SHASTA HARDWARE WORKSHOP

Bob Alverson, Cray
Wade Doll, Cray

May 6, 2019

bob@cray.com
wdoll@cray.com

AGENDA

- Shasta hardware overview
- Shasta infrastructure
- Shasta interconnect packaging
- Shasta liquid cooling deep dive
- Slingshot network
- Q&A

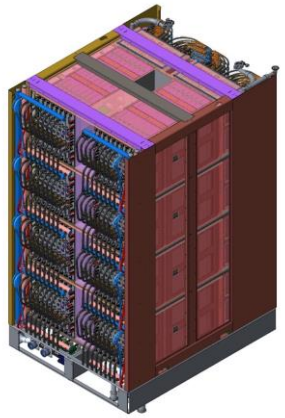
SHASTA HARDWARE WORKSHOP

- The purpose of this workshop is to expose users to the new Shasta hardware infrastructure and Slingshot network
- With a focus on the flexibility, advances, and technical capabilities of the infrastructure and network
- This deeper understanding will leave the user with the knowledge of differentiated features and superior hardware infrastructure

SHASTA HARDWARE OVERVIEW

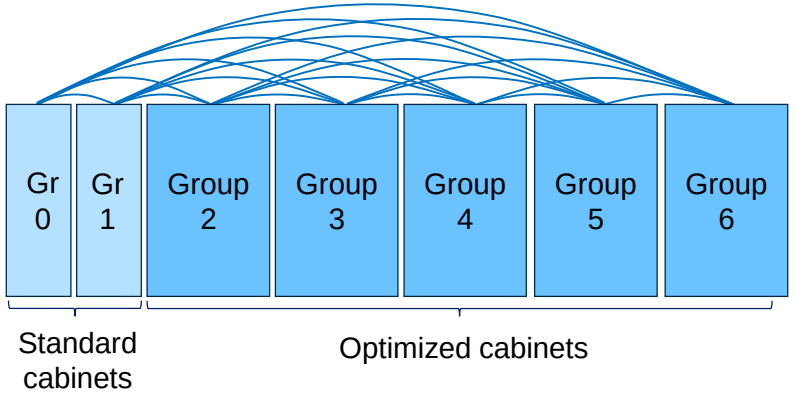
CRAY SHASTA SYSTEM

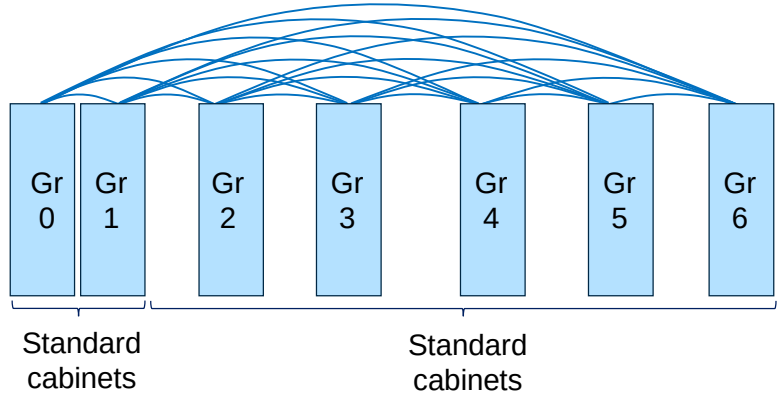
- Support diversity of processors
- Scale-optimized cabinets for density, cooling, and high network bandwidth
- Standard cabinets for flexibility
- Cray SW stack with improved modularity
- Unified by a single, high-performance interconnect



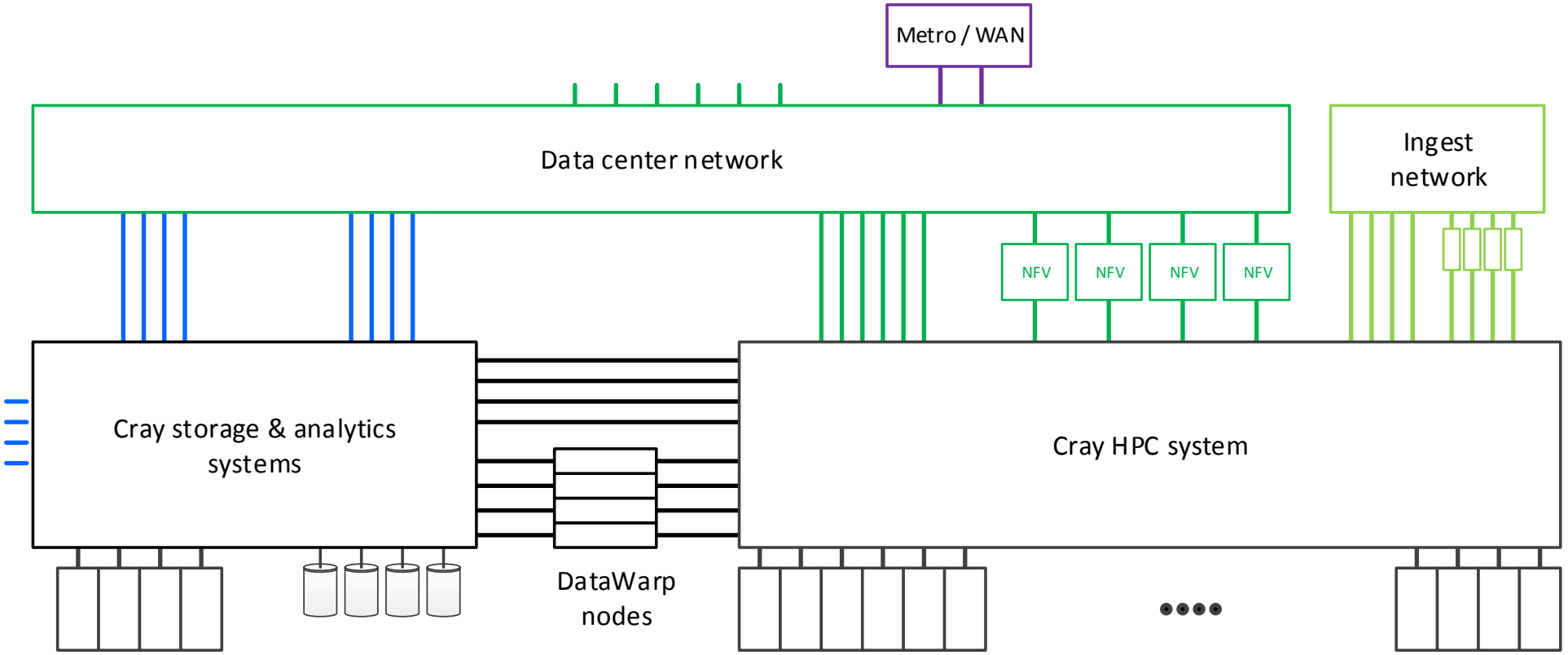
Optimized rack "Mountain"

Standard rack "River"





CRAY CONVERGED SYSTEM



SHASTA NETWORK OVERVIEW

- 3 Hop Dragonfly with switches integrated into same cabinets as nodes
- Host and Local links passive electrical cables
- Global links active optical cables
  - Flexible global bandwidth for network by varying amount of cables installed
  - Typically about 15% of links optical
- Warm swap switch
  - Quiesce not needed

SHASTA COMPUTE OVERVIEW

- High density Cray blades in Mountain cabinets
  - 64 compute blades per cabinet
  - 2/4/8/16 NICs per blade
- Flexible support for arbitrary nodes in River cabinets
- Hot swap compute blades
  - Separate boards for NIC, Switch

Mountain

River

SHASTA INFRASTRUCTURE

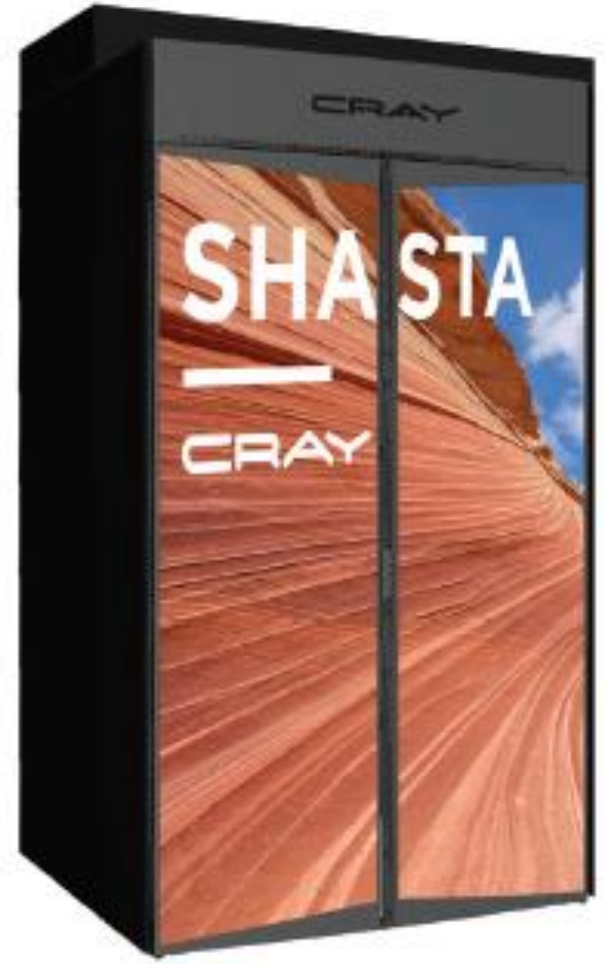
Wade Doll

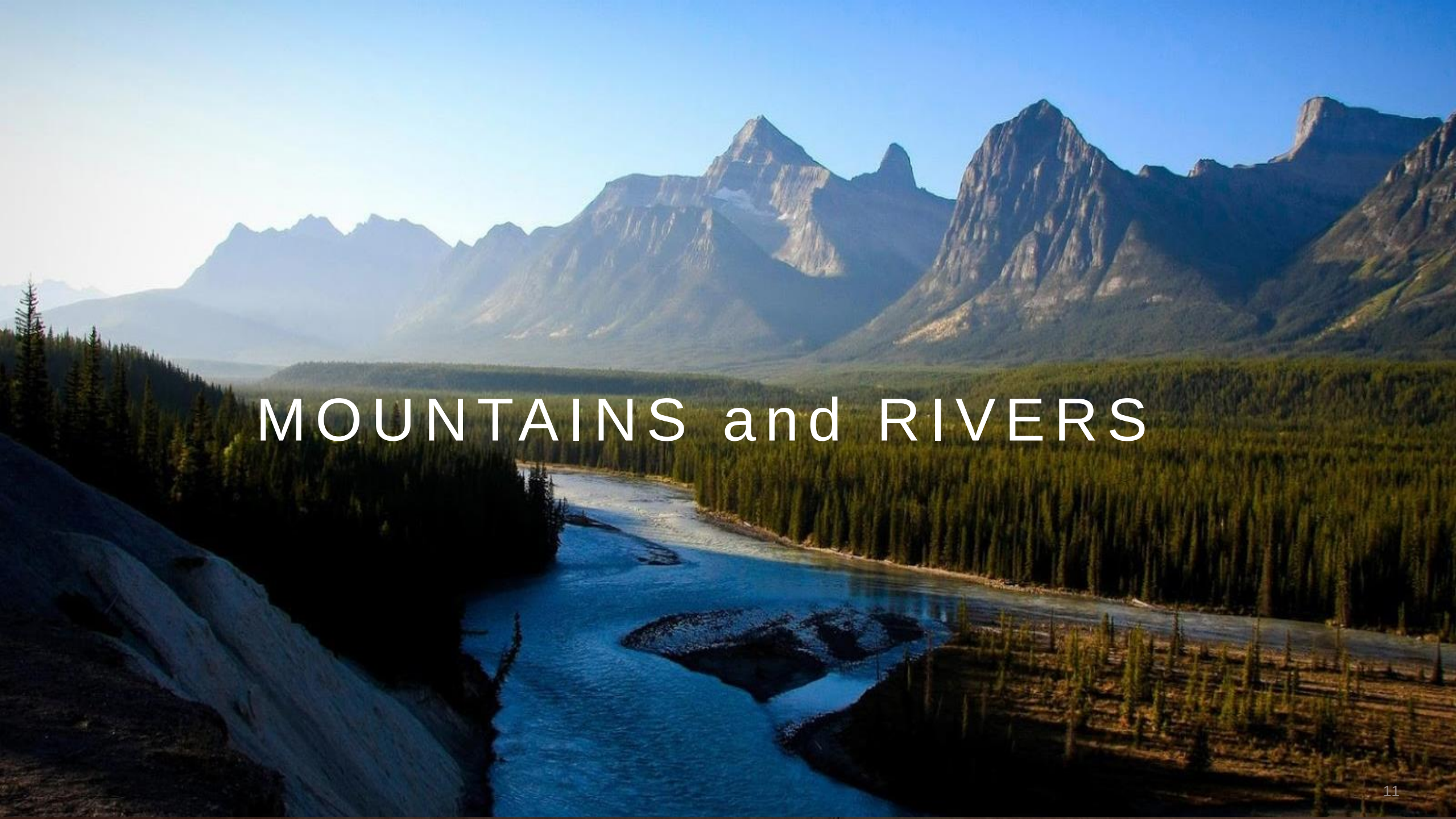
Principal Infrastructure Architect

Cray Inc.

