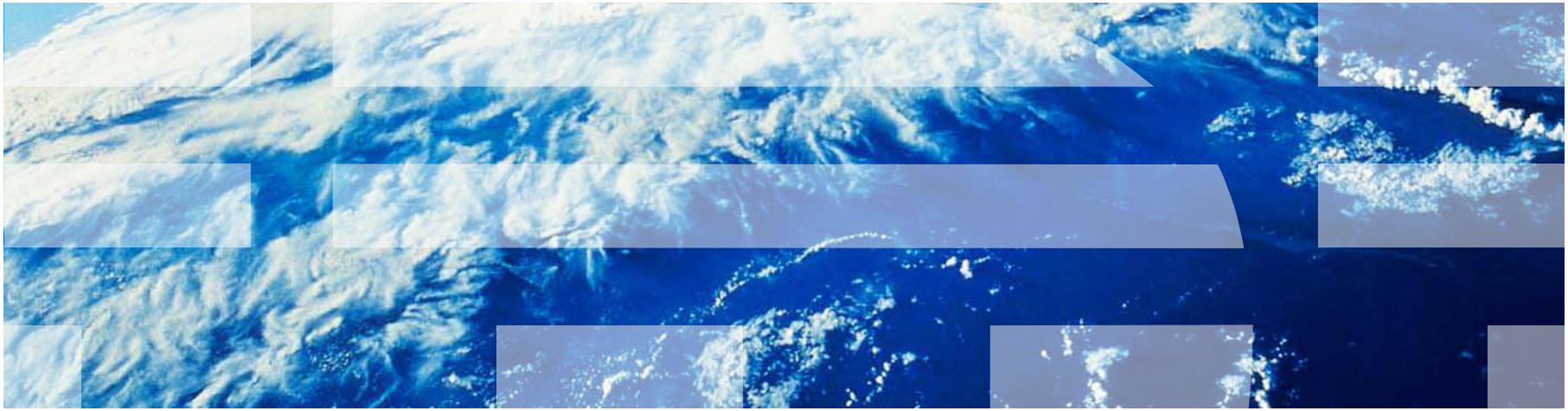
The IBM POWER7 HUB Module: A Terabyte Interconnect Switch for High-Performance Computer Systems

Baba Arimilli: Chief Architect

Steve Baumgartner: Senior Design Engineer



Scott Clark, Dan Dreps, Dave Siljenberg & Andrew Maki

Outline

- **PERCS Overview**
  - HPCS Program Background
  - Design Point
  - Hierarchical Structure and Interconnect

- **POWER7 Hub Chip**
  - Overview
  - Key Functional Units
  - Chip Floorplan / Metrics
  - Summary

- **POWER7 Hub Module and Off-chip Interconnect**
  - Summary of I/Os and PLL's
  - 10Gb/s Physical Transport Circuit Architectures
  - Hardware Characterization
  - Summary

"This design represents a tremendous increase in the use of optics in systems, and a disruptive transition from datacom- to computercom-style optical interconnect technologies."

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002.

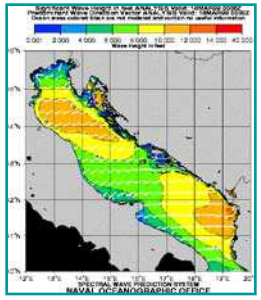
Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

DARPA's "High Productivity Computing Systems" Program

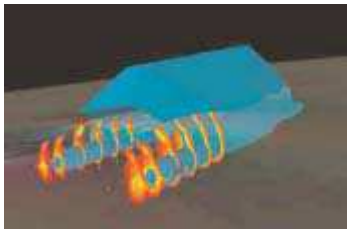
Goal: Provide a new generation of economically viable high productivity computing systems for the national security and industrial user community

- **Impact**:
  - **Performance** (time-to-solution): speedup by **10X to 40X**
  - **Programmability** (idea-to-first-solution): dramatically reduce cost & development time
  - **Portability** (transparency): insulate software from system
  - **Robustness** (reliability): continue operating in the presence of localized hardware failure, contain the impact of software defects, & minimize likelihood of operator error
- **Applications**:

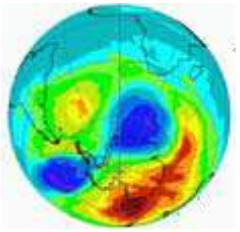
Weather Prediction



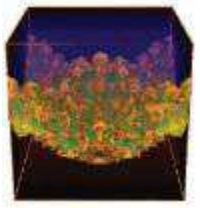
Ocean/wave Forecasting



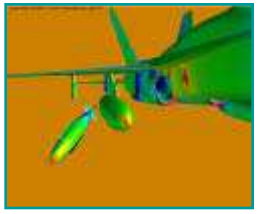
Ship Design



Climate Modeling



Nuclear Stockpile Stewardship



Weapons Integration

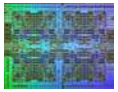
PERCS – Productive, Easy-to-use, Reliable Computing System is IBM's response to DARPA's HPCS Program

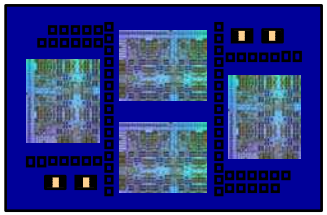
High-level Requirements and Design Point

- **High Bisection Bandwidth**
  - Hub chip with high fan-out
- **Low Latency**
  - Two-Level Direct Connect Topology for Interconnect
- **High Interconnect Bandwidth (even for packets < 50 bytes)**
  - Architect switch pipeline to handle small packets
  - Automatically (in hardware) aggregate and disaggregate small packets

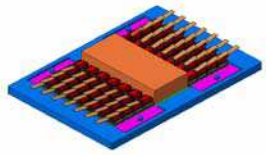
PERCS POWER7 Hierarchical Structure

- **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 Module: Hub Chip with Optics
- **POWER7 HPC Node**
  - 2U Node
  - 8 QCMs, 8 Hub Chip Modules
    - •256 Cores
- **POWER7 'Super Node'**
  - Multiple Nodes per 'Super Node'
  - Basic building block
- **Full System**
  - Multiple 'Super Nodes'

