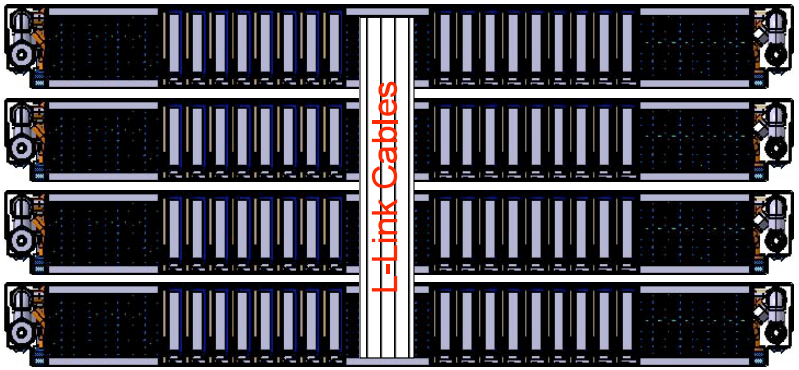
4 drawers, 32 MCMs, 128 chips, 1024 cores

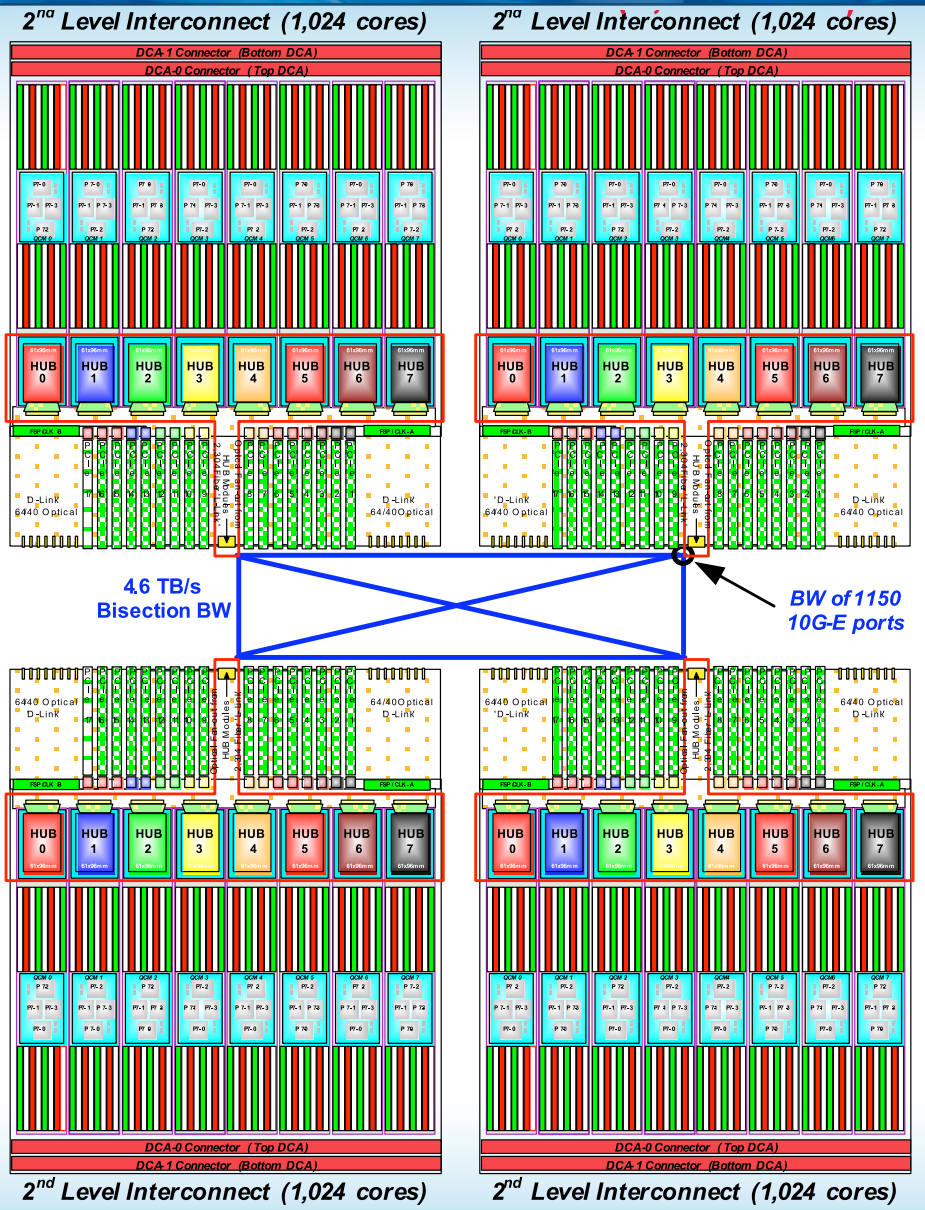
ONE SUPERNODE

Second Level Interconnect

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

Super Node
(32 Nodes / 4 CEC)