KEY SHASTA INFRASTRUCTURE ATTRIBUTES

Performance	Dense solution with high performance processors Supports highest TDP processors Single cabinet to Exaflops
Upgradeability	Infrastructure designed to last a decade Power & cooling headroom for future processors Multiple interconnect generations
Flexibility	Flexible cabinets options Wide choice of node types Configurable network bandwidth
Green	Direct liquid-cooling, up to ASHRAE W4 water Efficient, high voltage DC internal distribution Fine-granularity power management




MOUNTAINS and RIVERS

SHASTA COMPUTE INFRASTRUCTURES

MOUNTAIN
Optimized Cabinet

RIVER
Standard 19” Rack

Same Interconnect - Same Software Environment

SHASTA MOUNTAIN CABINET

- 100% direct liquid cooling
- 64 compute blades with 4-16 sockets/blade
- 512+ high-performance processors
- Lowest TCO/up to 10 years w/o forklift upgrade
- Scales to >100 cabinets
- 300kVA with warm water cooling
- Flexible, high-density interconnect

SHASTA MOUNTAIN ADVANTAGE

Performance

- **Highest power CPUs (500W+) supported via direct liquid cooling**
- **Up to 16 Slingshot injection ports per compute blade**
- **Hardware & Software scalable to Exascale class systems**

TCO

- **Warm water cooling (ASHRAE W3 and W4 temps supported)**
- **Efficient power conversion from mains to point-of-load**
- **Upgradeable for multiple technology generations**

SHASTA RIVER RACK

- Standard 19” rack
- Front to back cooling
- Air cooled with liquid cooling options
- Fits into every data center
- 42U/48U, 60cm/75cm options
- Scales to >100 racks
- Wide range of available compute and storage

SHASTA RIVER RACK ADVANTAGES

Standards-Based

- **Standard 19” Rack**
- **Standard front to back air cooling**
- **Standard datacenter integration**
- **Wide selection of compute nodes**
- **Flexible infrastructure**

Supercomputer

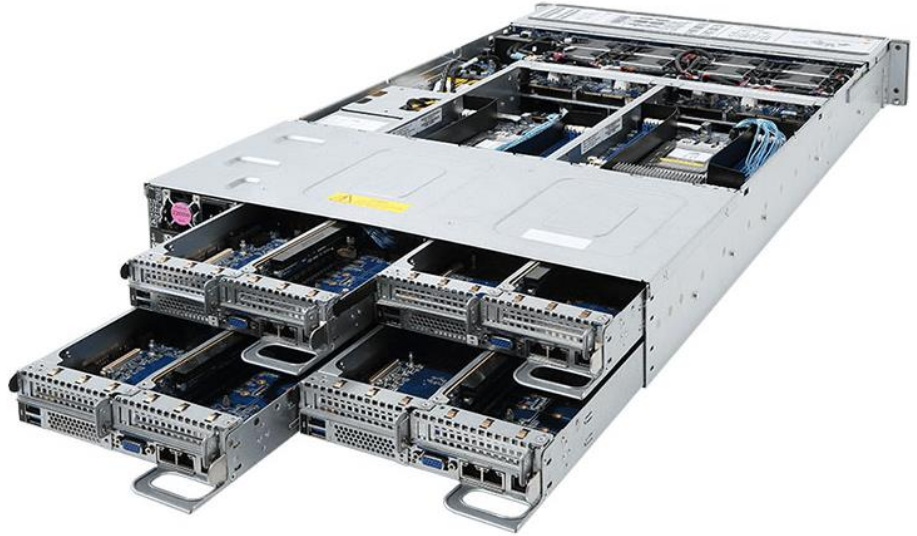
THE SUPERCOMPUTER COMPANY

- **Same powerful Cray interconnect fabric**
- **Same Cray Linux-based SW Stack**
- **Same Cray Programming Environment**
- **Scalable from 1 to 100’s of racks**

SHASTA COMPUTE NODES

MOUNTAIN
Customized Blades

RIVER
Standard Server Chassis



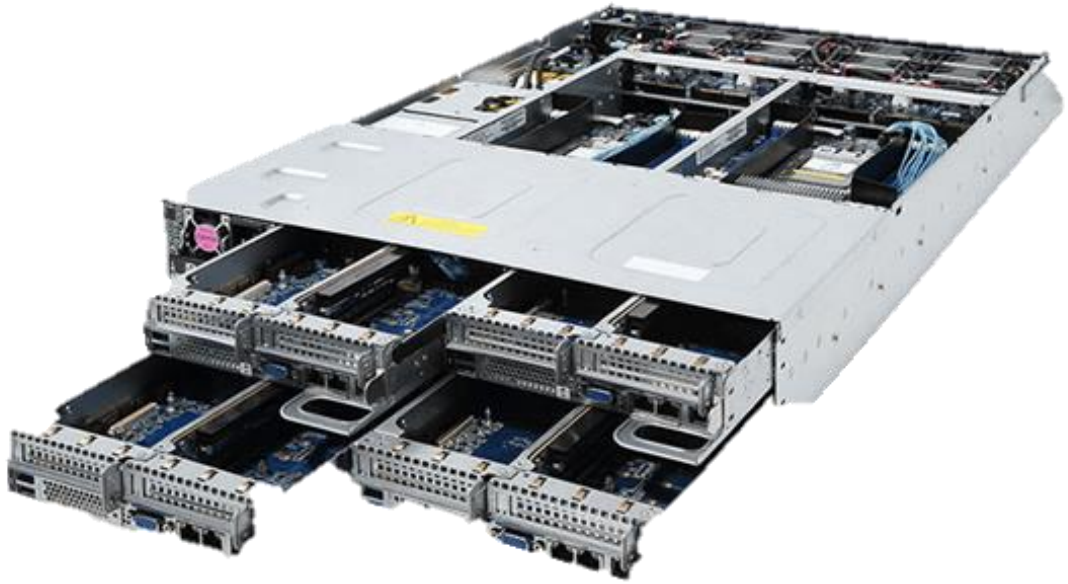
Same Compute Functionality

SHASTA MOUNTAIN BLADES

- High density packaging
- Supports highest TDP SKUs
- Supports highest speed DDR
- ~2x density of standard ½ SSI boards
- Board construction for reliability, signal integrity and power integrity performance
- Fully liquid cooled
- High voltage dc power architecture
- Network NIC and Hardware Management
- 1U x 91cm x 43cm
- Warm swappable

SHASTA RIVER CHASSIS

- Any standard 19” rack mount chassis
- Ex: 2U/4Node compute, 4 CPU/U
- Air cooled
- Standard 120Vac/240Vac power architecture
- PCIe card host adapter



SHASTA COOLING TECHNOLOGIES

MOUNTAIN
Full Liquid Cooling

RIVER
Air or Hybrid Cooling

Different Technologies & Capabilities

SHASTA MOUNTAIN COOLING TECHNOLOGY

- Direct contact liquid cold plates
- CPU, Memory, Voltage regulators…etc.
- Flexible and rigid tubing
- Serviceable cold plates
- Dry break quick disconnects
- Capable of cooling 500W+ processors
- Support for warm water cooling (ASHRAE W3, W4)
- Copper and stainless steel fluid contact
- In blade leak detection

Liquid Cooling deep dive later in the presentation

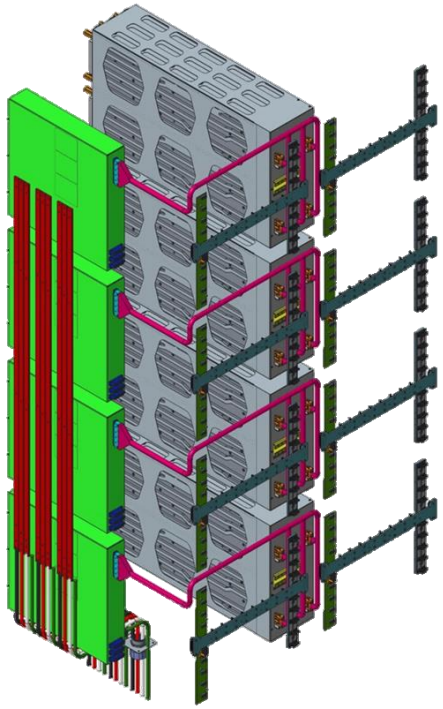
SHASTA RIVER COOLING TECHNOLOGY

- Standard front to back air cooling
- Support for warm air cooling (ASHRAE A2)
- Room neutral rear door heat exchanger option
- Limited direct liquid cooling options

SHASTA POWER ARCHITECTURE

MOUNTAIN
High Voltage Distribution

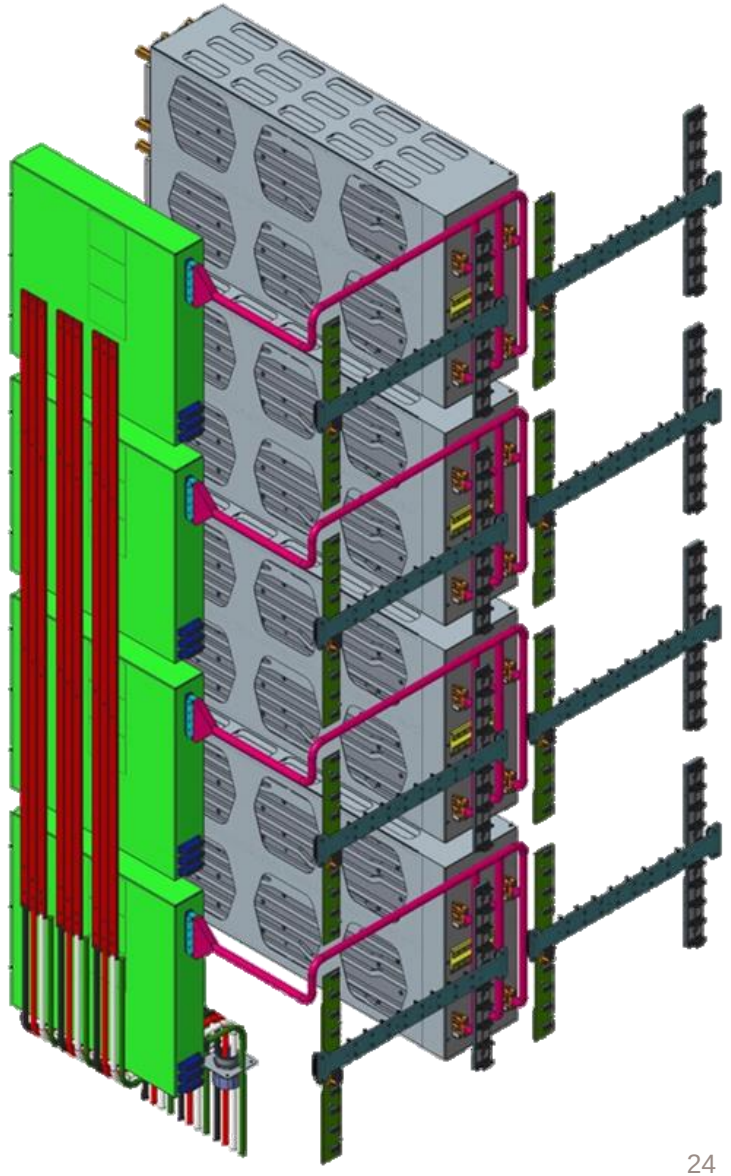
RIVER
Flexible Voltage Distribution



Different Capabilities

SHASTA MOUNTAIN POWER ARCHITECTURE

- High voltage 3-phase rack feeds (400Vac/480Vac) to minimize whips
- 80A/100A/120A cabinet ratings options to minimize stranded power
- Redundant, hot swap ac/dc rectifiers
- High voltage dc rack distribution to minimize losses
- Integrated PDU with fuses, breakers, and filters
- Supports cabinet powers of 300kVA