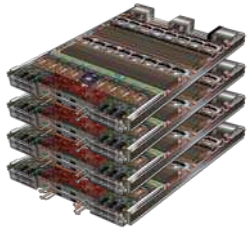


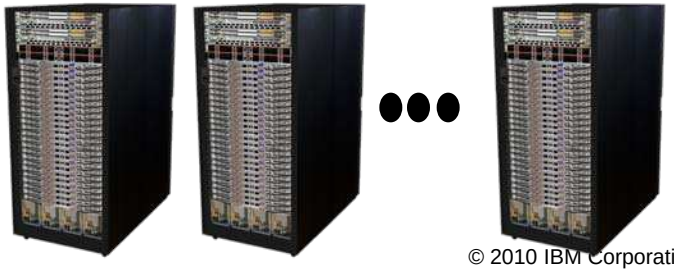


Hub Chip

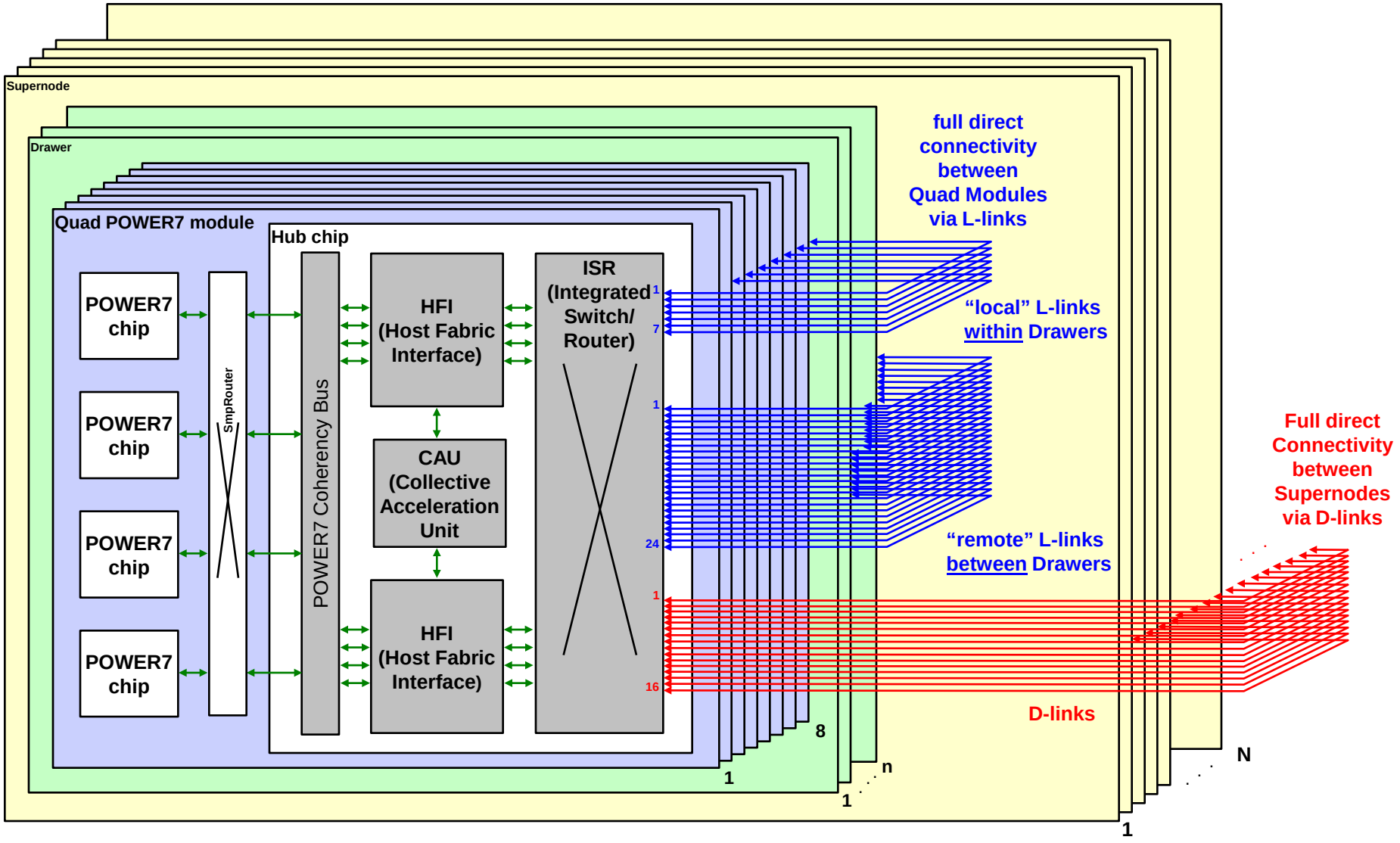


Hub Module with Optics





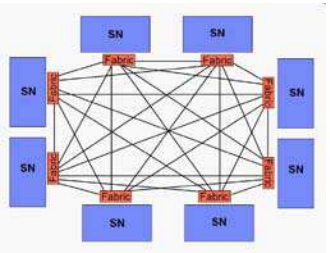
Logical View of PERCS Interconnect



POWER7 Hub Chip

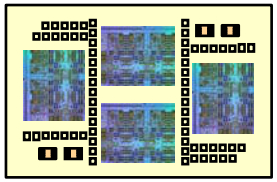
POWER7 Hub Chip Overview

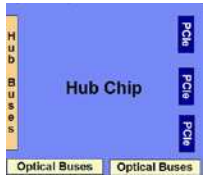
- Extends POWER7 capability for high performance cluster optimized systems
- Replaces external switching and routing functions in prior networks



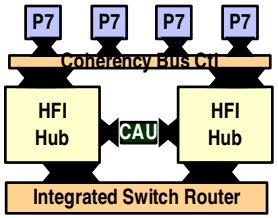
- Low diameter Two-tier Direct graph network topology is used to interconnect tens of thousands of POWER7 8-core processor chips to dramatically improve bi-section bandwidth and latency

- Highly Integrated Design
  - Integrated Switch/Router (ISR)
  - Integrated HCA (HFI)
  - Integrated MMU (NMMU)
  - Integrated PCIe channel controllers
  - Distributed function across the POWER7 and Hub chipset
  - Chip and optical interconnect on module
    - Enables maximum packaging density

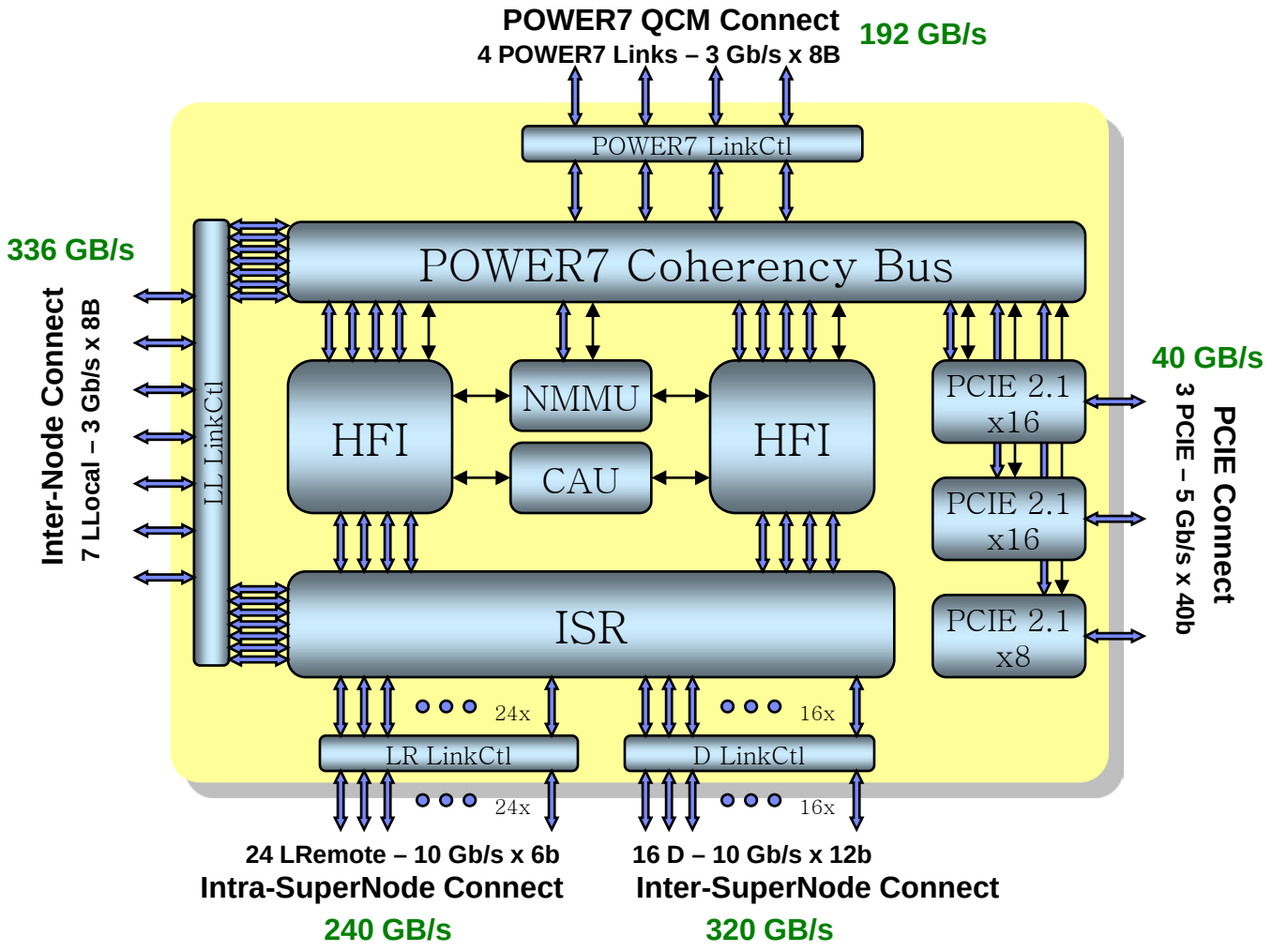




- Hardware Acceleration of key functions
  - Collective Acceleration
    - No CPU overhead at each intermediate stage of the spanning tree
  - Global Shared Memory
    - No CPU overhead for remote atomic updates
    - No CPU overhead at each intermediate stage for small packet disaggregation/aggregation
  - Virtual RDMA
    - No CPU overhead for address translation

