BLUE WATERS

SUSTAINED PETASCALE COMPUTING

2010 Blue Waters Performance Modeling Workshop – Opening and Introduction

Torsten Hoefler

With slides from: William Kramer, Marc Snir, William Gropp, IBM, and the Blue Waters team

NCSA IBM GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Introduction and Overview

- **My slides** contain **only public information** and will be available online after the workshop
  - No need to take pictures or notes!
- Parts of **tomorrow** will contain **IBM confidential** information
  - You may only attend the NDA session if your institution signed and cleared all NDAs for you!
  - You are responsible to maintain the confidentiality of the information!

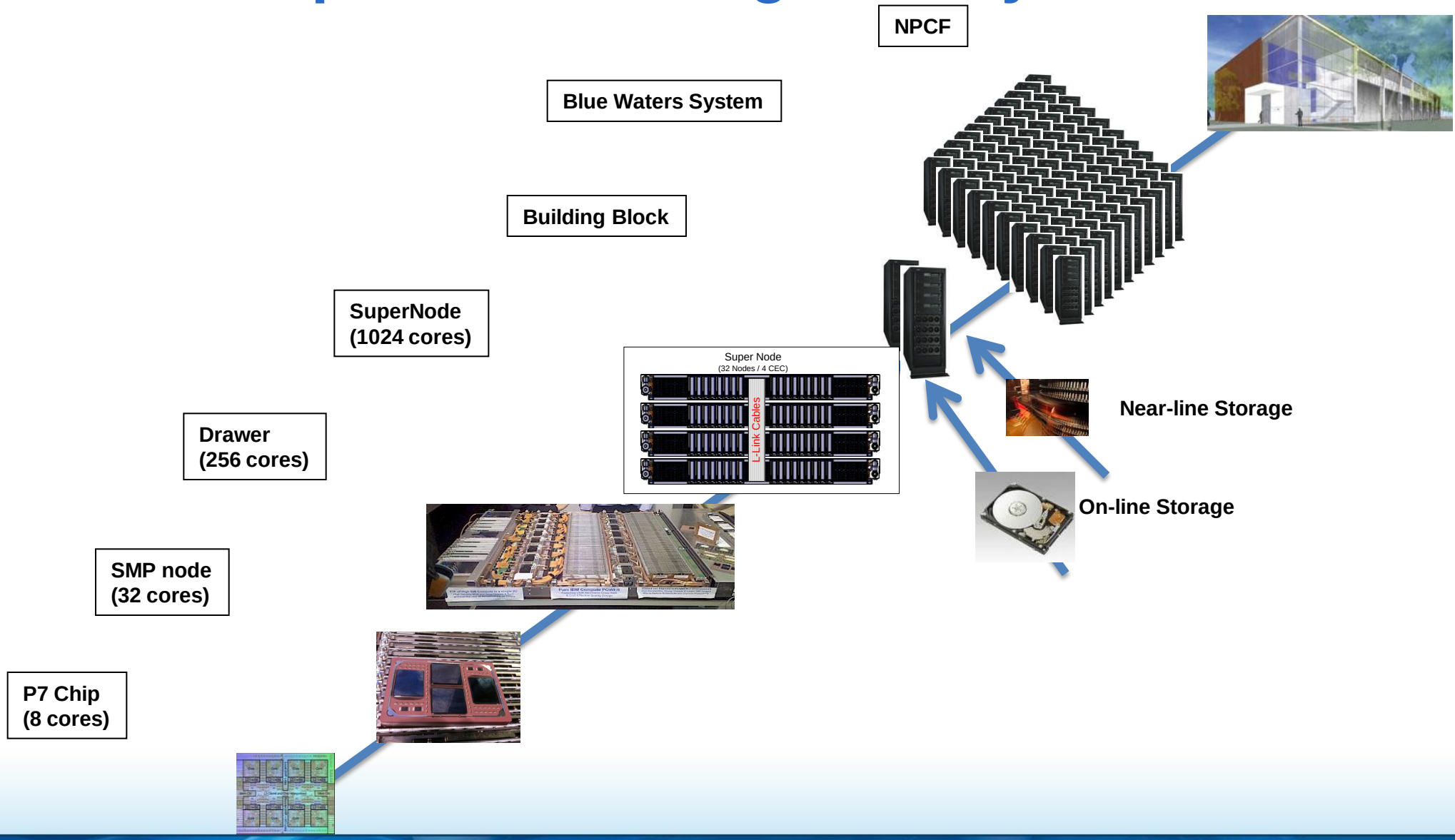
Blue Waters in a Nutshell

- >300.000 compute cores
  - based on Power7
- 10 PF/s peak
  - 1 PF/s sustained
- >1 PiB RAM
- >10 PiB disk storage
- >0.5 EiB archival storage