SHASTA RIVER POWER ARCHITECTURE

- Standard 3-phase ac PDU options
- 208Vac/400Vac/415Vac
- Intelligent PDU
- A or A/B redundant feed options
- 1-4 PDU per rack

SHASTA DATA CENTER INTEGRATION

MOUNTAIN
Modern Datacenter Requirements

RIVER
Fits in Any Datacenter

Different Facility Requirements

SHASTA MOUNTAIN CABINET

- High voltage 400Vac/480Vac feeds (wye or delta)
- Larger than typical footprint
- 48U cabinet
- High floor loading
- Liquid cooling only
- No casters

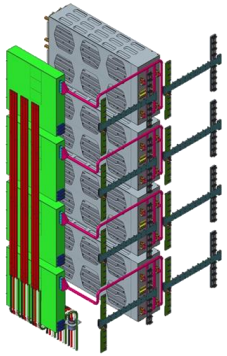
SHASTA RIVER RACK

- Standard 3 phase 208Vac/400Vac power feeds (wye or delta)
- Standard 42U 19” rack
- Front to back air cooling
- Castered rack

SHASTA TCO ENABLEMENT

MOUNTAIN
TCO Optimized

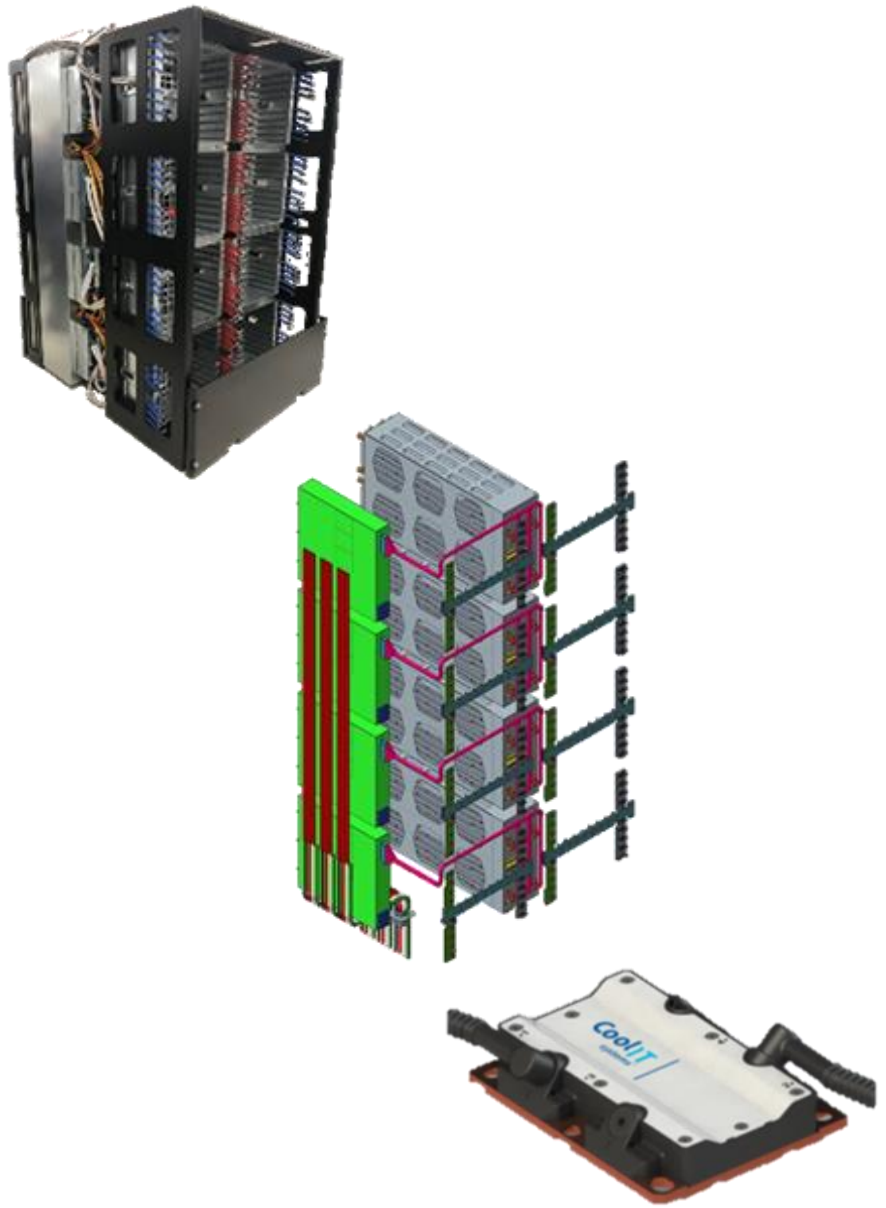
RIVER
Standards based



Different Technologies and Capabilities

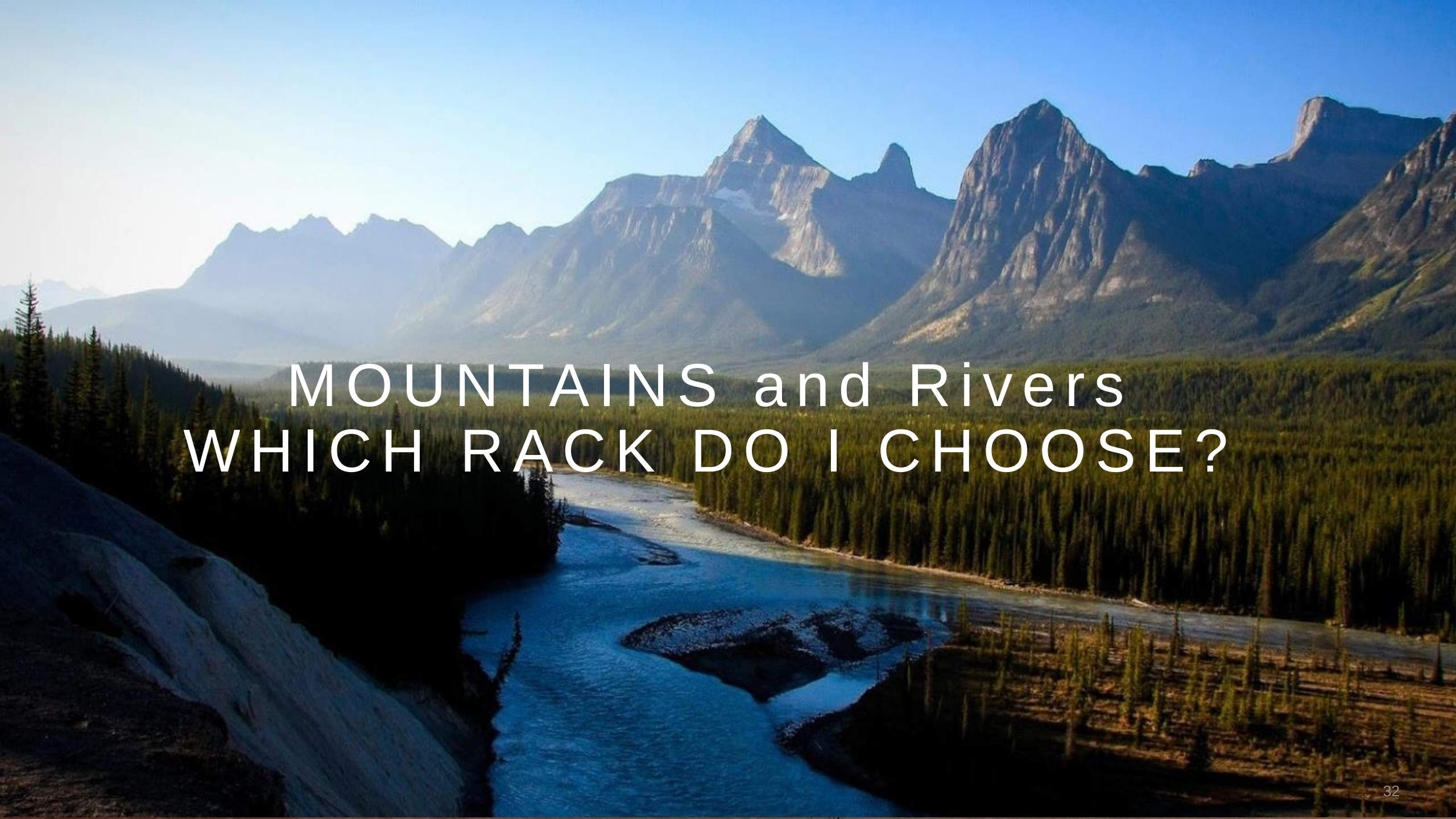
SHASTA MOUNTAIN CABINET

- High density packaging to reduce data center footprint
- Warm water cooling to support chiller-less cooling
- No heat transferred into the data center
- High voltage power architecture to reduce power losses
- Architecture flexibility designed for 10yr life



SHASTA RIVER RACK

- Driven by standards technology
- Flexibility allows integration into all datacenters
- No expensive facility upgrades

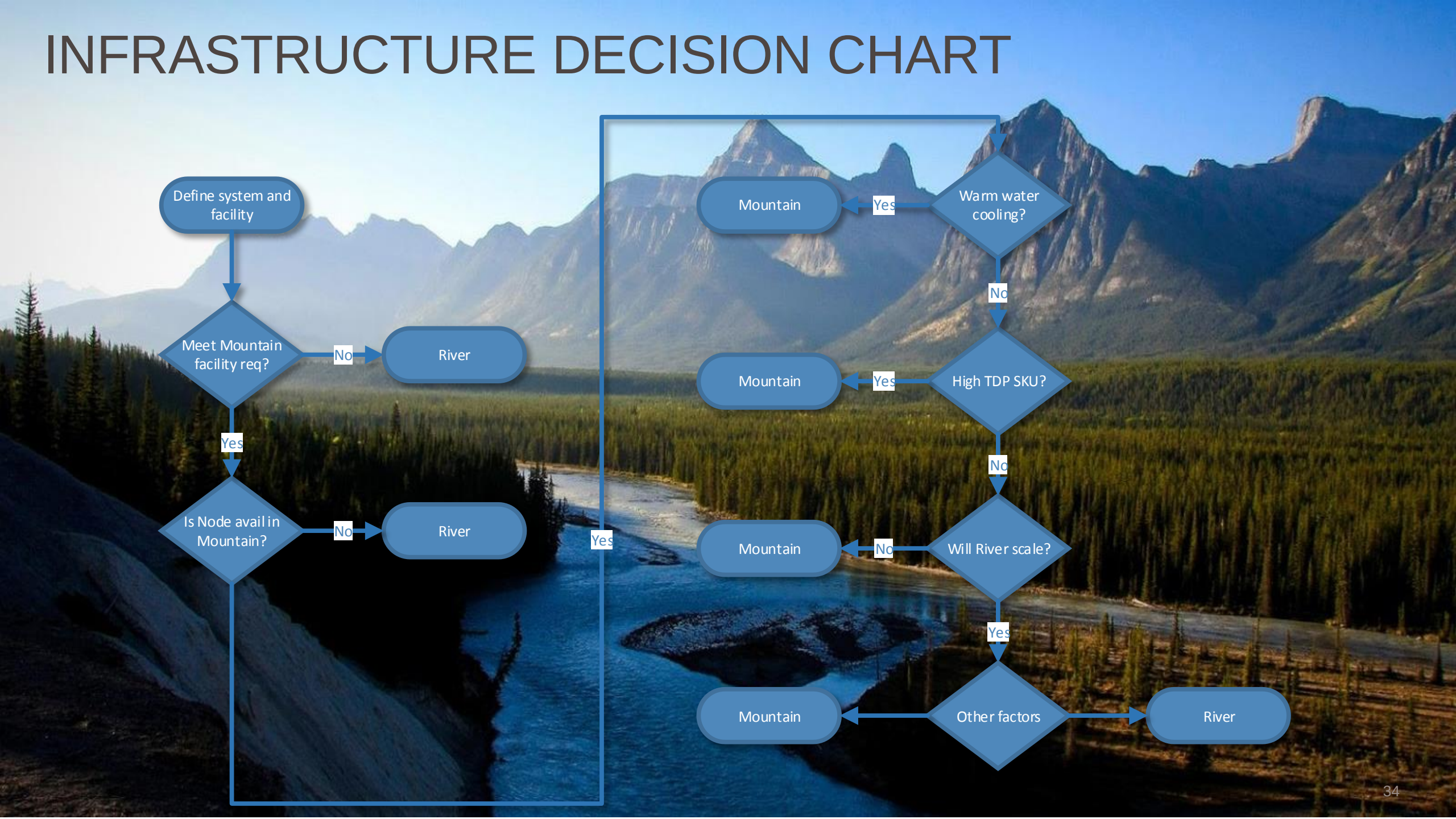

MOUNTAINS and Rivers
WHICH RACK DO I CHOOSE?