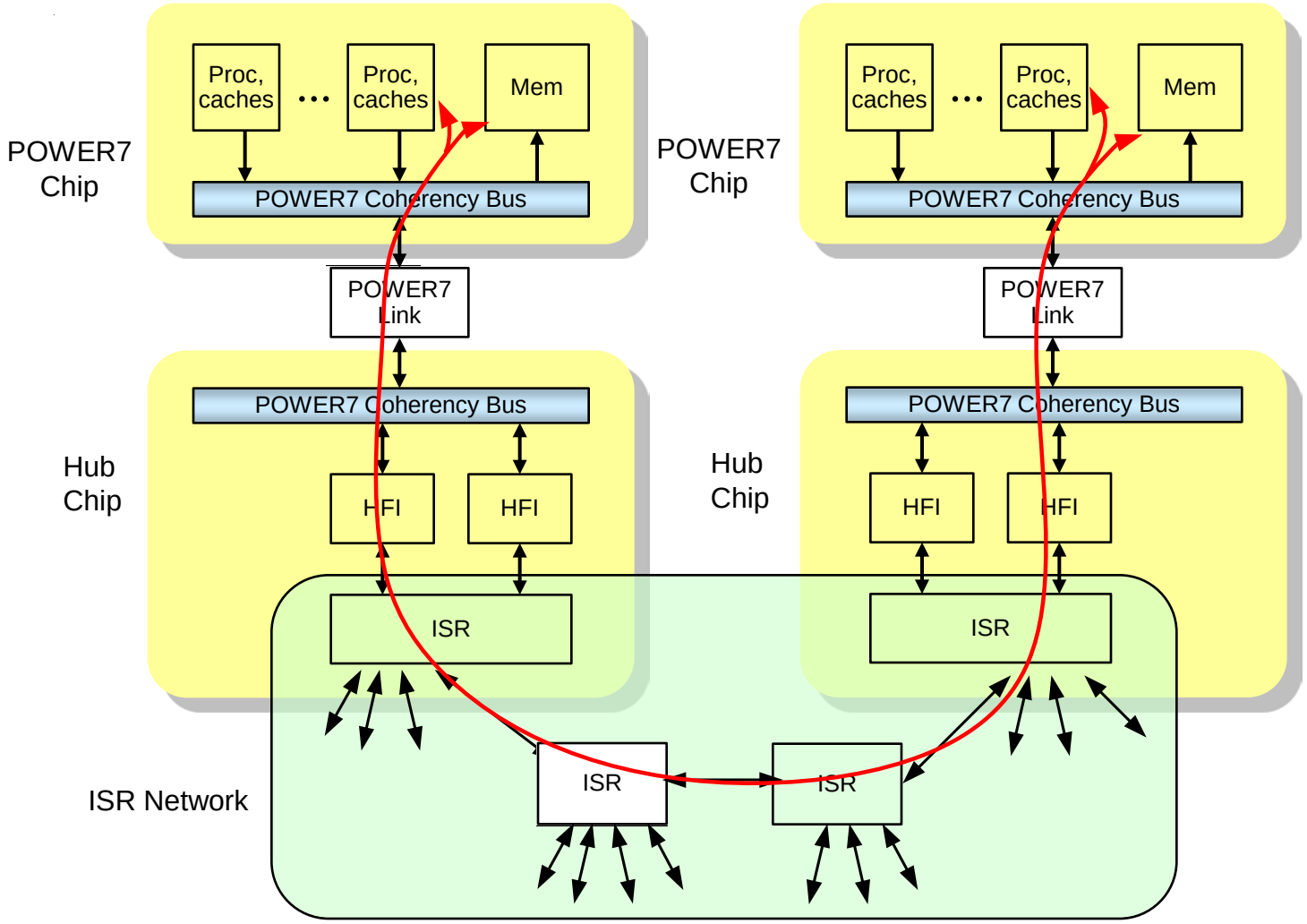


POWER7 Hub Chip Block Diagram



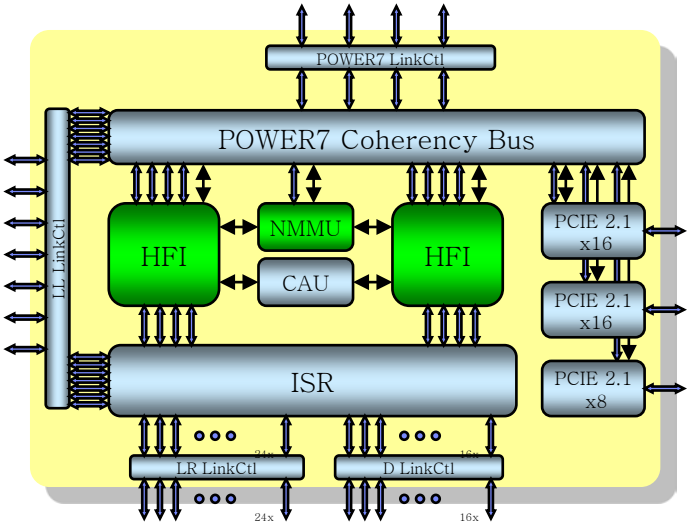
1.128 TB/s of off-chip interconnect bandwidth

End-to-End Packet Flow



Host Fabric Interface (HFI) Features

- **Non-coherent interface between the POWER7 QCM and the ISR**
  - Four ramps/ports from each HFI to the ISR
- **Communication controlled through “windows”**
  - Multiple supported per HFI
- **Address Translation provided by NMMU**
  - HFI provides EA, LPID, Key, Protection Domain
  - Multiple page sizes supported
- **POWER7 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)

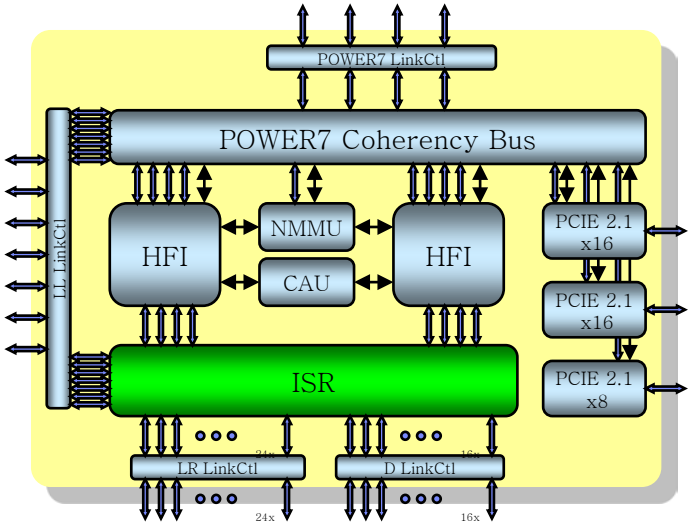
- **Supports three APIs**
  - Message Passing Interface (MPI)
  - Global Shared Memory (GSM)
    - Support for active messaging in HFI (and POWER7 Memory Controller)
  - Internet Protocol (IP)
- **Supports five primary packet formats**
  - Immediate Send
    - ICSWX instruction for low latency
  - FIFO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

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

- **GSM/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 receive/send FIFO)
    - Write, Read, Completion
    - Single packet per message
  - Small-RDMA (FIFO to memory)
    - Atomic updates
    - ADD, AND, OR, XOR, and Cmp & Swap with and without Data Fetch
  - Remote Atomic Update (FIFO to memory)
    - Multiple independent remote atomic updates
    - ADD, AND, OR, XOR
    - Hardware guaranteed reliability mode

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
- **Link Reliability**
  - CRC based link-level retry
  - Lane steering for failed links



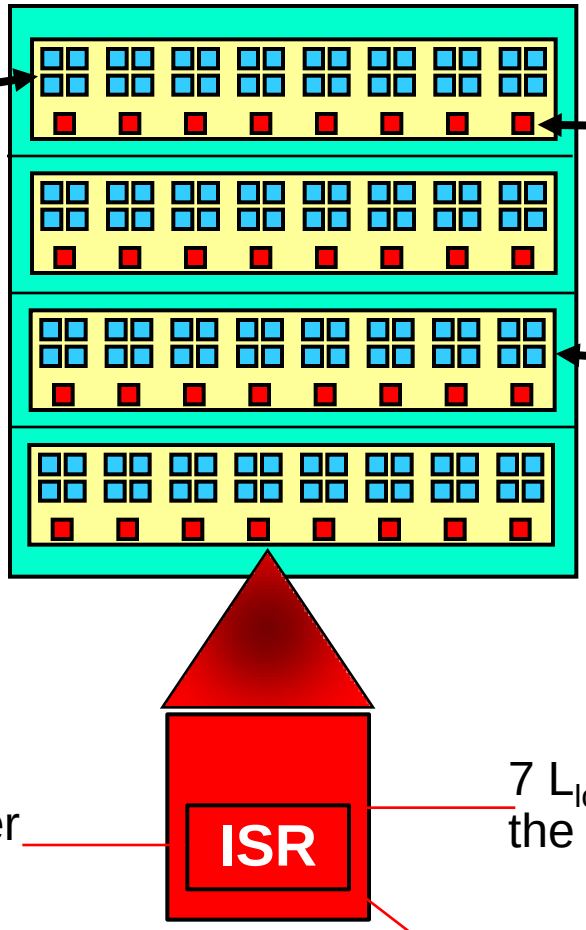
- **IP Multicast Support**
  - Multicast route tables per ISR for replicating and forwarding multicast packets
- **Global Counter Support**
  - ISR compensates for link latencies as counter information is propagated
  - HW synchronization with Network Management setup and maintenance