Performance Modeling for Blue Waters

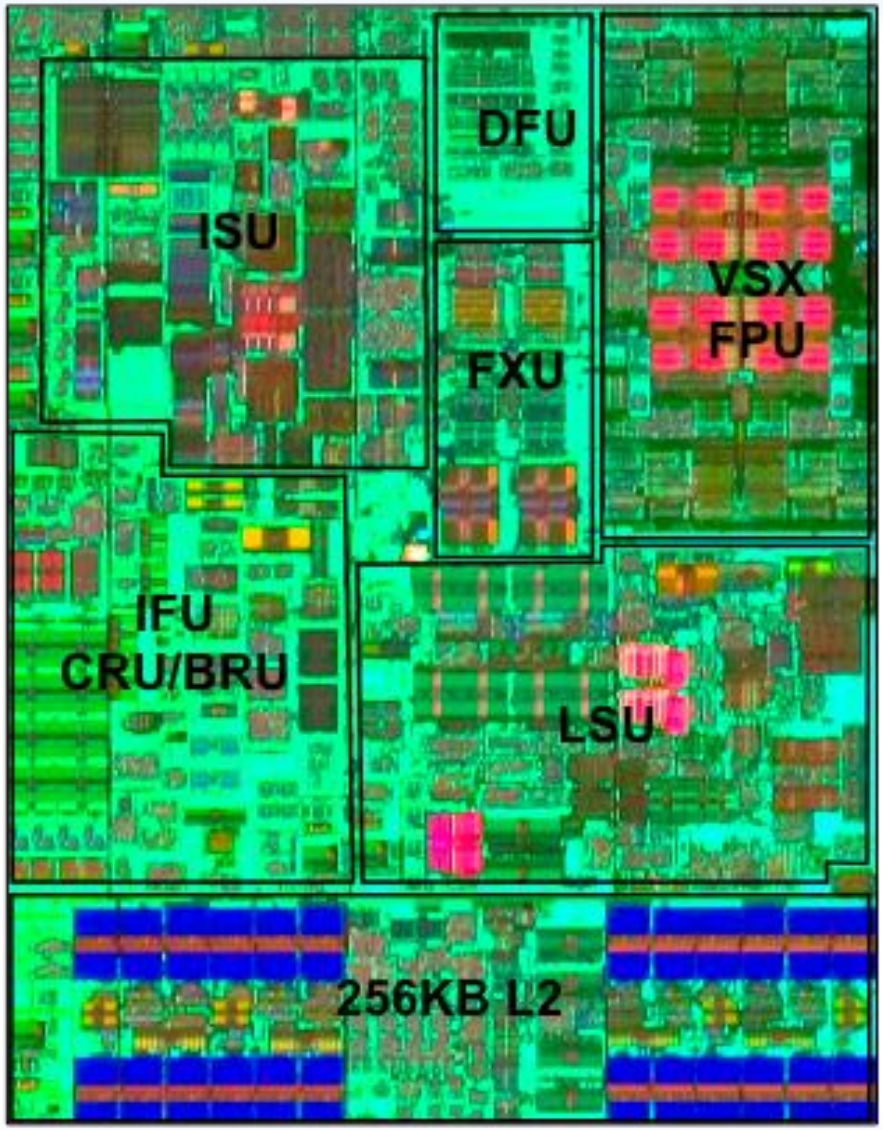
- Most users have only experience at comparatively “small” scale (<8000 cores)
- Applications should be ready to run on the full system
- Needs a clear understanding before system is deployed (run, tweak, rerun loop not possible)
- ➢ Programmers need to develop a deep understanding of the application scaling and bottlenecks at scale by performance modeling!