WHICH INFRASTRUCTURE IS RIGHT FOR YOU?

MOUNTAIN	RIVER
❑ *Specific data center requirements*	❑ *Standard data center requirements*
❑ Supports highest power processors	❑ Supports standard power processors
❑ Supports warm water cooling (no air cooling support)	❑ Supports typical room air cooling
❑ Supports lowest TCO, PUE, and iTUE	❑ Supports typical TCO, PUE, and iTUE
❑ Scales to 500 cabinets (250k endpoints)	❑ Scales to 162 racks (10k endpoints)
❑ Support for Slingshot network	❑ Support for Slingshot network

- Generally, if the facility can support the requirements, and the node is available, Mountain is an overall better choice

INFRASTRUCTURE DECISION CHART



SHASTA INTERCONNECT PACKAGING

Wade Doll,

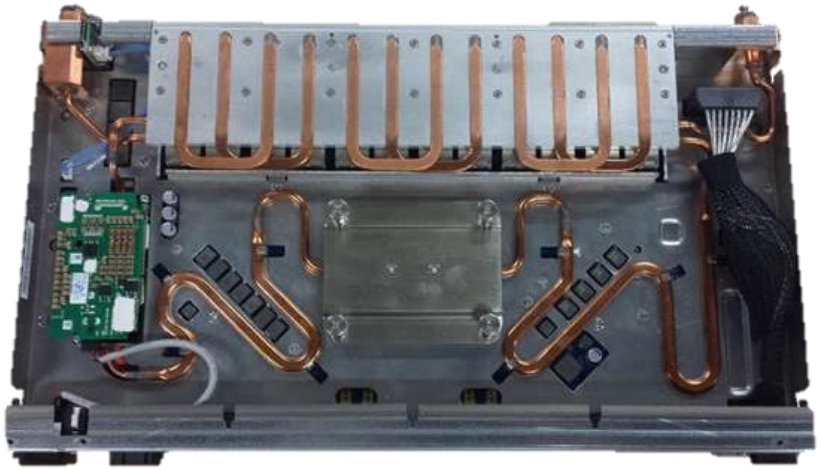
Principal Infrastructure Architect

Cray Inc.

SLINGSHOT SWITCH PACKAGING

MOUNTAIN
Custom

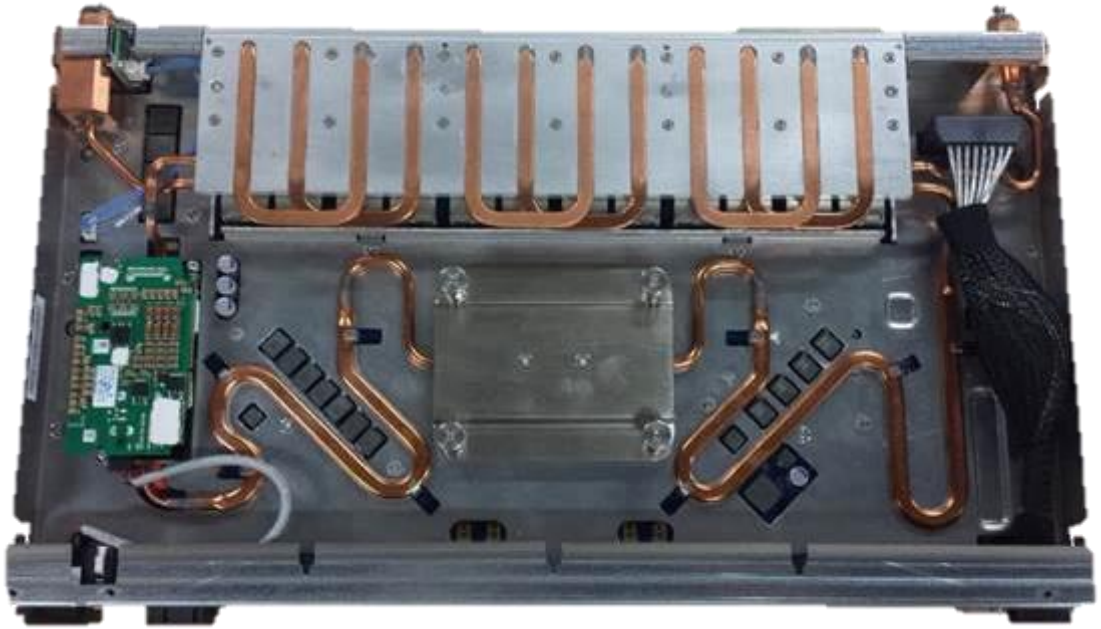
RIVER
Industry Standard



Same Functionality

SLINGSHOT MOUNTAIN SWITCH

- 64 port Rosetta switch
- Fully liquid cooled
- 16 Downlinks to compute blades/nodes
- 48 Ports for L1 and L2 Dragonfly exposed as 24 QSFP-DD cables
- Custom form factor for Mountain chassis
- High voltage dc power architecture



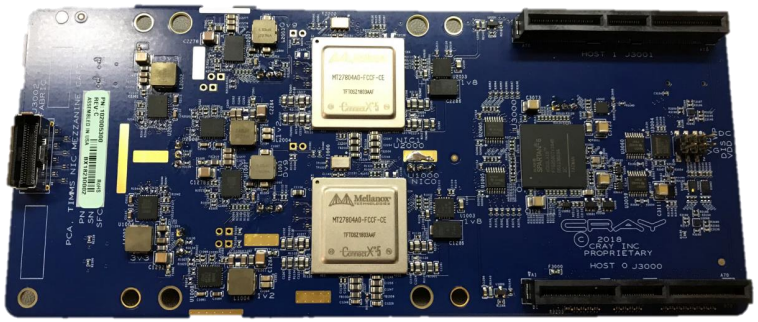
SLINGSHOT RIVER SWITCH

- 1U form factor
- 64 port Rosetta switch
- Air cooled (front-to-back or back-to-front)
- Hot swap fans
- 64 links exposed with 32 QSFP-DD cables
- Support for electrical or AOC cables
- Standard 19” rail mounted
- Universal 240Vac support

SLINGSHOT NIC PACKAGING

MOUNTAIN
Custom Form Factor

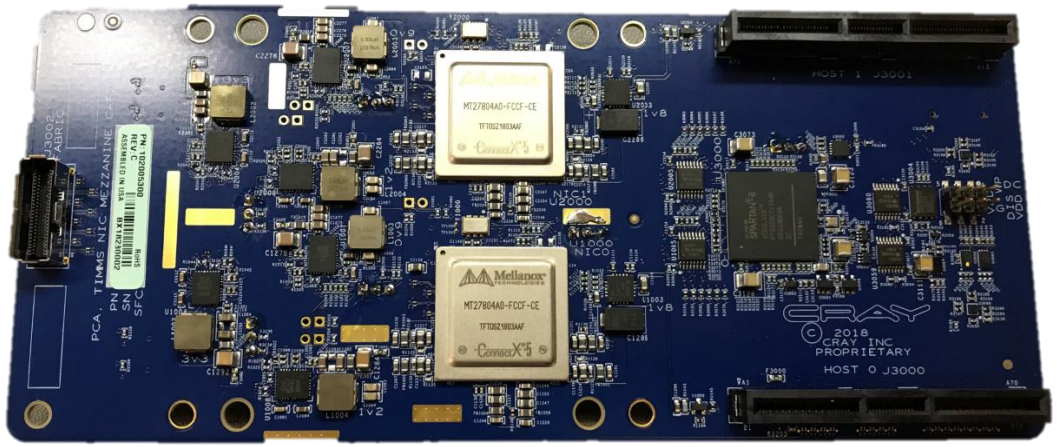
RIVER
Industry Standard Form Factor



Same Functionality

SLINGSHOT MOUNTAIN NIC MEZZANINE

- Custom form factor for dense Mountain blade
- Approximately 84mm x 190mm
- Liquid cooled
- Can be packaged on top of processor heat sink
- Two NIC ports per mezzanine card
- Two PCIe x16 ports from nodes
- Two Slingshot x4 ports cabled to switch



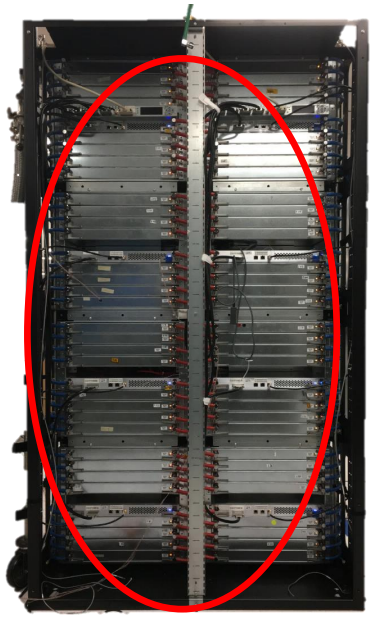
SHASTA RIVER NIC ADAPTER CARD

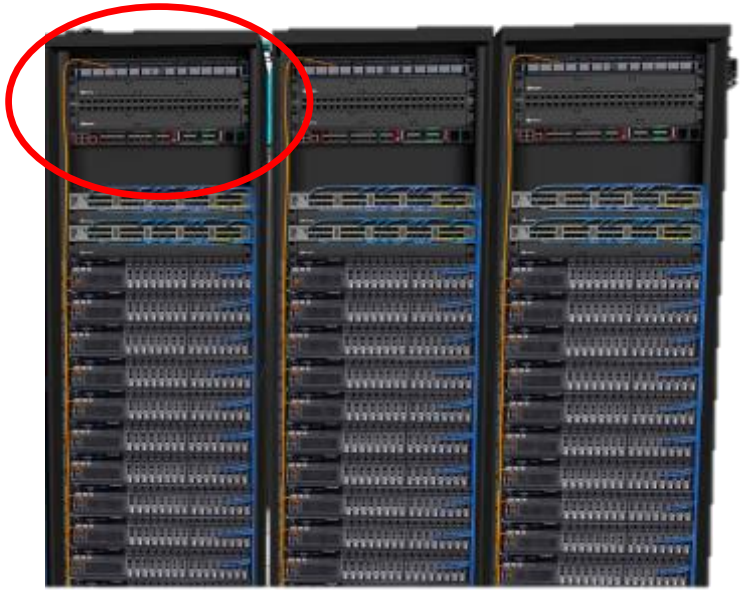
- Standard PCIe card, (Half Height Half Length)
- One x16 PCIe port from node
- One x4 Slingshot port cabled to TOR switch, QSFP28
- Air cooled

SLINGSHOT SWITCH & RACK INTEGRATION

MOUNTAIN
Distributed Switches

RIVER
TOR Co-Located Switches





Same Functionality

SLINGSHOT MOUNTAIN SWITCHES

- Switches distributed horizontally and vertically behind compute blades
- Reduces length of downlink cables
- Facilitates cable management
- Enables switch removal without disturbing cables
- Supports 8 to 64 switches per cabinet to vary the network performance

SLINGSHOT RIVER TOR SWITCH

- Switches co-located at top of rack
- Facilitates multi-rack groups with electrical group cables
- Supports 1 to 4 switches per rack to vary the network performance