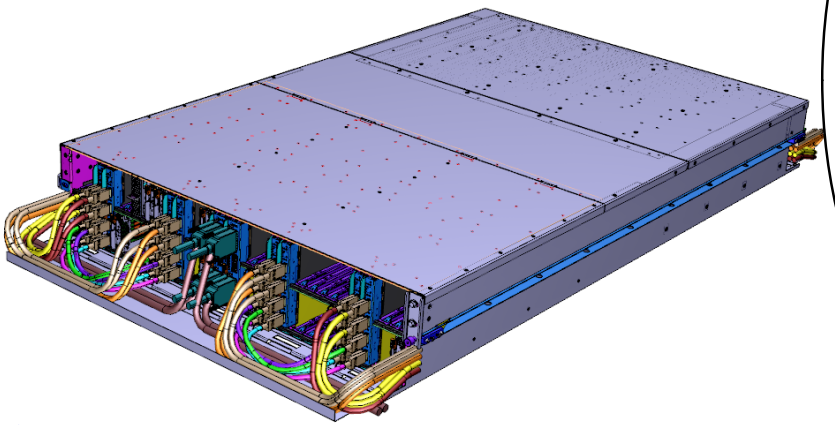


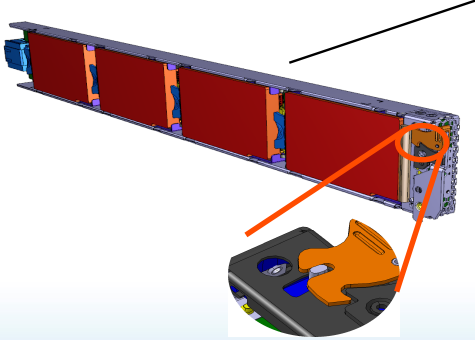


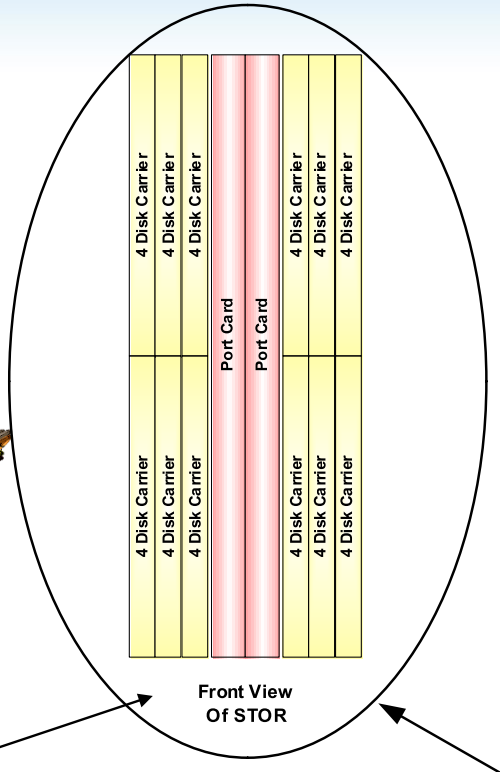
DISKS

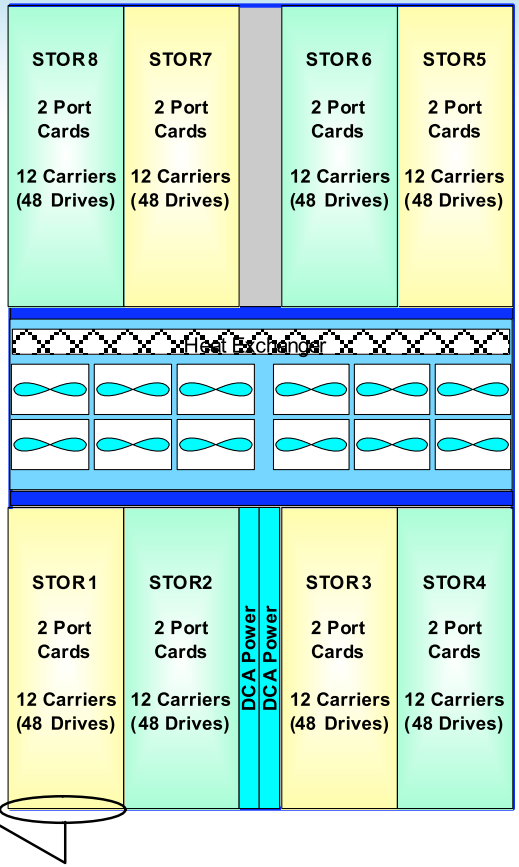
Disk Storage

- Disk Enclosure is Connected to CEC via SAS PCIe Cards
- Redundant DCA / DE
- Up to 384 SFF DASD drives / DE
- 8 Storage Groups (STOR 1-8) with 48 drives each
  - ✓ Two Port Cards per STOR
  - ✓ 12 carriers per STOR
  - ✓ Carriers contains up to 4 drives each
- Data Rates:
  - ✓ Serial Attach SCSI (SAS) SDR = 3.0 Gbps per lane (SEC to Drive)
  - ✓ Serial Attach SCSI (SAS) DDR = 6.0 Gbps per lane (SAS Adapter in Node to SEC)






4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
Port Card
Port Card
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
4 Disk Carrier
Front View
Of STOR


STOR 8
2 Port Cards
12 Carriers (48 Drives)
STOR7
2 Port Cards
12 Carriers (48 Drives)
STOR 6
2 Port Cards
12 Carriers (48 Drives)
STOR5
2 Port Cards
12 Carriers (48 Drives)
Heat Exchanger
STOR 1
2 Port Cards
12 Carriers (48 Drives)
STOR2
2 Port Cards
12 Carriers (48 Drives)
DCA Power
DCA Power
STOR 3
2 Port Cards
12 Carriers (48 Drives)
STOR4
2 Port Cards
12 Carriers (48 Drives)

1-12 2U CECs (256 – 3073 cores)

0-6 4U drawers (0-2304 disks)

ONE RACK

Rack

- 990.6w x 1828.8d x 2108.2
- 39”w x 72”d x 83”h
- ~2948kg (~6500lbs)