From Chip to Entire Integrated System



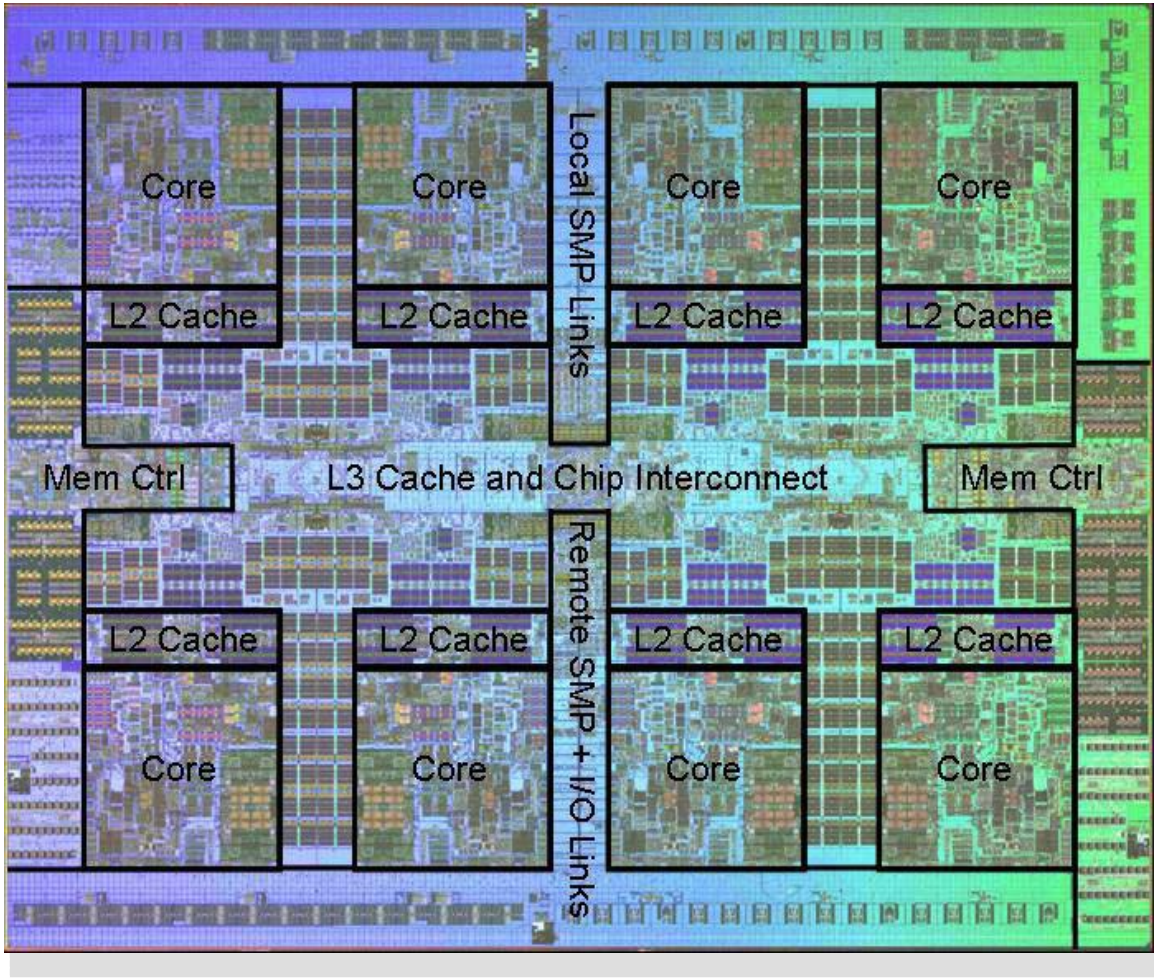
POWER7: Core

- 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



Power7 Chip (8 cores)

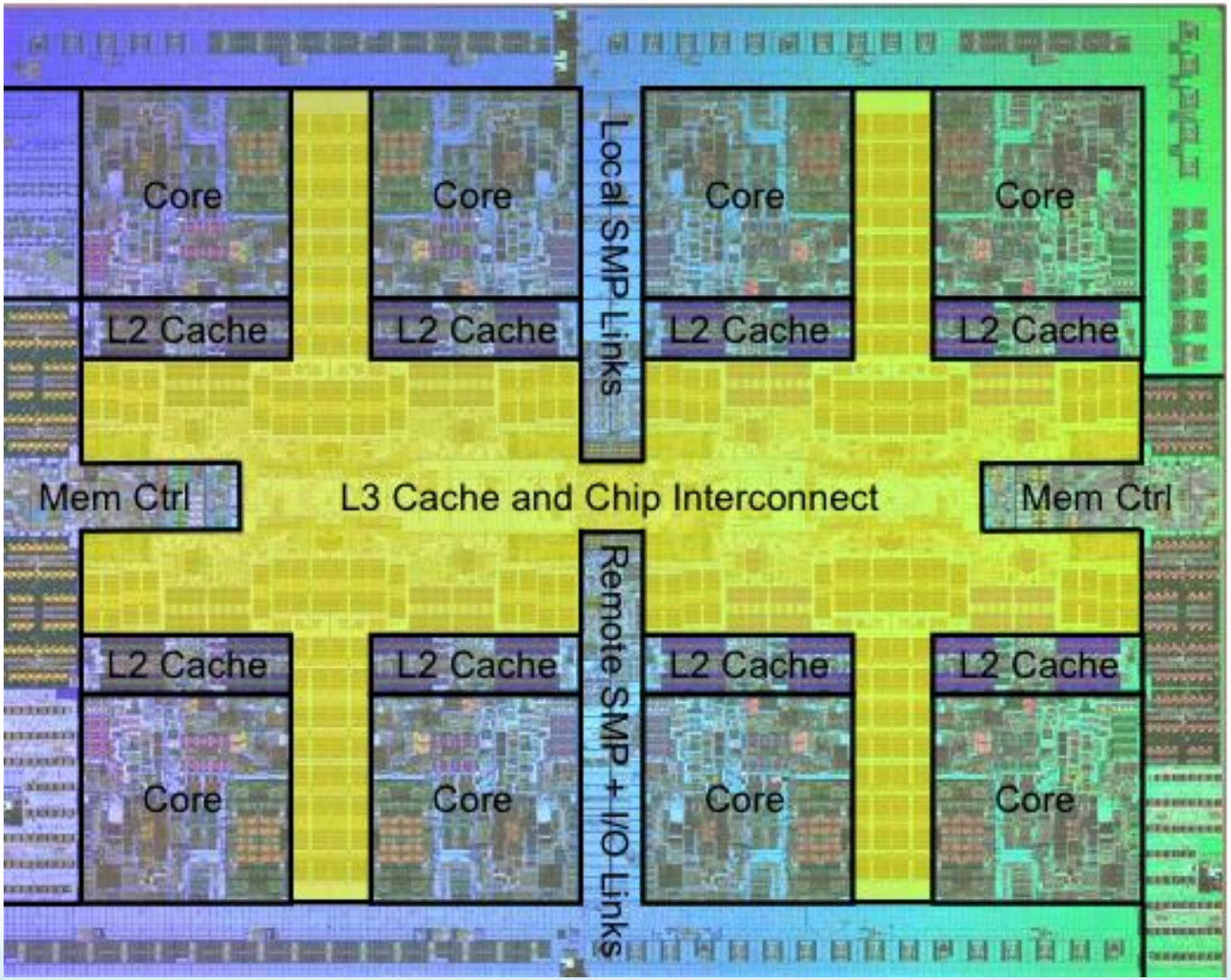
- Base Technology
  - 45 nm, 576 mm2
  - 1.2 B transistors
- Chip
  - 8 cores
  - 4 FMAs/cycle/core
  - 32 MB L3 (private/shared)
  - Dual DDR3 memory
    - 128 GiB/s peak bandwidth
    - (1/2 byte/flop)
  - Clock range of 3.5 – 4 GHz