Integrated Switch Router (ISR) - Routing

- **Routing Characteristics**
  - 3-hop L-D-L longest direct route
  - 5-hop L-D-L-D-L longest indirect route
  - Cut-through Wormhole routing
  - Full hardware routing using distributed route tables across the ISRs
    - 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
  - FLITs of a packet arrive in order, packets of a message can arrive out of order

- **Routing Modes**
  - Hardware Single Direct Routing
  - Hardware Multiple Direct Routing
    - For less than full-up system where more than one direct path exists
  - Hardware Indirect Routing for data striping and failover
    - Round-Robin, Random
  - Software controlled indirect routing through hardware route tables

SuperNode



QCM: 4 chip POWER7 processor module

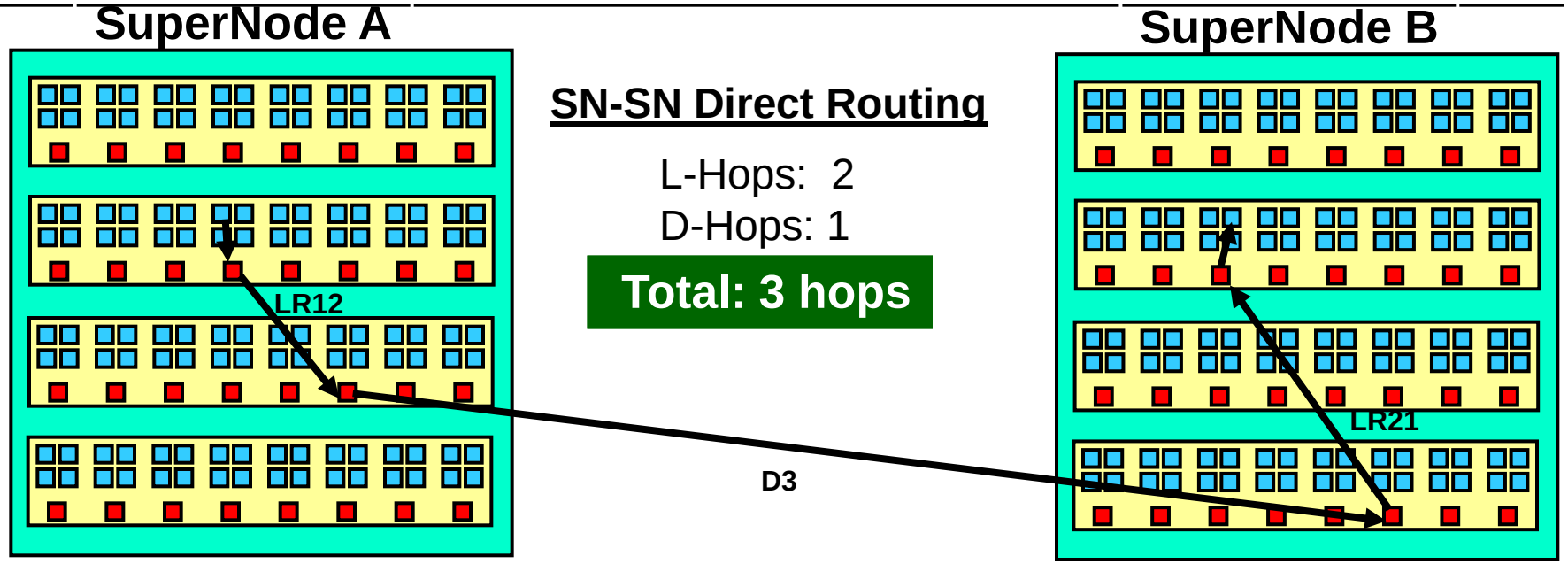
Hub chip with Integrated Switch Router (ISR)

POWER7 IH Drawer with 8 octants (QCM-Hub chip pairs)

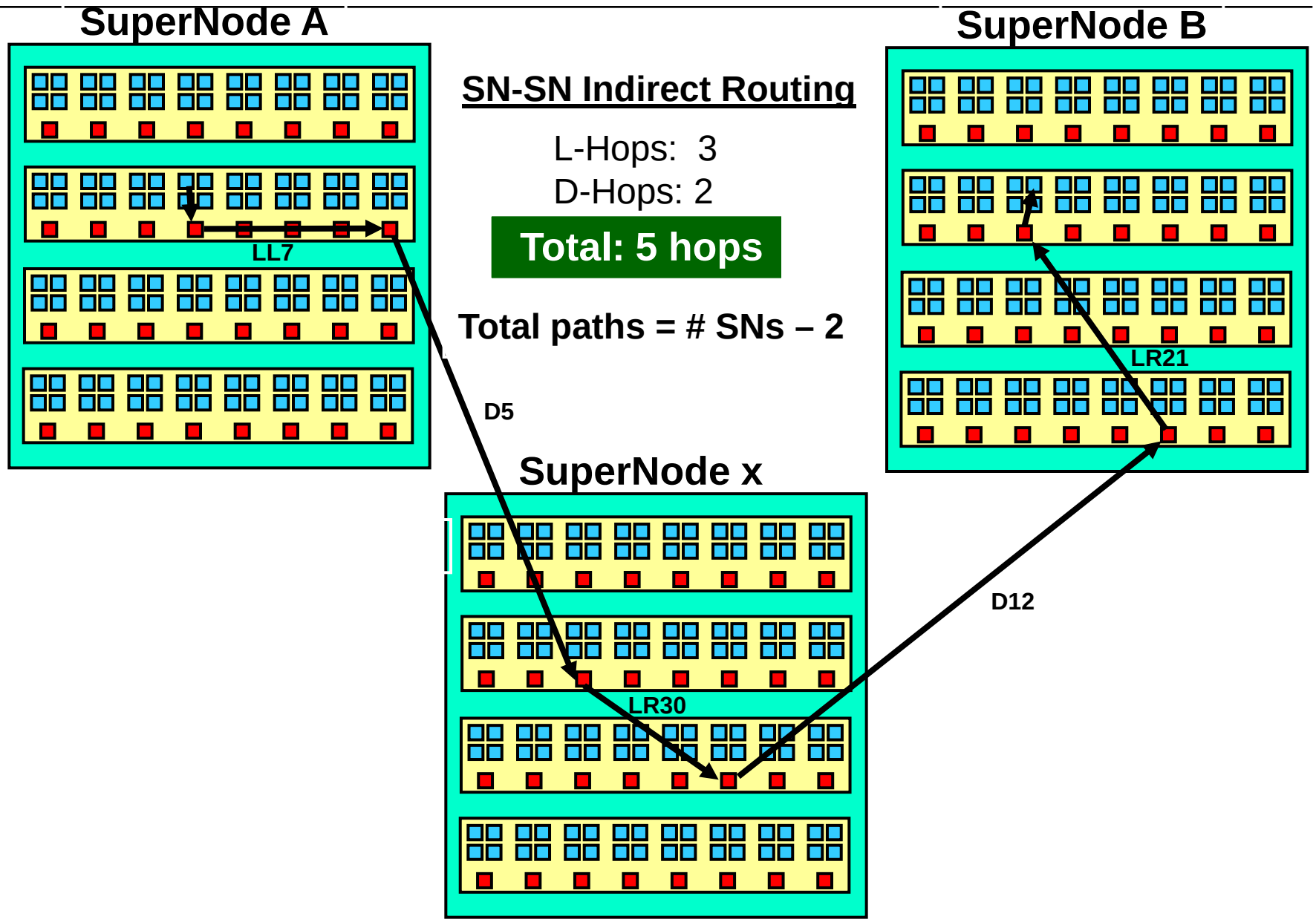
16 D links to other supernodes

7 L_{local} Links to Hubs within the drawer

24 L_{remote} Links to Hubs in other three drawers of the supernode

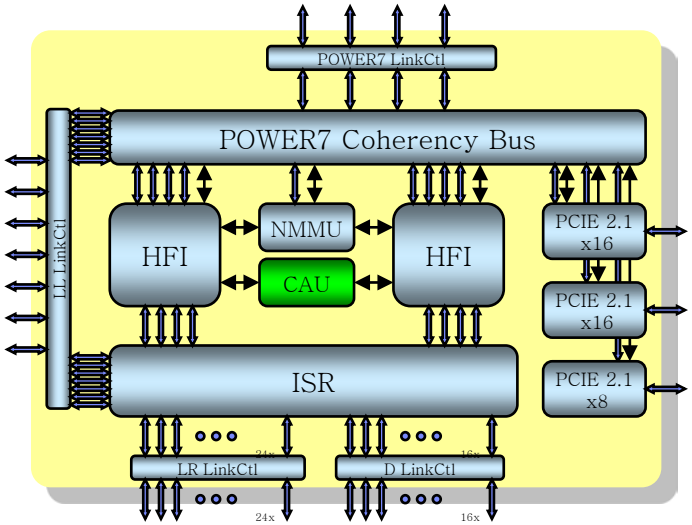


- **Max Bisection BW**
- **One to many paths depending on system size**



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
  - Each tree has one and only one participating HFI window on any involved node
  - It's up to the software to setup the topology