Data Center In a Rack

Compute

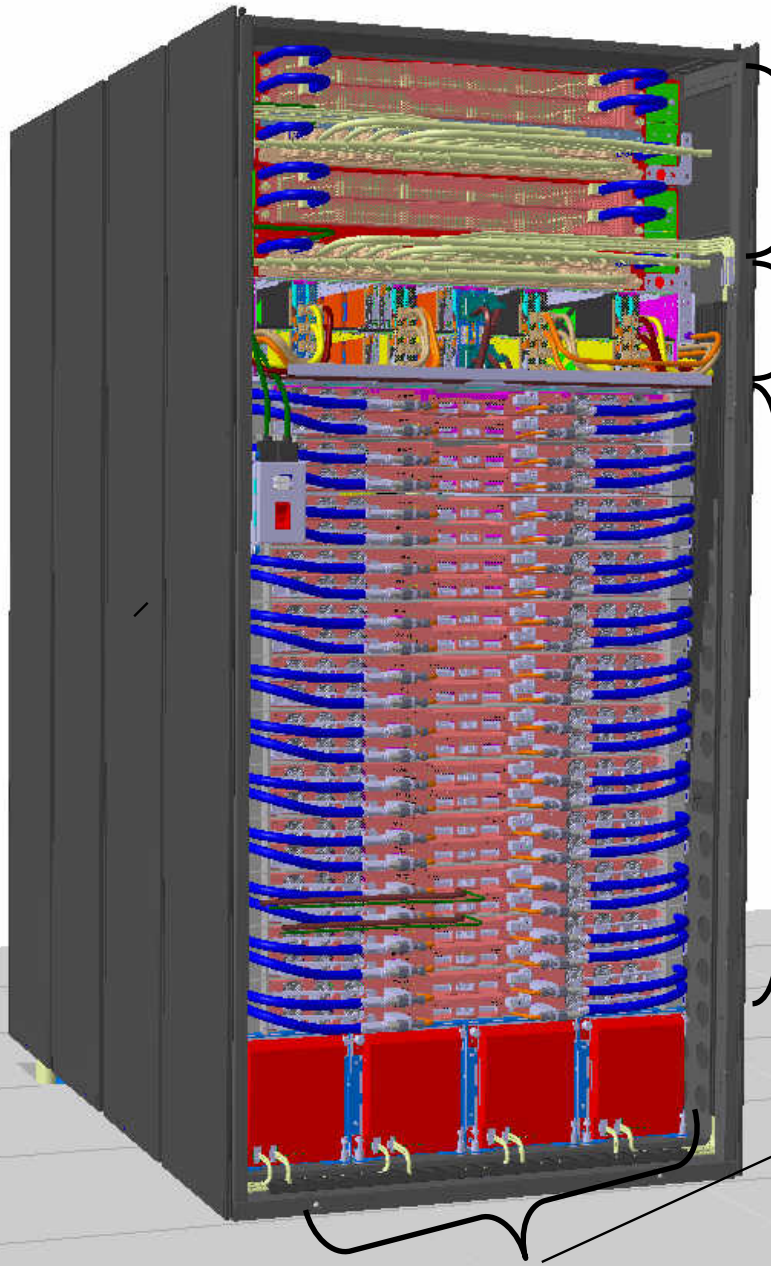
Storage

Switch

100% Cooling

PDU Eliminated

Input: 8 Water Lines, 4 Power Cords
Out: ~100TFLOPs / 24.6TB / 153.5TB
192 PCI-e 16x / 12 PCI-e 8x



BPA

- 200 to 480Vac
- 370 to 575Vdc
- Redundant Power
- Direct Site Power Feed
- PDU Elimination

Storage Unit

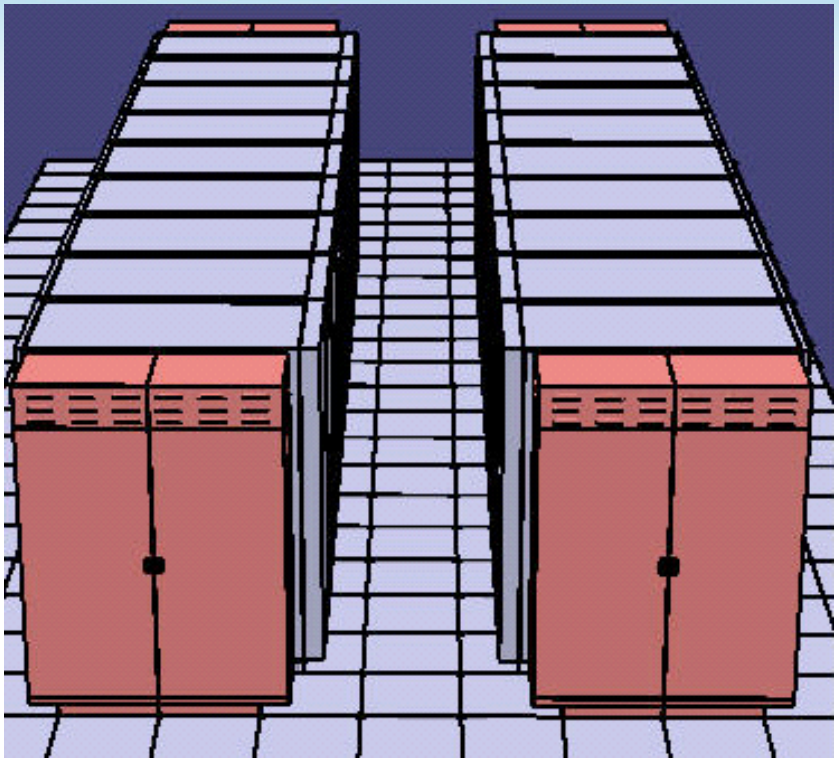
- 4U
- 0-6 / Rack
- Up To 384 SFF DASD / Unit
- File System