Quad-chip MCM

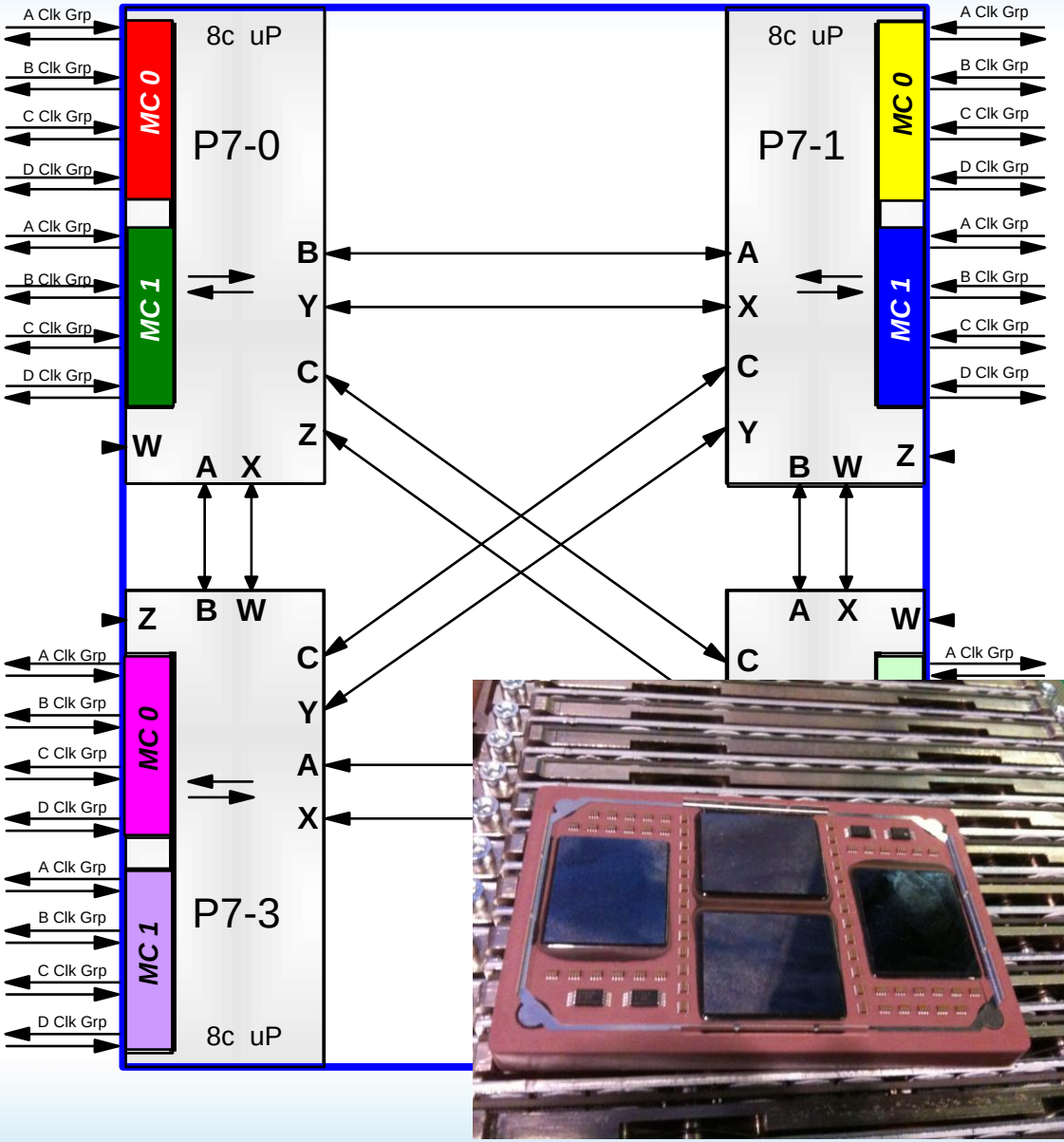
L3 Cache/On-Chip Communication

- L1 32KB Instruction / core
- L1 32KB Data / core
- L2 = 256KB / core
- L3 = 4MB eDRAM / core
- Fast private and shared region