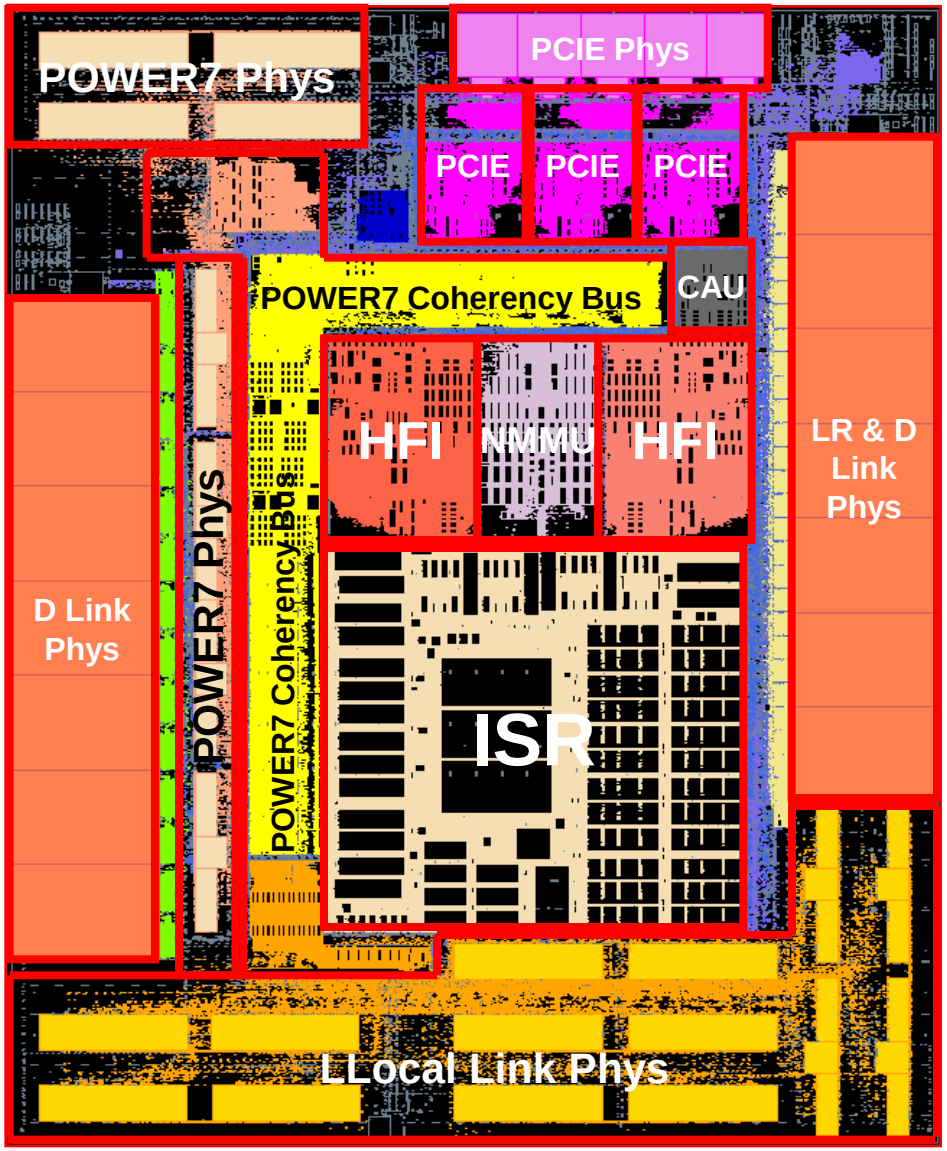
Collectives Acceleration Unit (CAU) Features (cont'd)

- **Sequence Numbers for Reliability and Pipelining**
  - Software driven retransmission protocol if credit return is delayed
  - The previous collective is saved on CAU for retransmission
    - Allows retransmit from point of failure vs. restart of entire operation

- **Reproducibility**
  - Binary trees for reproducibility
  - Wider trees for better performance

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



Key System Benefits Enabled by the POWER7 Hub Chip

- **A PetaScale System with Global Shared Memory**
- **Dramatic improvement in Performance and *Sustained* Performance**
  - Scale Out Application Performance (Sockets, MPI, GSM)
  - Ultra-low latency, enormous bandwidth and very low CPU utilization
- **Elimination of the Traditional Infrastructure**
  - HPC Network: No HCAs and External Switches, 50% less PHYs and cables of equivalent fat-tree structure with the same bisection bandwidth
  - Storage: No FCS HBAs, External Switches, Storage Controllers, DASD within the compute node
  - I/O: No External PCI-Express Controllers
- **Dramatic cost reduction**
  - Reduce the overall Bill of Material (BOM) costs in the System
- **A step function improvement in Data Center reliability**
  - Compared to commodity clusters with external storage controllers, routers/switches, etc.
- **Full virtualization of all hardware in the data center**
- **Robust end-to-end systems management**
- **Dramatic reduction in Data Center power compared to commodity clusters**

Outline

- **Background, Goals, and Design Point**
  - HPCS Program Background
  - Design Point and Topology
  - Hierarchical Structure and Interconnect

- **POWER7 Hub Chip**
  - Design Overview
  - Key Functional Units
  - Chip Floorplan / Metrics
  - Summary

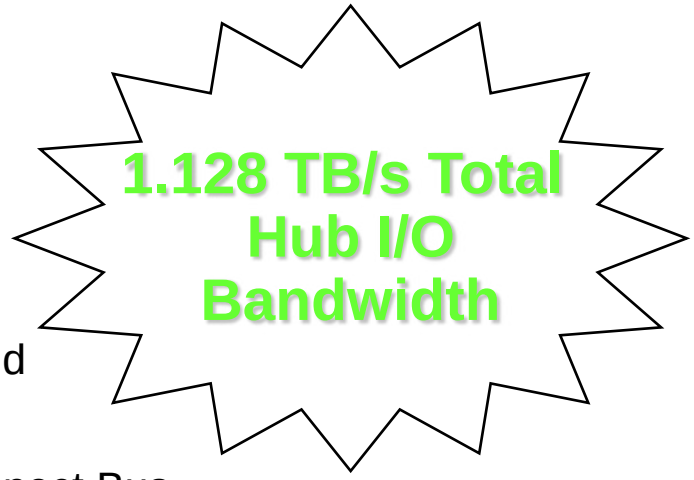
- **POWER7 Hub Module and Off-chip Interconnect**
  - Summary of I/Os and PLL's
  - 10Gb/s Physical Transport Circuit Architectures
  - Hardware Characterization
  - Summary

"This design represents a tremendous increase in the use of optics in systems, and a disruptive transition from datacom- to computercom-style optical interconnect technologies."