Liquid Cooling Deep Dive

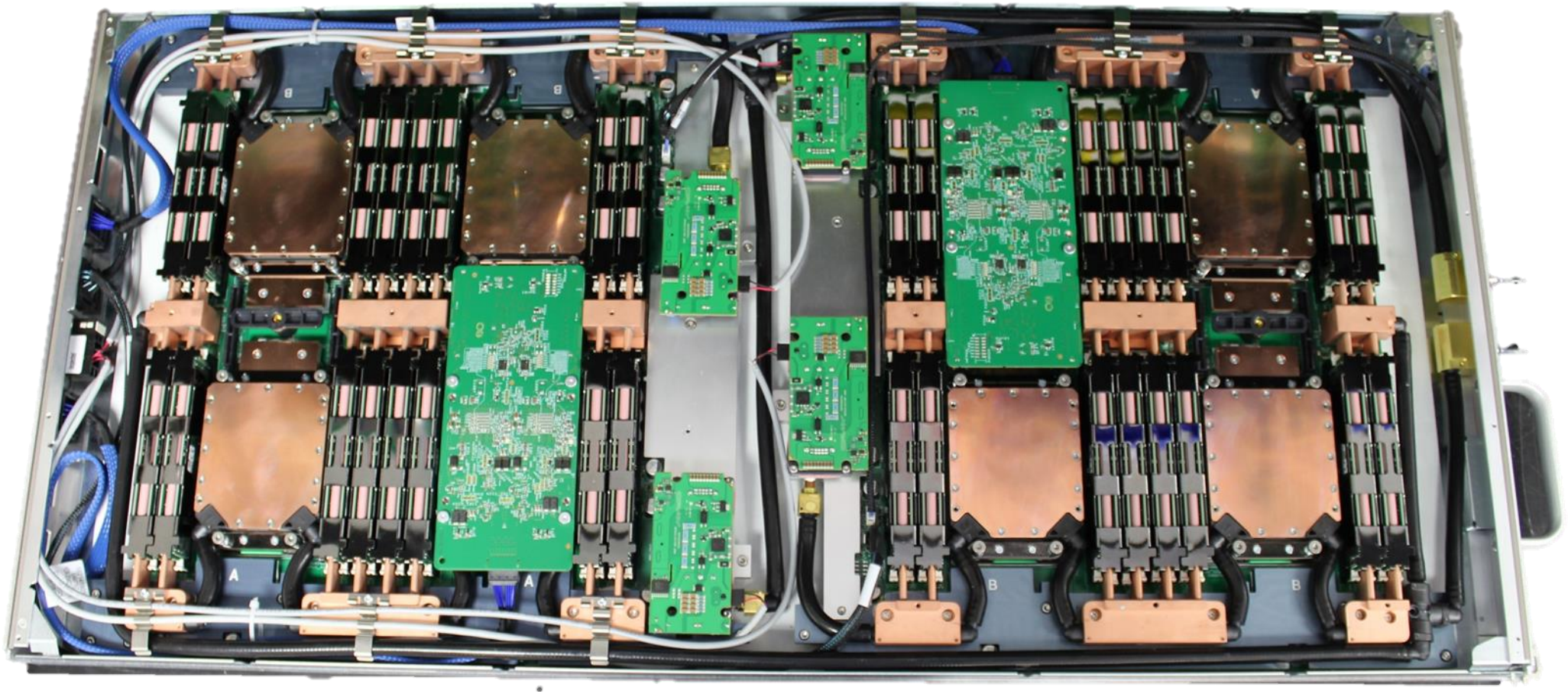
Wade Doll,

Principal Infrastructure Architect

Cray Inc.

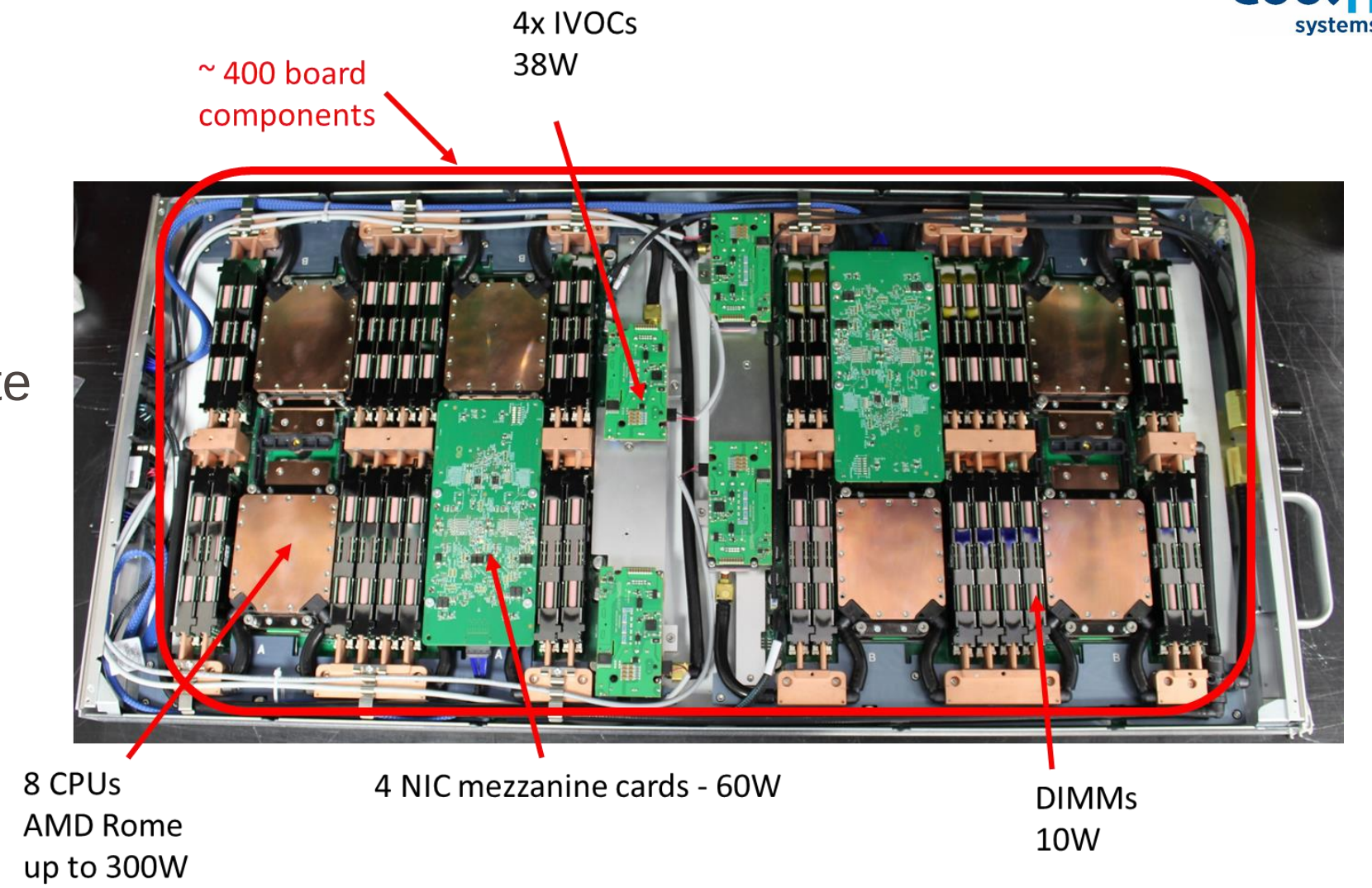
100% LIQUID COOLED MOUNTAIN BLADE

- Co-development with liquid cooling partner CoolIT Systems and Cray Inc.



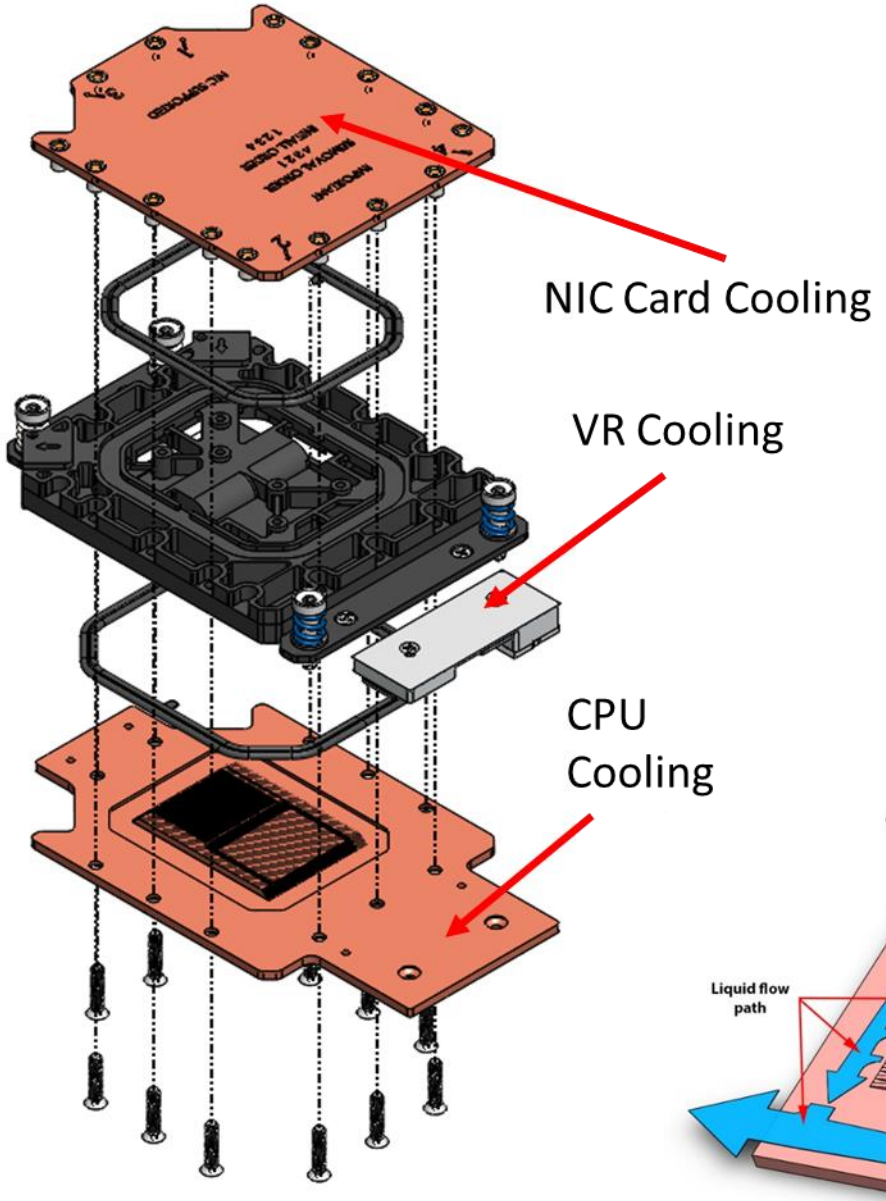
LIQUID COOLING SCOPE

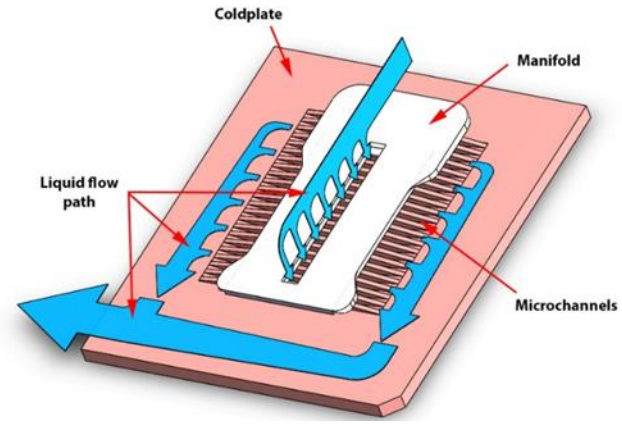
- 4kW total heat capture
- Treated water coolant
- Closed loop cooling system
- 2.8 lpm coolant flow rate
- Up to 50C inlet coolant temperature

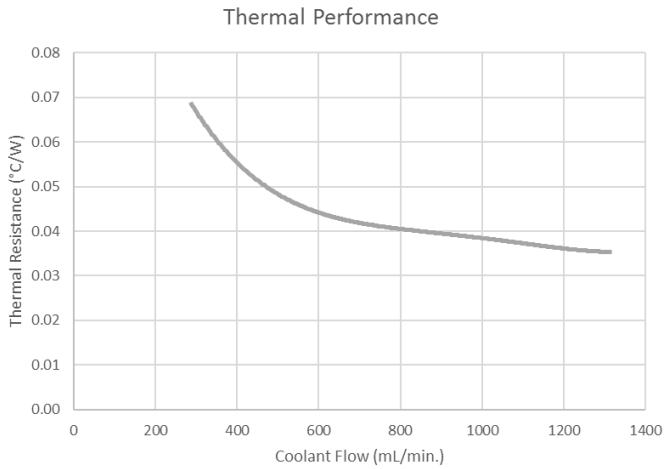


CPU COLD PLATE

- Double-sided cooling cold plate
- Cooling CPU and NIC card at the same time
- Low profile
- Patented Split-flow technology for CPU cooling