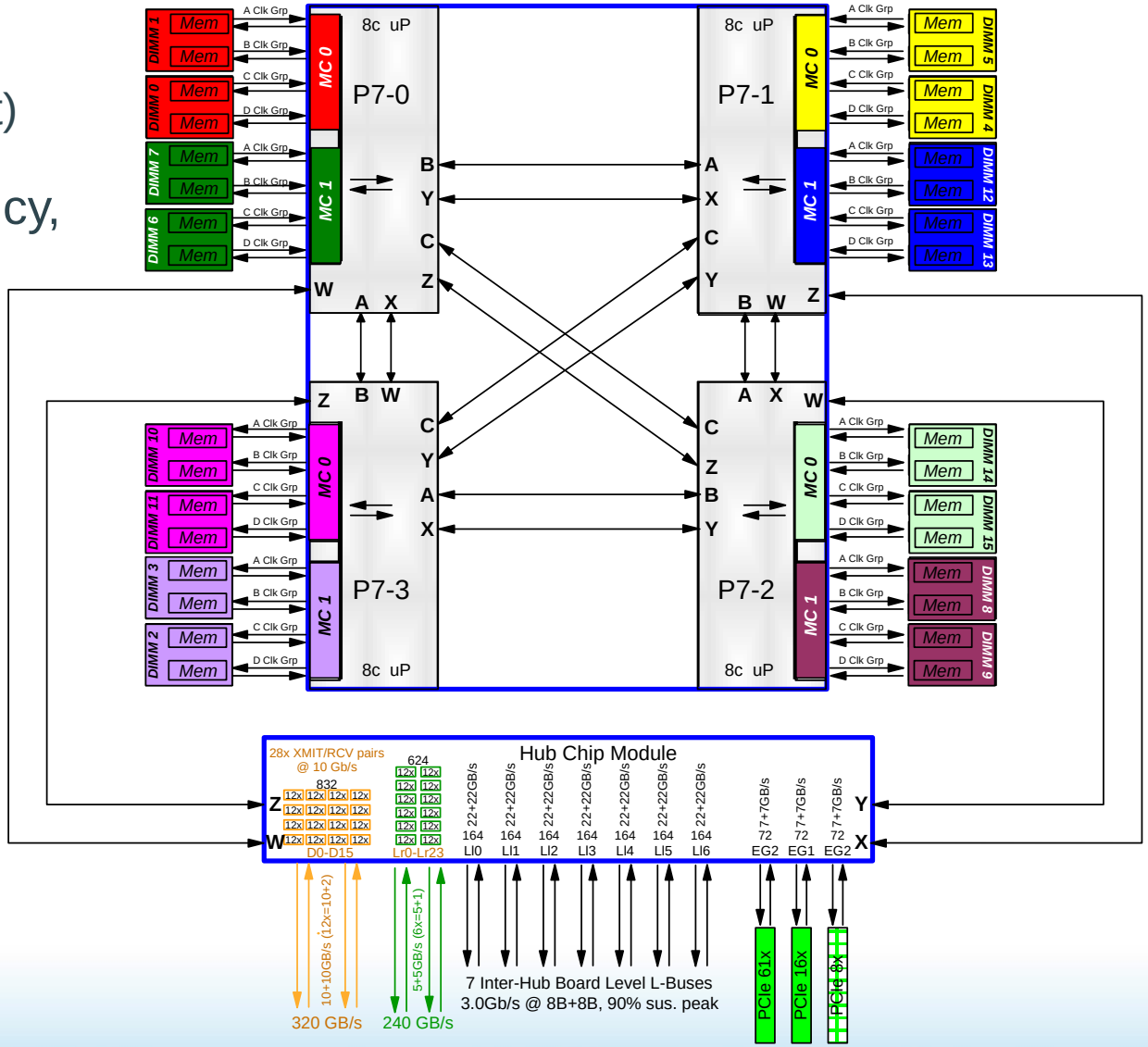
Quad Chip Module (4 chips)

- 32 cores!
  - 32 cores*8 F/core*4 GHz = 1 TF
- 4 threads per core (max)
- 4x32 MiB L3 cache
- 512 GB/s RAM BW (0.5 B/F)
- 800 W (0.8 W/F)
- Flat shared memory!



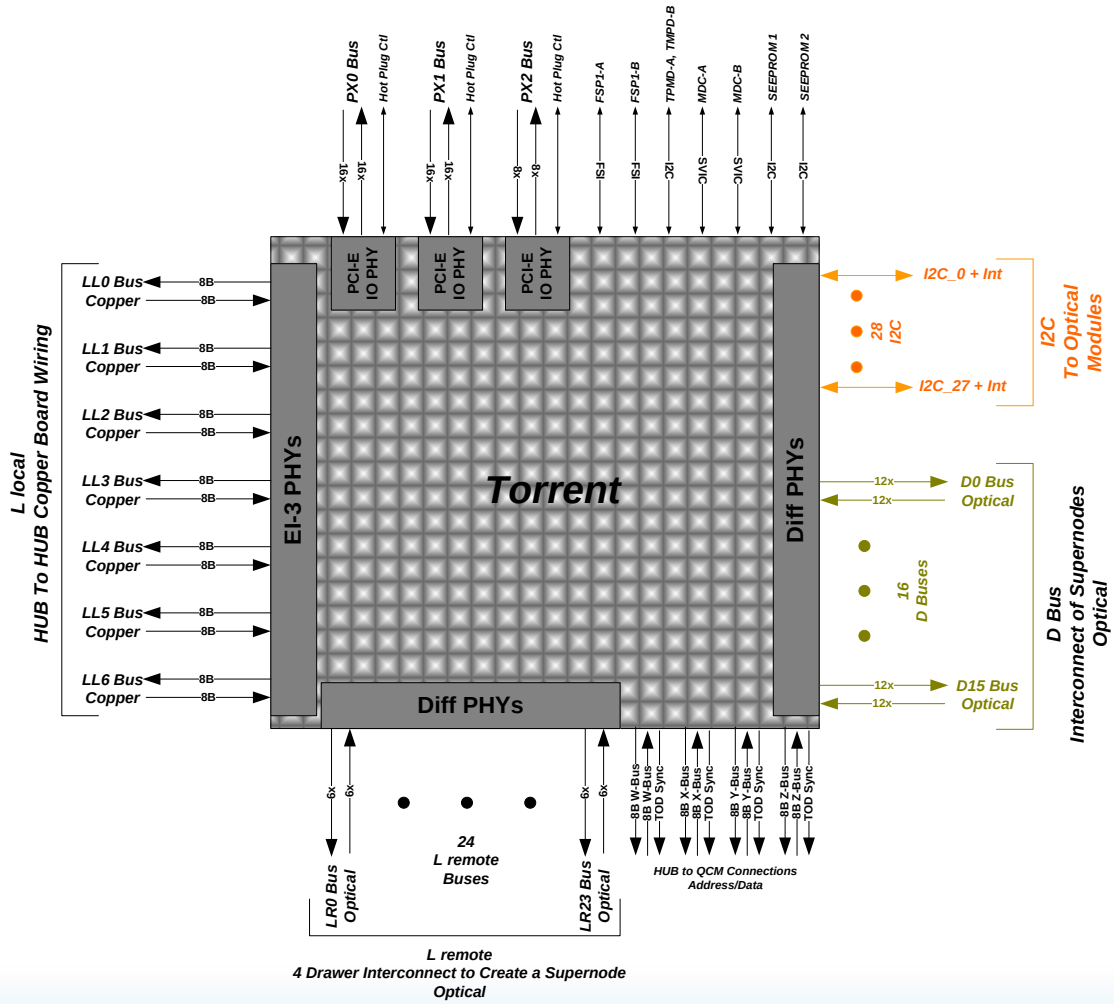
Adding a Network Interface (Torrent)