CECs

- 2U
- 1-12 CECs/Rack
- 256 Cores
- 128 SN DIMM Slots / CEC
- 8,16, (32) GB DIMMs
- 17 PCI-e Slots
- Imbedded Switch
- Redundant DCA
- NW Fabric
- Up to:3072 cores, 24.6TB (49.2TB)

WCU

- Facility Water Input
- 100% Heat to Water
- Redundant Cooling
- CRAH Eliminated

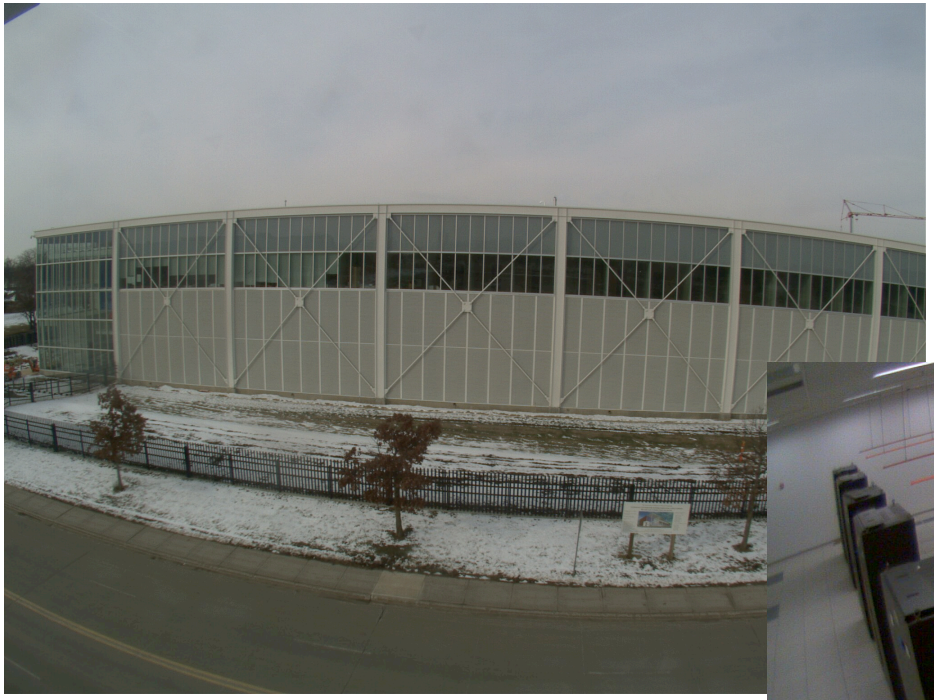


ONE SYSTEM: BLUE WATERS

Blue Waters

- Approximately 10 PF/s peak
- More than 300,000 cores
- More than 1 PetaByte memory
- More than 10 Petabyte disk storage
- More than 0.5 Exabyte archival storage
- More than 1 PF/s sustained on scientific applications

National Petascale Computing Facility



A facility dedicated to Blue Waters