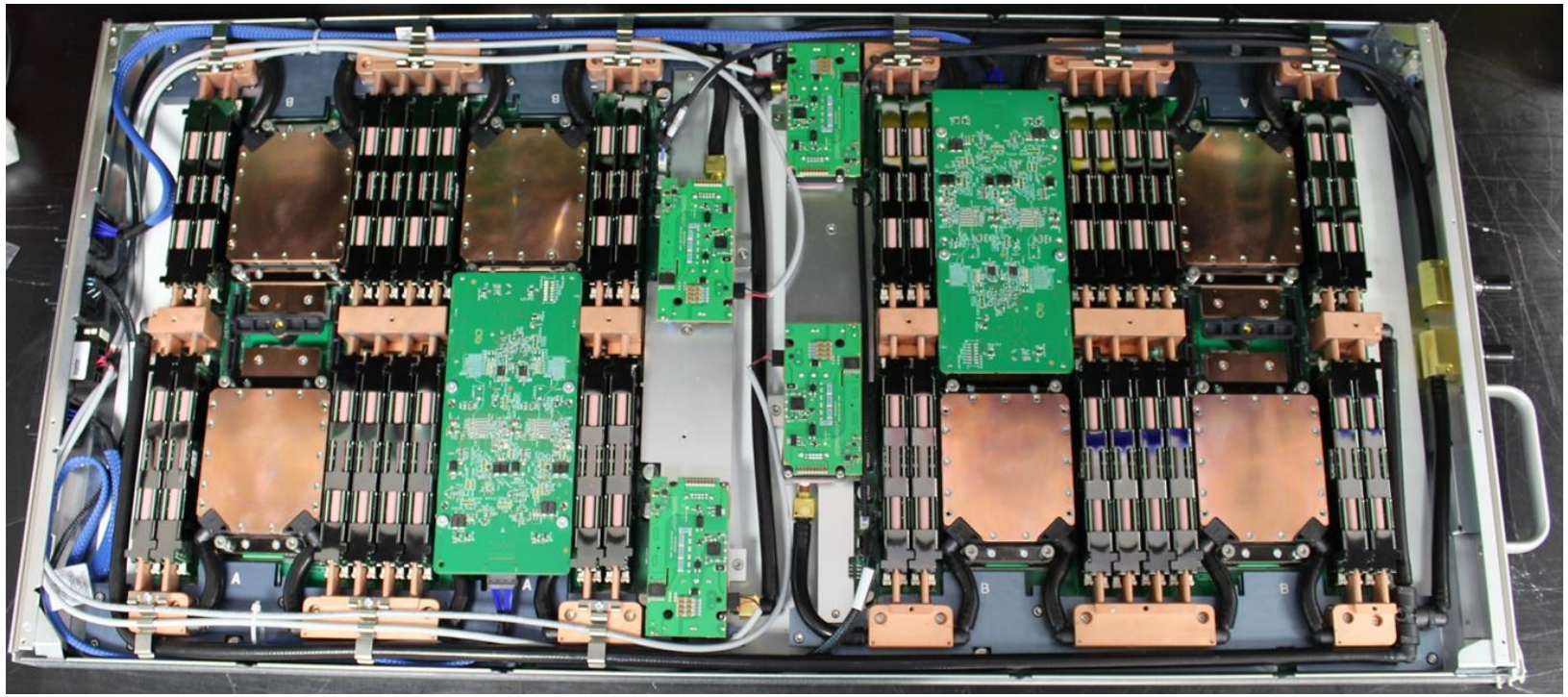






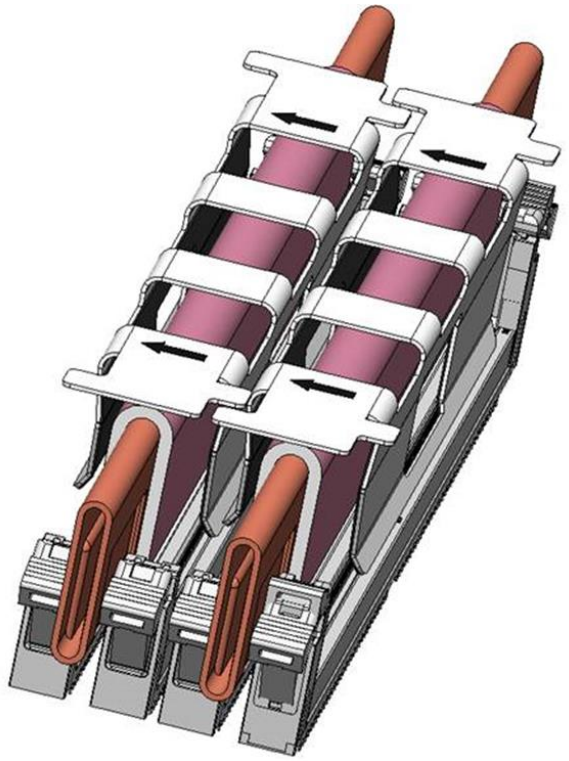
COMPONENT SERVICABILITY

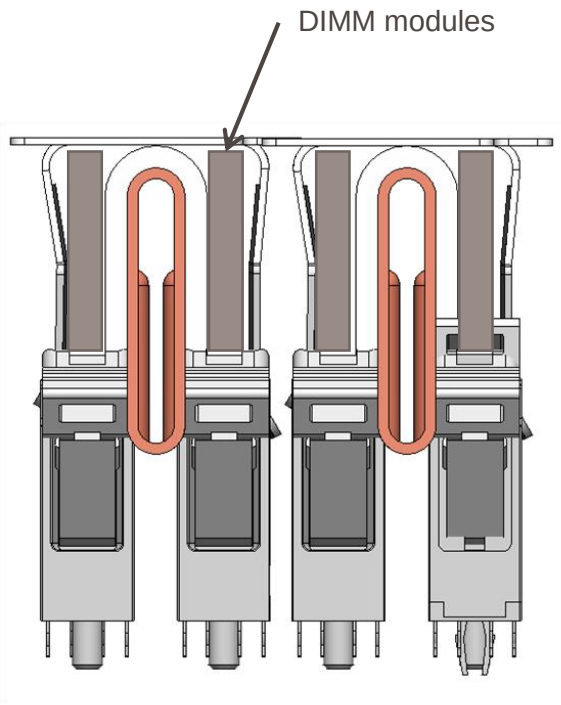
- Rubber (flexible) tubing makes the system serviceable



DIMM COOLING

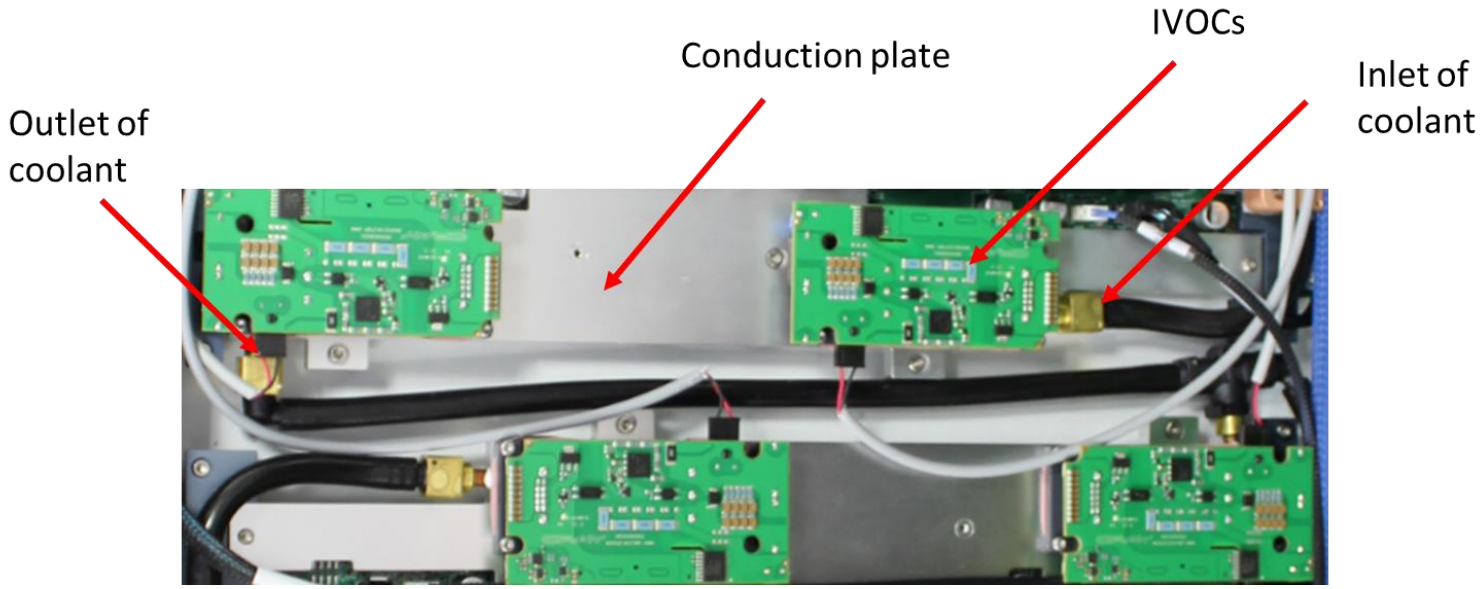
- Direct contact memory coolers
- Tool-less memory service
- Fluid circuit and cold plate structure left untouched



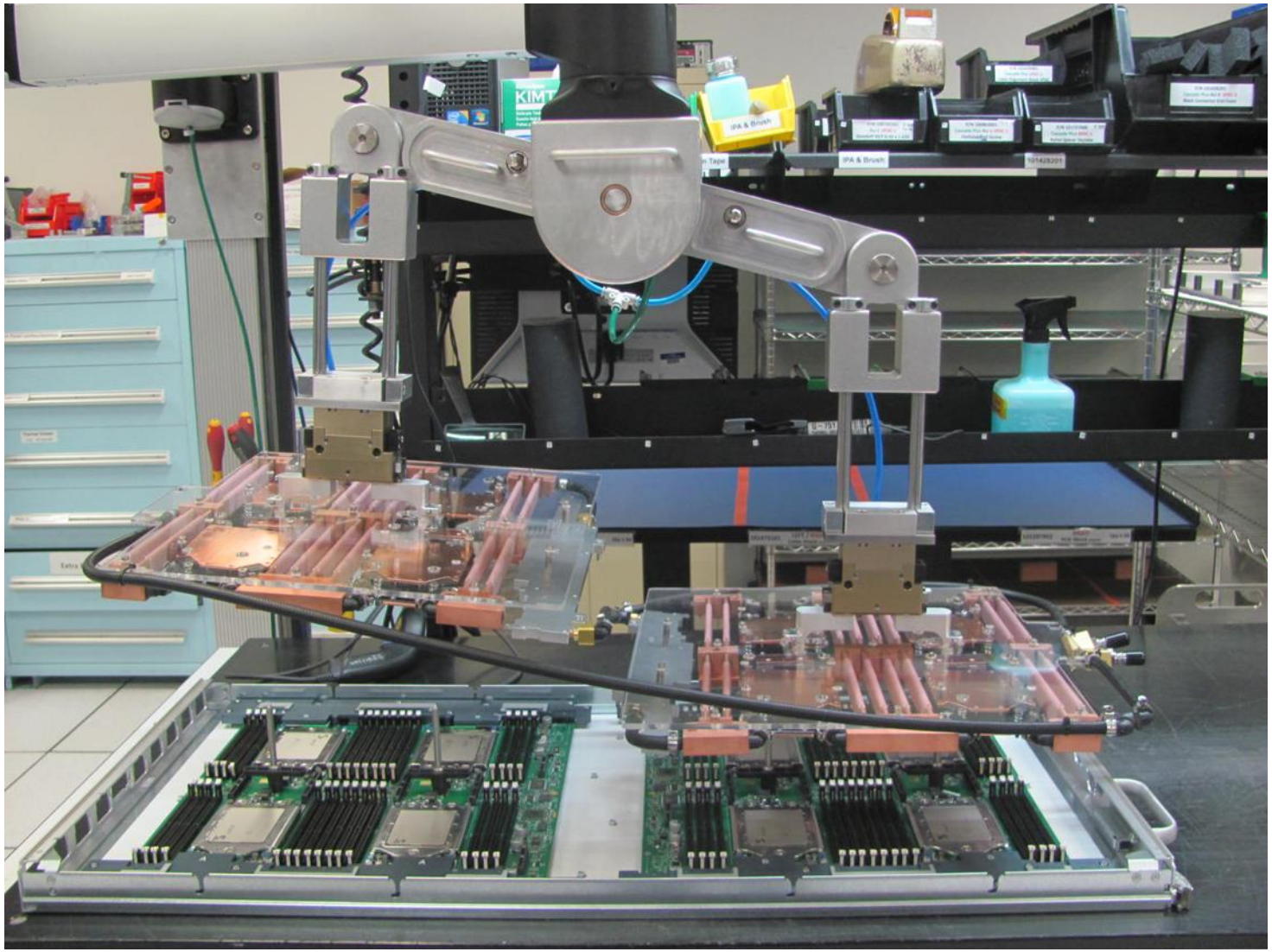


VOLTAGE CONVERTER COOLING (IVOC)