- Connects QCM to PCI-e
  - (two 16x and one 8x PCI-e slot)
- Connects 8 QCM's via low latency, high bandwidth, copper fabric.
  - Provides a message passing mechanism with very high bandwidth
  - Provides the lowest possible latency between 8 QCM's



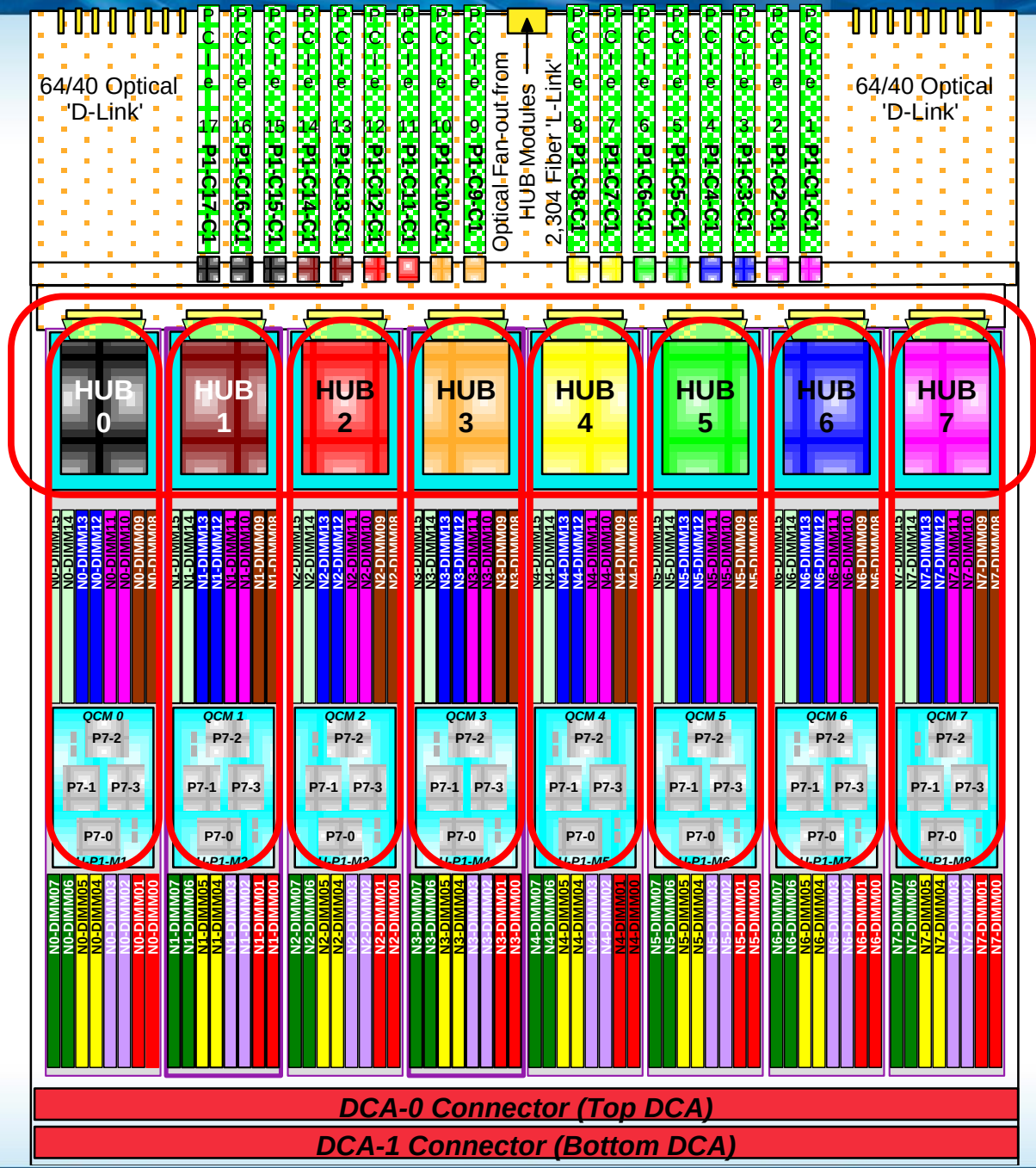
1.1 TB/s HUB

- 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



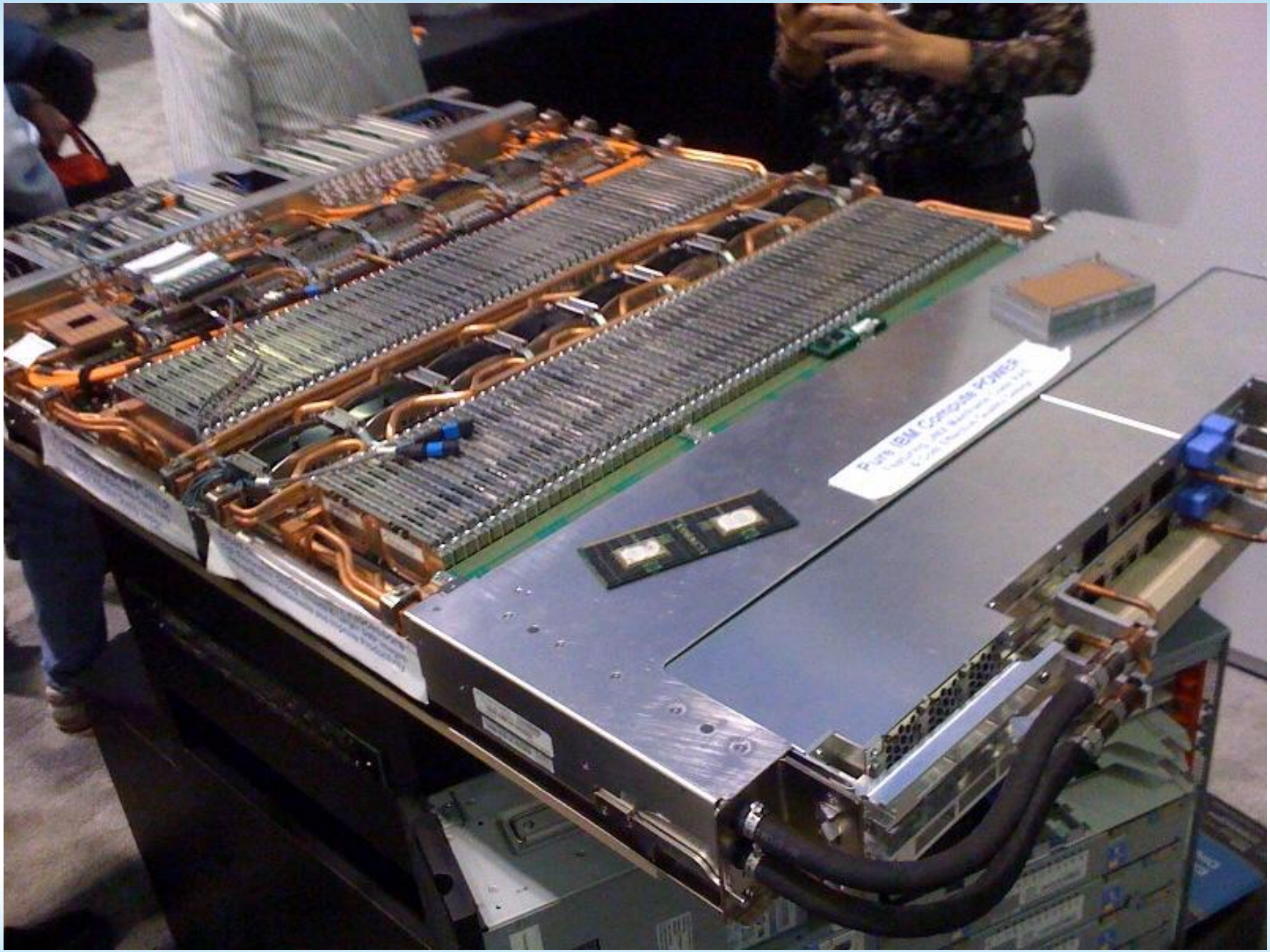
Drawer

- 8 nodes
  - 32 chips
  - 256 cores



First Level Interconnect

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

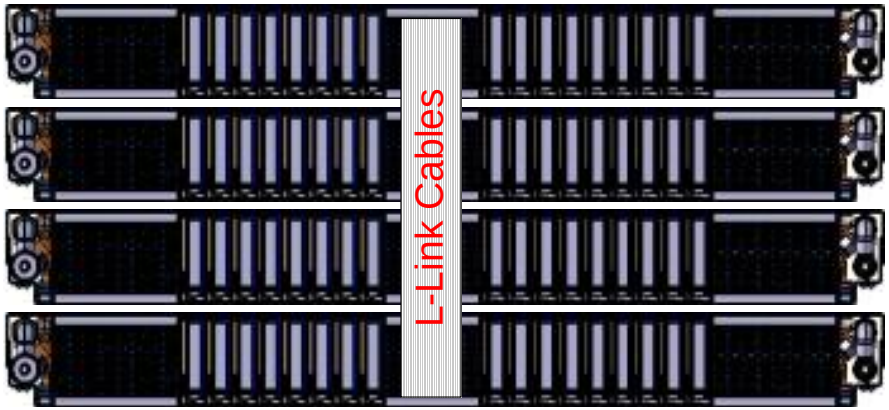


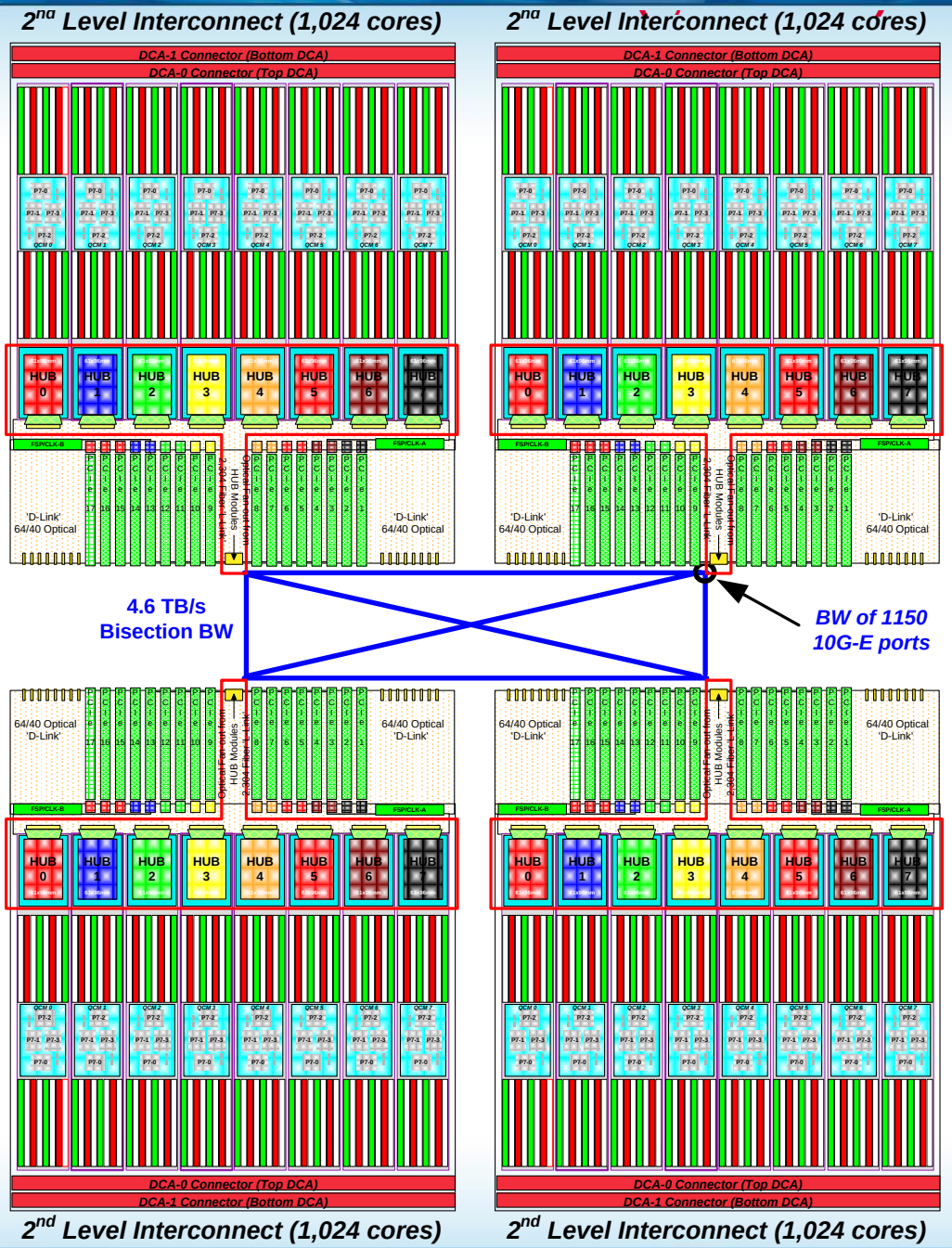
Supernode

Second Level Interconnect