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Off-chip Interconnect and PLL's



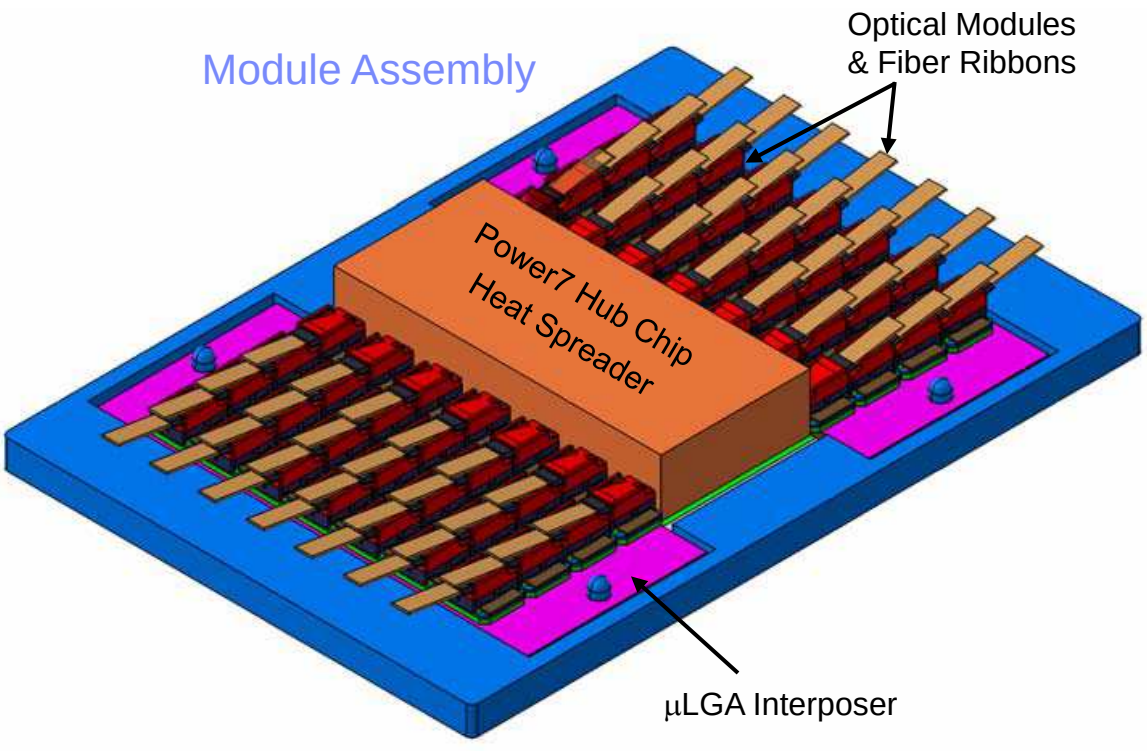
- (4) 8B W,X,Y,Z Interconnect Bus to POWER7
  - 3.0 Gb/s Single Ended EI-3 → **192 GB/s** throughput
- (7) 8B L-Local (LL) Interconnect Busses to Hub chips
  - 3.0 Gb/s Single Ended EI-3 → **336 GB/s** throughput
    - Shared physical transport for Cluster Interconnect and POWER7 Coherency Bus protocol
- (24) L-Remote: LR[0..23] within Drawer Hub Optical Interconnect Bus
  - 6b @ 10Gb/s Differential, 8/10 encoded → **240GB/s** throughput
- (16) D[0..15] between Drawer Hub Optical Interconnect Bus
  - 10b @ 10Gb/s Differential 8/10 encoded → **320 GB/s** throughput
- PCI General Purpose I/O → **40GB/s** throughput

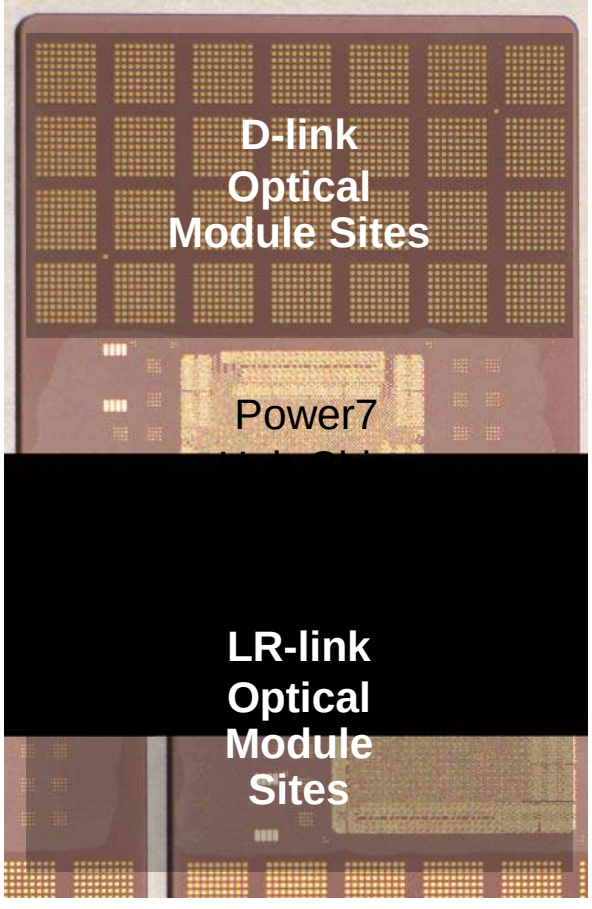
- 24 total PLL's
  - (3) "Core" PLLs
    - (1) Internal Logic • (1) W,X,Y,Z EI3 buses • (1) LL0 – LL6 EI3 buses
  - (2) "Intermediate Frequency "IF LC Tank" PLLs
    - (1) Optical LR buses • (1) Optical D buses
  - (14) High Frequency "HF LC Tank" PLLs
    - (6) Optical LR buses • (8) Optical D buses
  - (5) PCI-E "Combo PHY" PLLs

Hub Module Overview

Attribute	Definition
Technology	High Performance Glass Ceramic LGA
Body Size	61mm x 95.5mm
LGA Grid	Depopulated 58 x 89
Layer Count	90
Module BSM I/O	5139