- IVOCs cooled from bottom side with conduction plate
- Makes IVOCs serviceable without disrupting fluid circuit or cold plate structure



INTEGRATION WITH THE BLADE

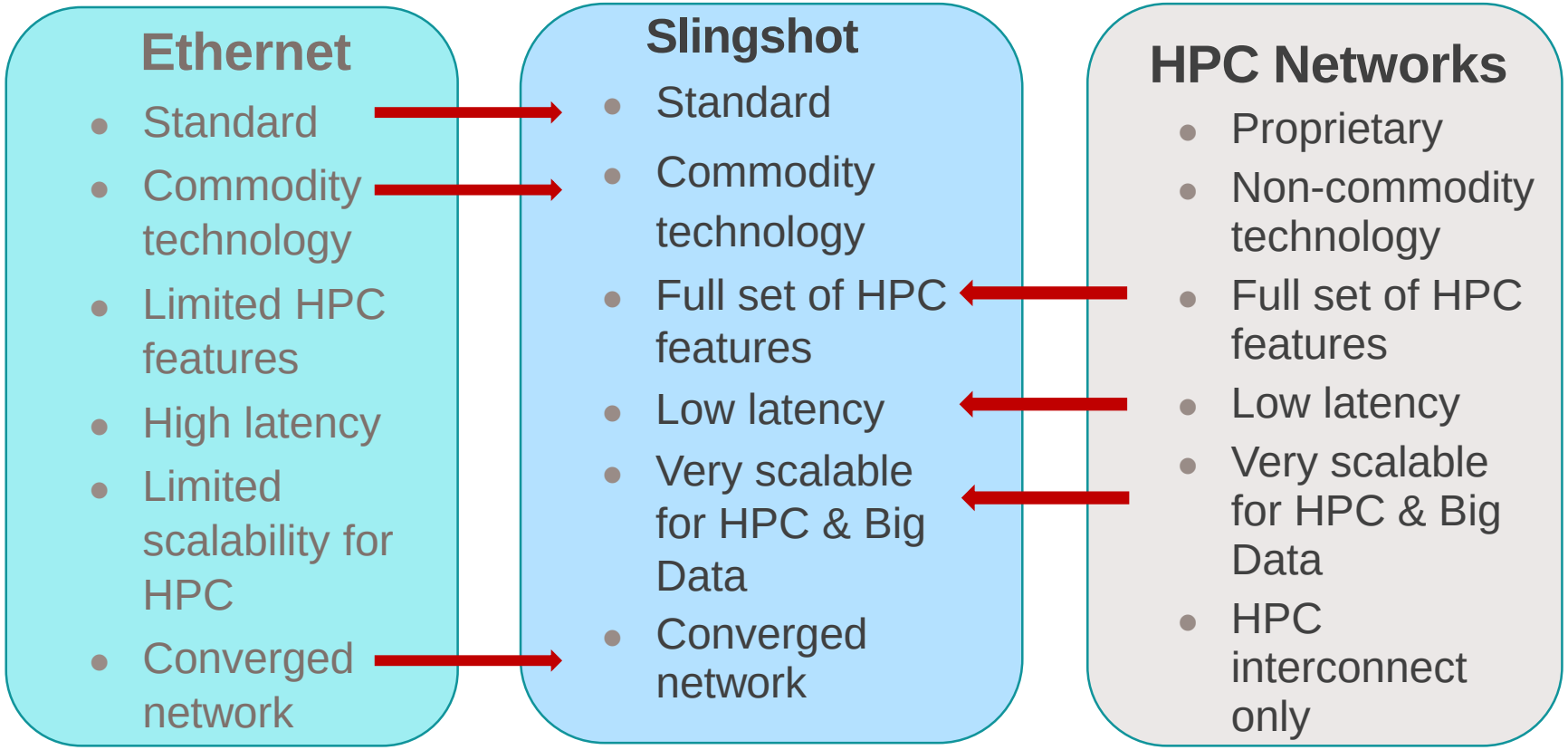


Slingshot Network

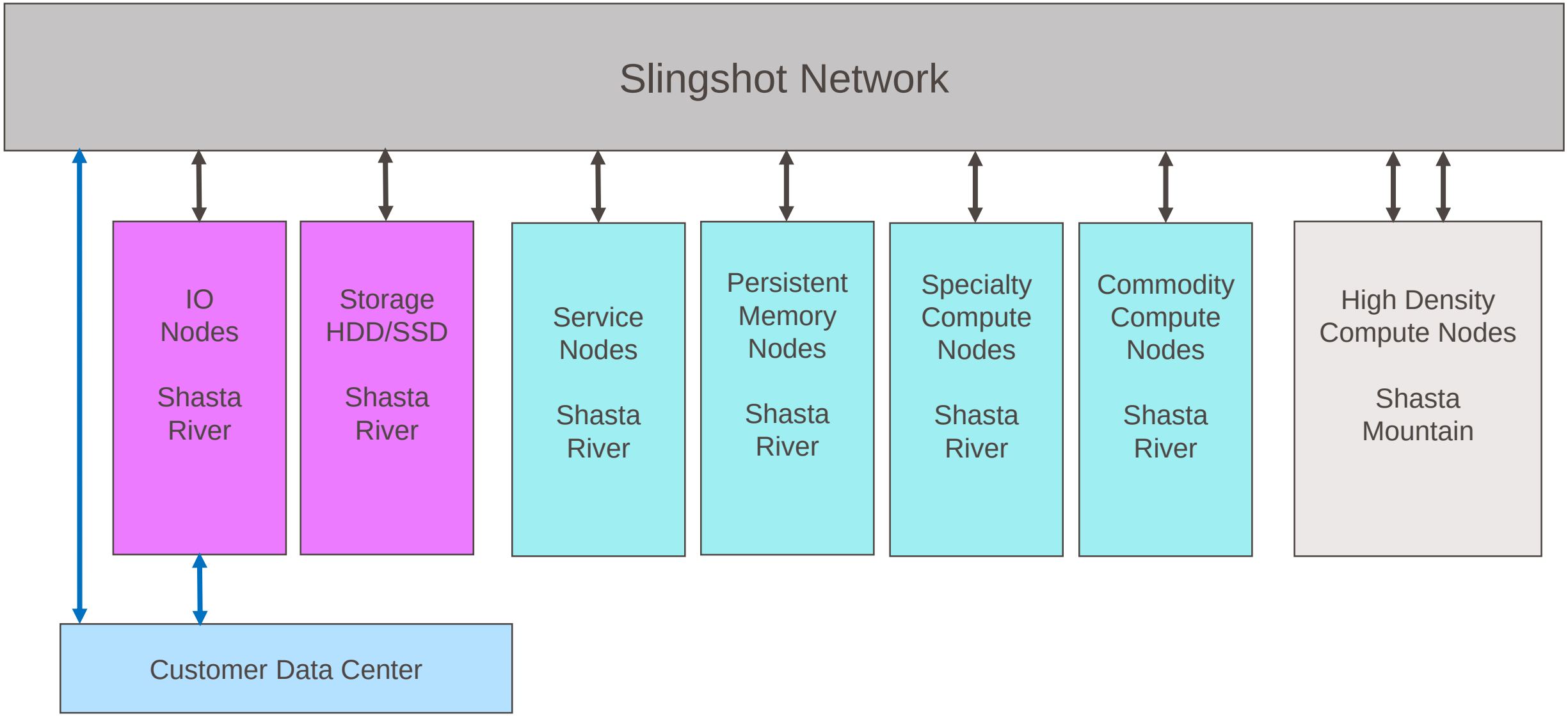
CRAY SLINGSHOT NETWORK

- Best of HPC and commodity Ethernet in one network
- Cray-designed, open HPC network for Shasta system
  - Supports Intel x86, AMD x86 and GPUs, Cavium ARM, and NVIDIA GPUs
- Interoperates with Ethernet ecosystem
  - Compatible with Ethernet NICs and datacenter switches
  - Leverage infrastructure (cables, optics, SerDes, etc.)
- Based on Cray Rosetta HPC Ethernet switch