- Optical 'L-Remote' Links from HUB
- 4 drawers
- 1,024 Cores

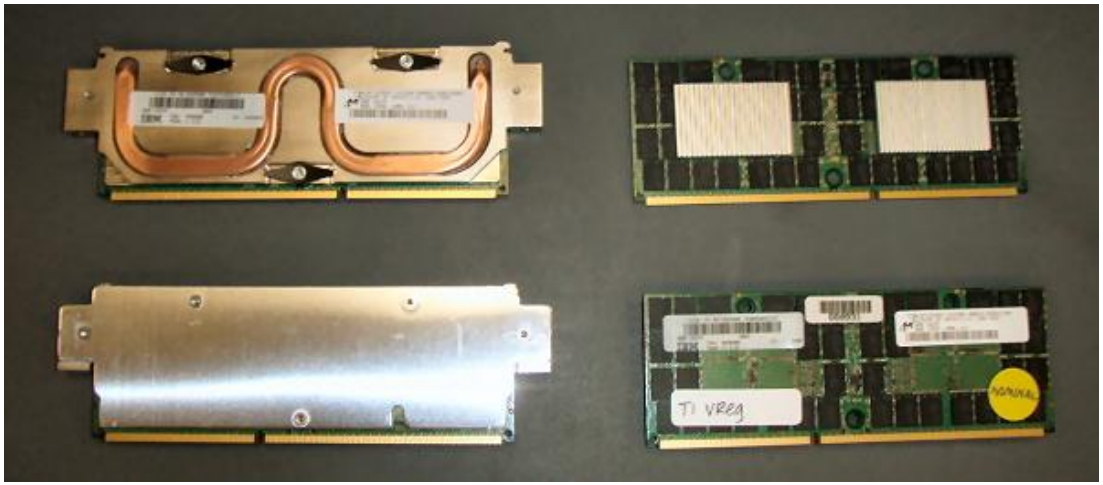
Super Node
(32 Nodes / 4 CEC)





Global Interconnection Network

- **This space is intentionally left blank**
  - More details in the NDA sessions



A photo of the RAM for distraction.

National Petascale Computing Facility

A facility dedicated to Blue Waters

Back to Performance Modeling

- **Main goals of this workshop:**
  - Ignite performance modeling efforts within all PRAC teams in collaboration with NCSA
  - Start to gather a deep understanding of the performance characteristics of all codes

Logistics

- **Today:**
  - A team from LANL will present a tutorial about performance modeling
    - Specific examples and use-cases
- **Tomorrow:**
  - Hands-on sessions to get modeling of applications started
    - Supported by LANL and NCSA teams
    - Try to work with your PoC

NDA issues

- **Not all participants are covered by all necessary NDAs**
  - Badges will be marked

- **Please be careful what you talk about**
  - You are responsible for the information
  - Everything in my slides can be communicated freely!

I'm here to help!

- **We have 15 training accounts on a Power 5 available for tomorrow**
  - It's AIX
  - Ask me if you need one

- **Let me know if you have questions, problems, or comments!**