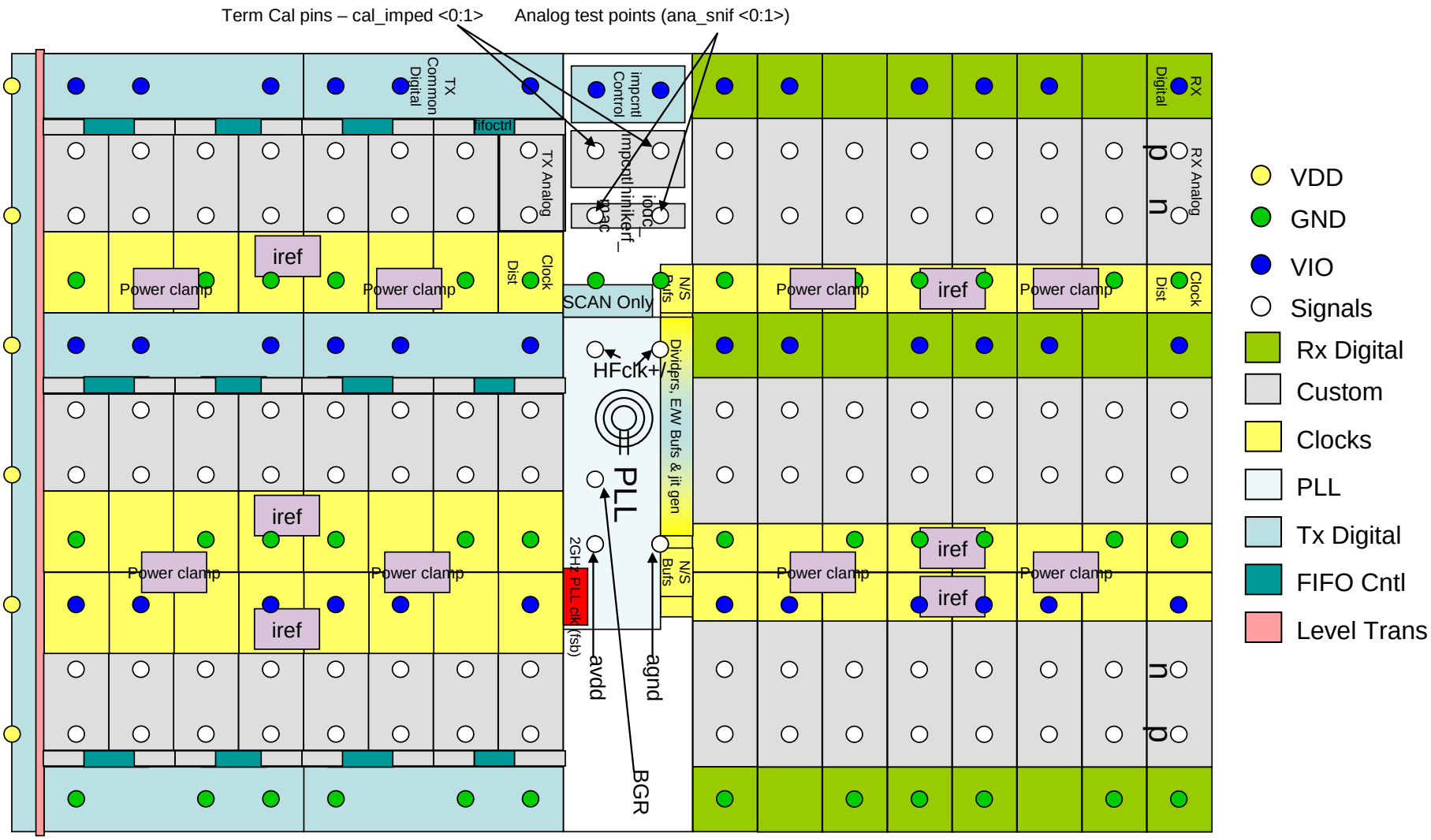
Module Assembly





10Gb/s D-link & LR-link PHY floorplan

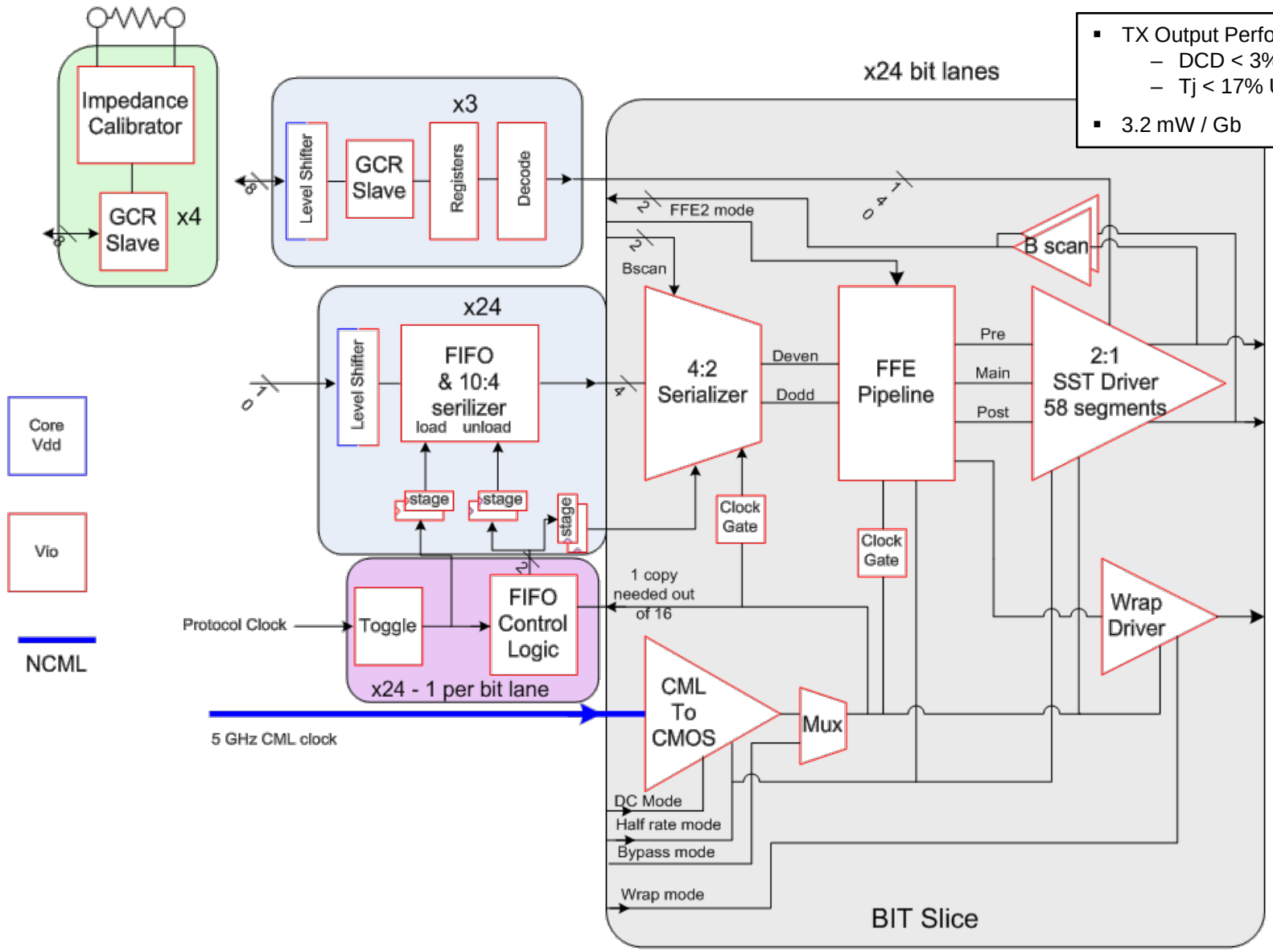
- 24 TX and 24 RX core sharing 1 HF PLL



10Gb/s D-link & LR-link PHY Transmitter Complex Features

- 10-to-1 serialization at 10.0 Gb/s.
- Driver Impedance calibration
  - Calibration at power-on determines the part specific number of SST driver segments to enable to achieve the desired source impedance
- On-chip T-coils are used for improved return loss
- Precursor / Post Cursor FFE levels with amplitude margining
  - 6 dB post cursor FFE & 3 dB precursor FFE capability
  - 6 dB margining capability
- FFE2 / FFE3 modes available
  - FFE2 mode has lower latency vs. FFE3
- Manufacturing test modes
  - Flush DC data to the c4 pads, boundary scan receiver & latches.
  - Provide clock to RX for test mode.
  - Individual SST segment test for fault coverage
  - On-die wrap test enabling at speed Built In Self Test
- Global Configuration Ring (GCR) support for transmitter configuration and impedance calibrator sensing

10Gb/s D-link & LR-link PHY Transmitter Block Diagram



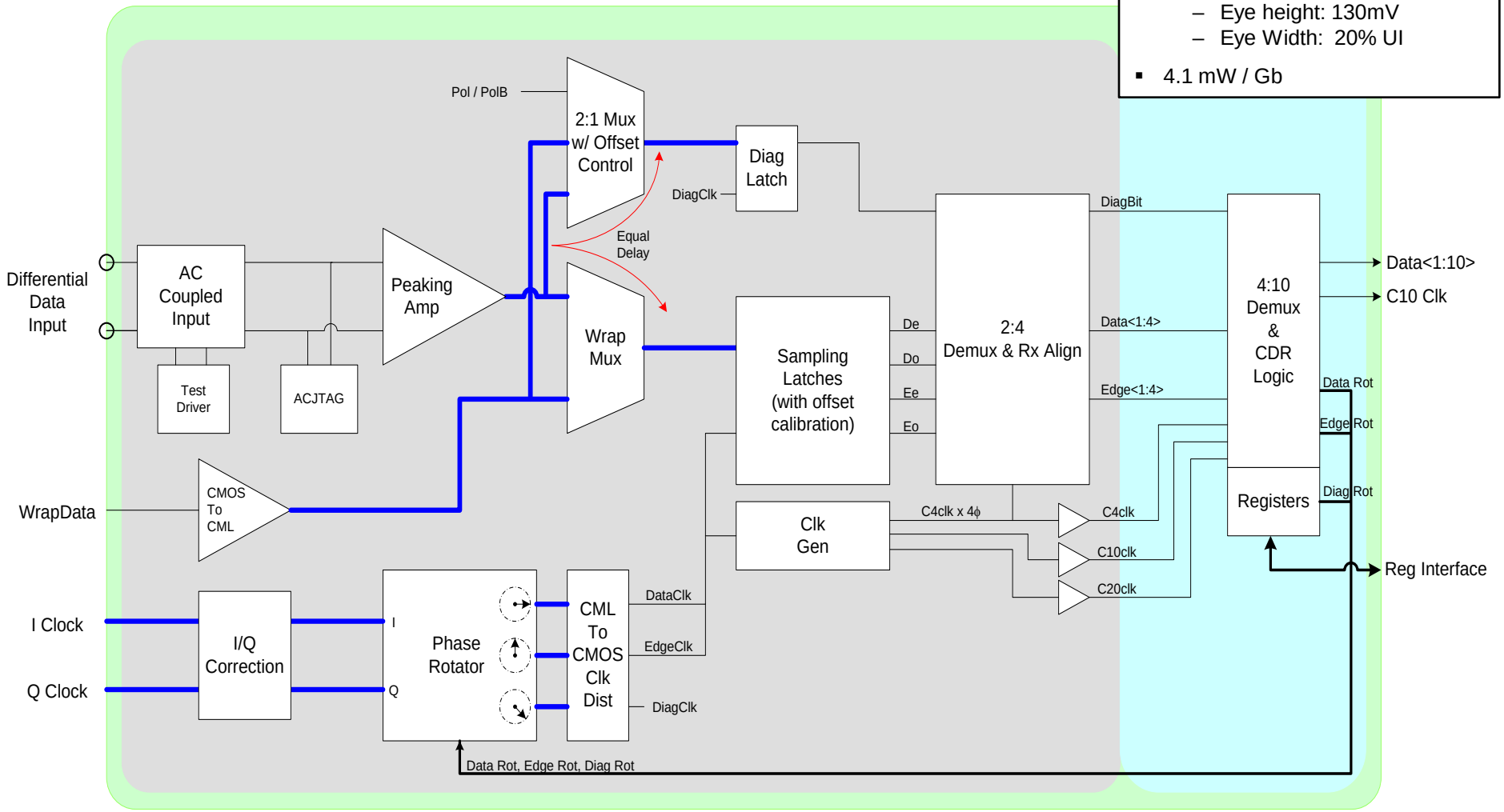
- TX Output Performance
  - DCD < 3% UI
  - Tj < 17% UI
- 3.2 mW / Gb

10Gb/s D-link & LR-link PHY Receiver Complex Features

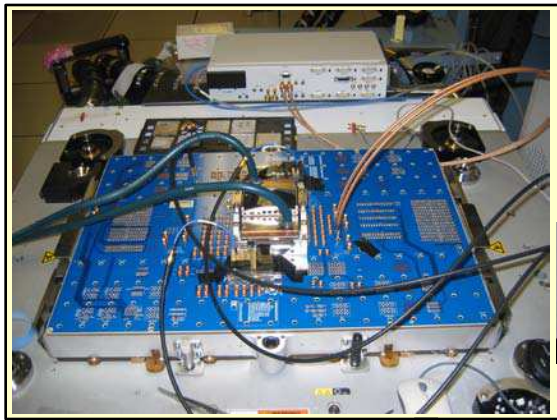
- 1-to-10 deserialization at 10.0 Gb/s
- CML Peaking Preamp, per lane offset calibration
  - Calibration performed at power-on to compensate for circuit offset caused by device mismatch
- On-chip T-coils for improved return loss
- I/Q Clock generation in the HF PLL
- Low power CML Phase Rotators
- DCVS sampling latches
- IEEE 1149.6 compliant JTAG test receiver
- Hardware Characterization modes:
  - Master Phase Rotator used to provide a controlled jitter injection to expand manufacturing test coverage
  - Diagnostic sampling latches to characterize margins on functional data
  - Per lane control of I/Q relationship for in-situ link margin testing
- Global Configuration Ring (GCR) support for receiver calibration, tune bit settings and test mode control

10Gb/s D-link & LR-link PHY Receiver Block Diagram

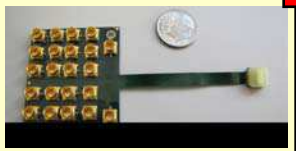
- RX Input Eye Mask Performance
  - Eye height: 130mV
  - Eye Width: 20% UI
- 4.1 mW / Gb



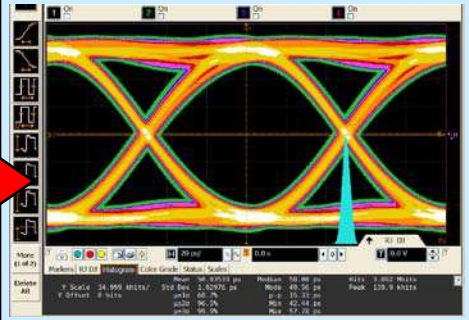
Hardware Characterization @ 10Gb/s

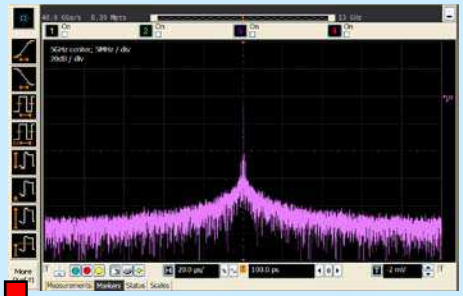


Designed Custom Characterization Hardware

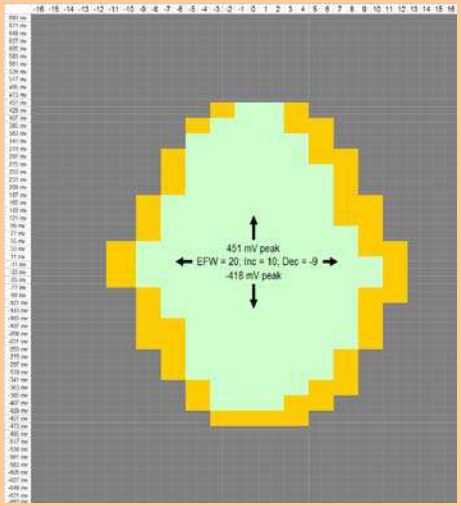


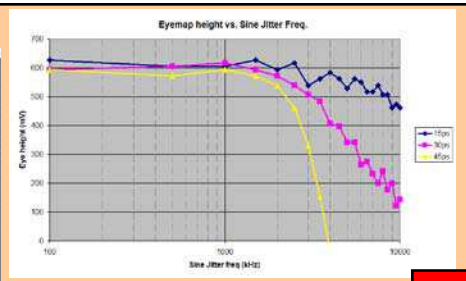
Perform Transmitter Characterization

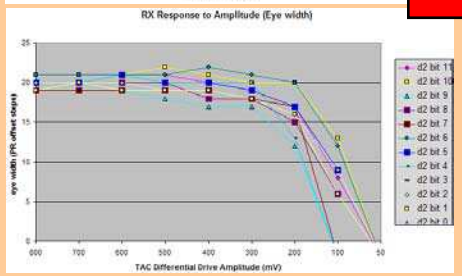




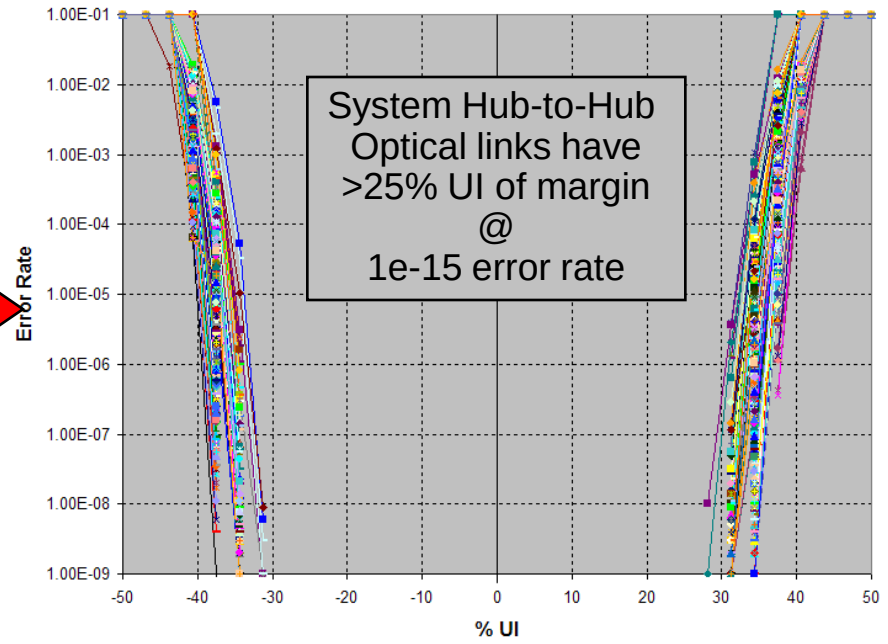
Perform Receiver Characterization







Perform Insitu End-to-end Link Characterization



Summary of Power7 Hub Chip Off-chip Interconnect

- **1.128 TB/s total Hub I/O Bandwidth achieved via Custom and Industry standard I/O**
  - 528 GB/s of Single-ended EI-3 at 3.0 Gb/s
  - 560 GB/s of Differential Optical Link PHYs at 10 Gb/s
- **Differential Optical Link PHYs achieved high performance at low power**
  - Transmitter achieved less than 17% UI Total Jitter @ 3.2mW / Gb
  - Receiver achieved greater than 80% UI Total Jitter Tolerance @ 4.1mW / Gb
- **End-to-end optical link performance in system operates with > 25% UI of margin at 1e-15 Error Rate**

Acknowledgements and References

- Authors
  - Scott Clark, Baba Arimilli, Ben Drerup, Jerry Lewis, John Irish, David Krolak, Kerry Imming, Joe McDonald, Andreas Koenig, Daniel Dreps, David Siljenberg, Steve Baumgartner, Glen Wiedemeier, Jim Strom, Dan OConnor, Andrew Maki, Dhaval Sejpal, Mark Ritter, Dave Friend and Charlie Geer

- References
  - HOT Chips 19: *The 3rd generation of IBM's Elastic Interface (EI-3) Implementation of POWER6*

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0002.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.