LEVERAGING THE BEST OF BOTH WORLDS

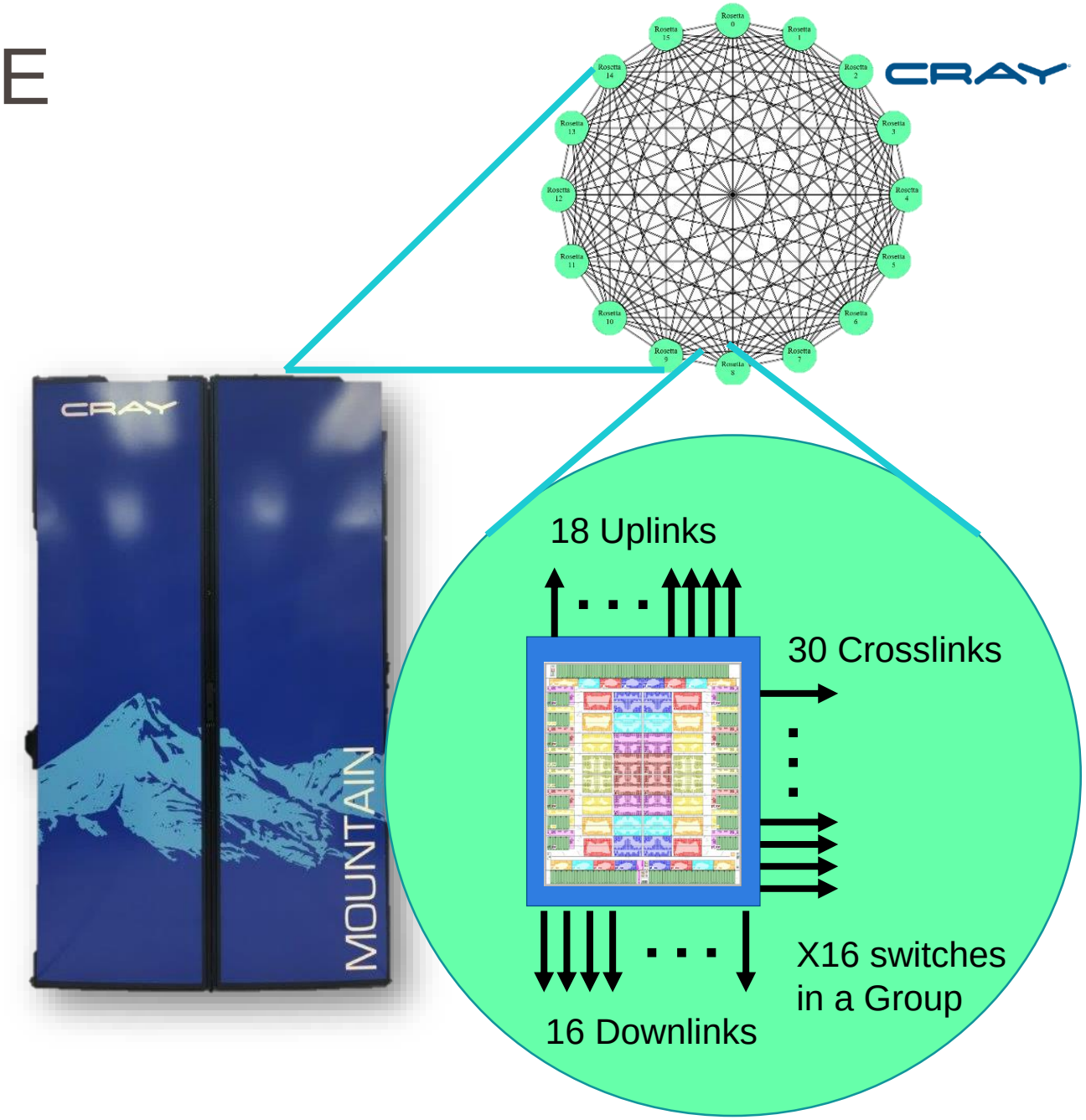


SLINGSHOT SYSTEM

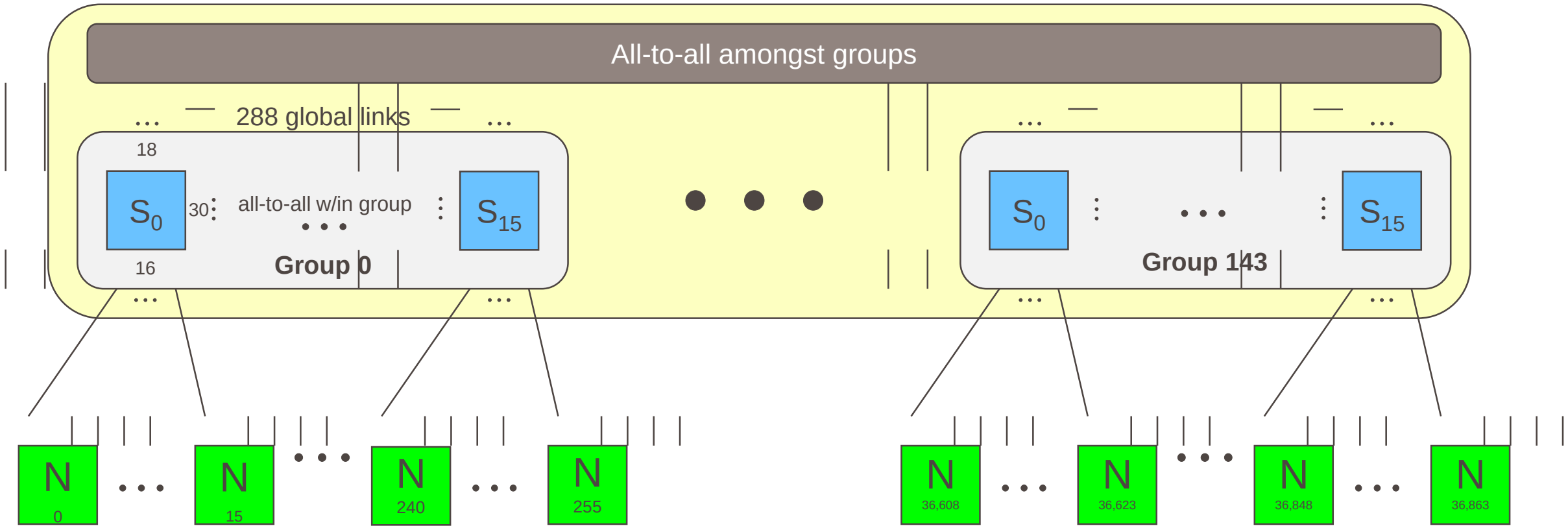


CPU COMPUTE EXAMPLE

- One DF group per CPU cabinet
- Four injection ports per compute blade
  - One per node
- All copper cables within group
- 16 Rosetta Switches per cabinet
- 256 CPU Compute nodes
- Scales to 144 groups
  - 144 CPU cabinets (36k nodes)



SLINGSHOT 16 SWITCH GROUP



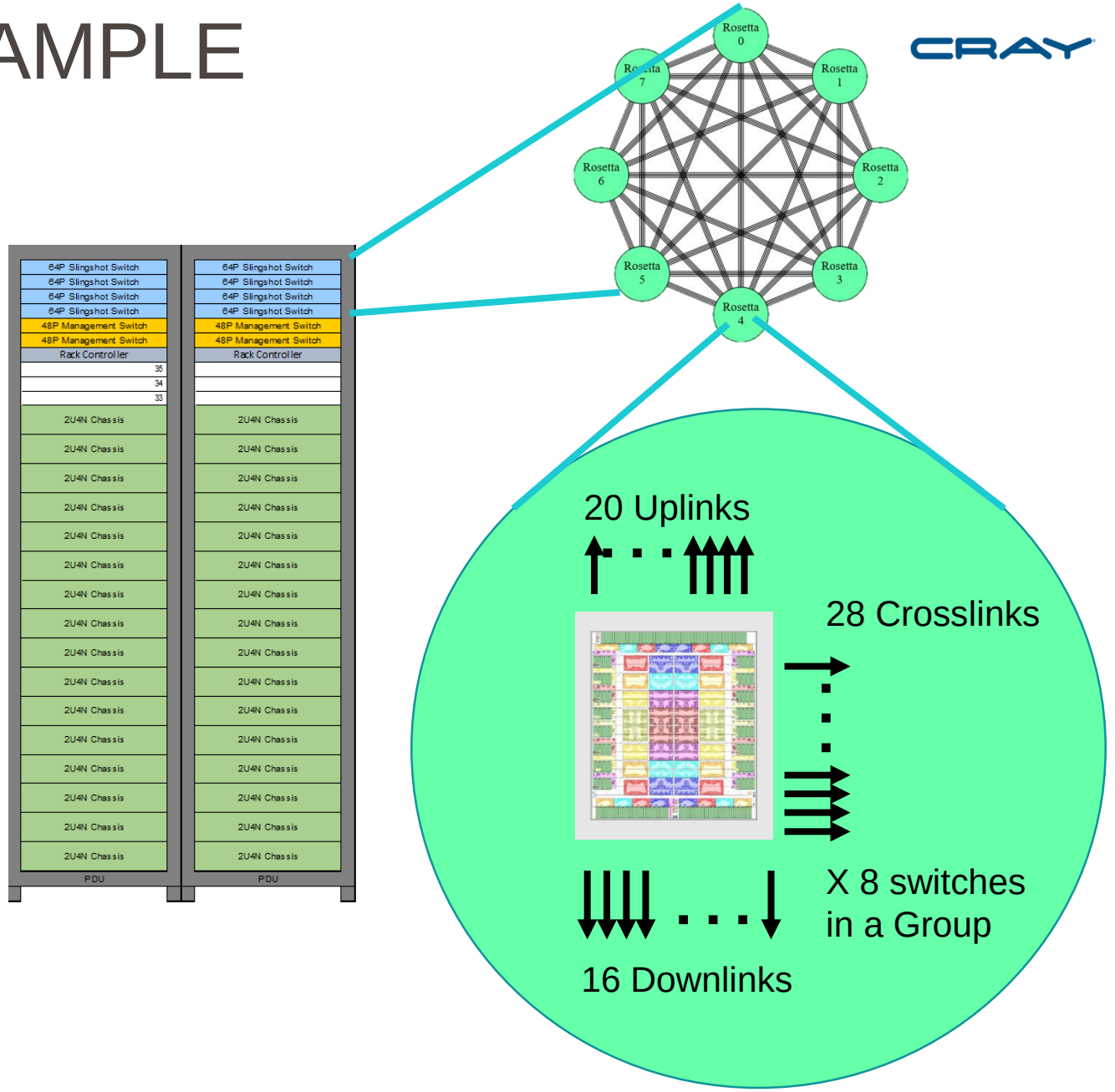
- 16 Switch Group – 1 Group per cabinet
- 256 nodes per cabinet – 512 sockets
- 16 Switch modules per cabinet

CABLING THE MOUNTAIN NETWORK

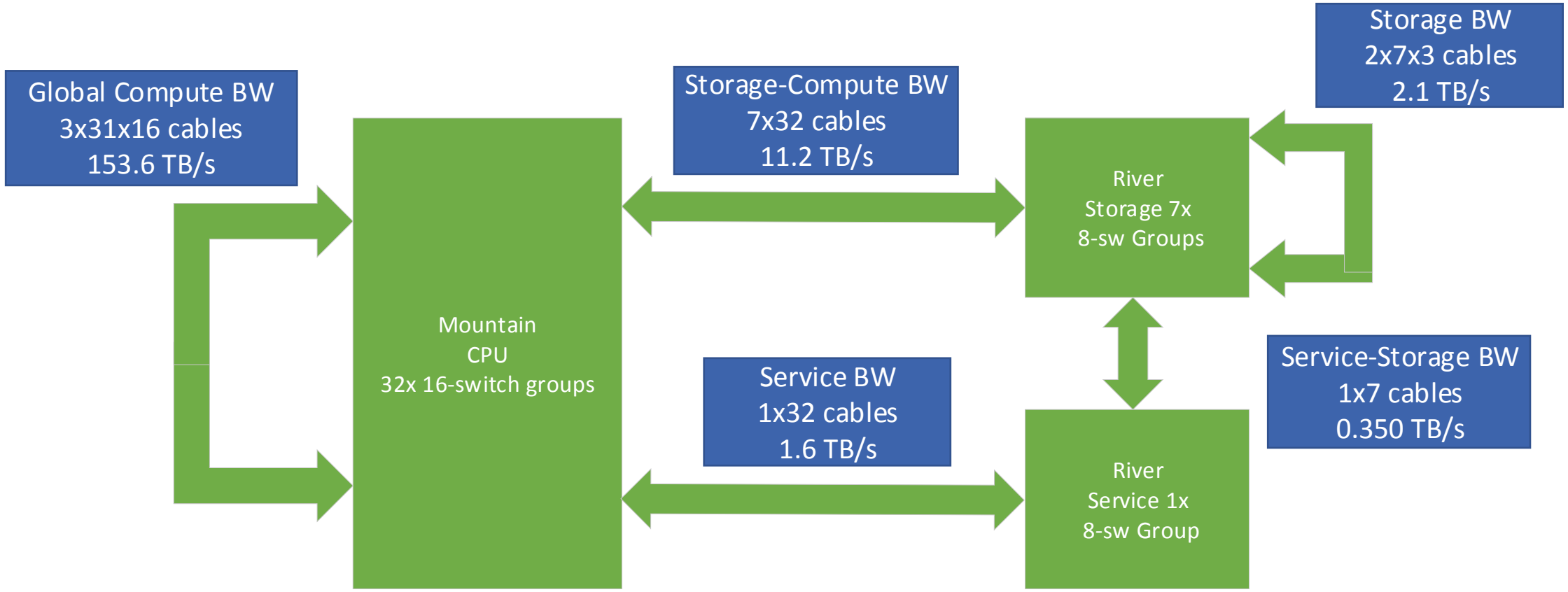
- All network cabling within the cabinet is copper (no optics $$)
- Optics reserved for cabling leaving the cabinet.
- Accessed from the rear
- Switch blades are extracted without removing extra cables for improved serviceability
- Standard QSFP-DD cables to leverage market volumes
- EMI shielded cables

STORAGE/SERVICE EXAMPLE

- 2 Cabinet DF group
- All Copper cables within Group
- 8 Rosetta Switches
- Up to 128 storage/service nodes
  - Compute also supported
- Scales to 162 Cabinets (10k nodes), 81 groups



EXAMPLE SLINGSHOT SYSTEM



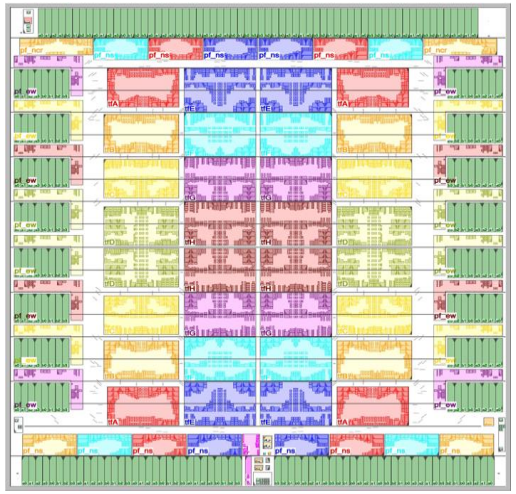
Cray Rosetta Switch

Bob Alverson,
Network Architect
Cray Inc.

CRAY SWITCH – ROSETTA

- 16nm design
  - Broadcom IP
  - Size 24mx28.5mm – 685mm^2
  - Each port is 4x56Gpbs – 200Gbps
- Protocol
  - Supports standard Ethernet protocol
  - HPC enhancements
    - Designed by Cray and Broadcom
  - Will be open to third parties in time
  - Added reliability and latency features for fabric links

Cray's Slingshot switch



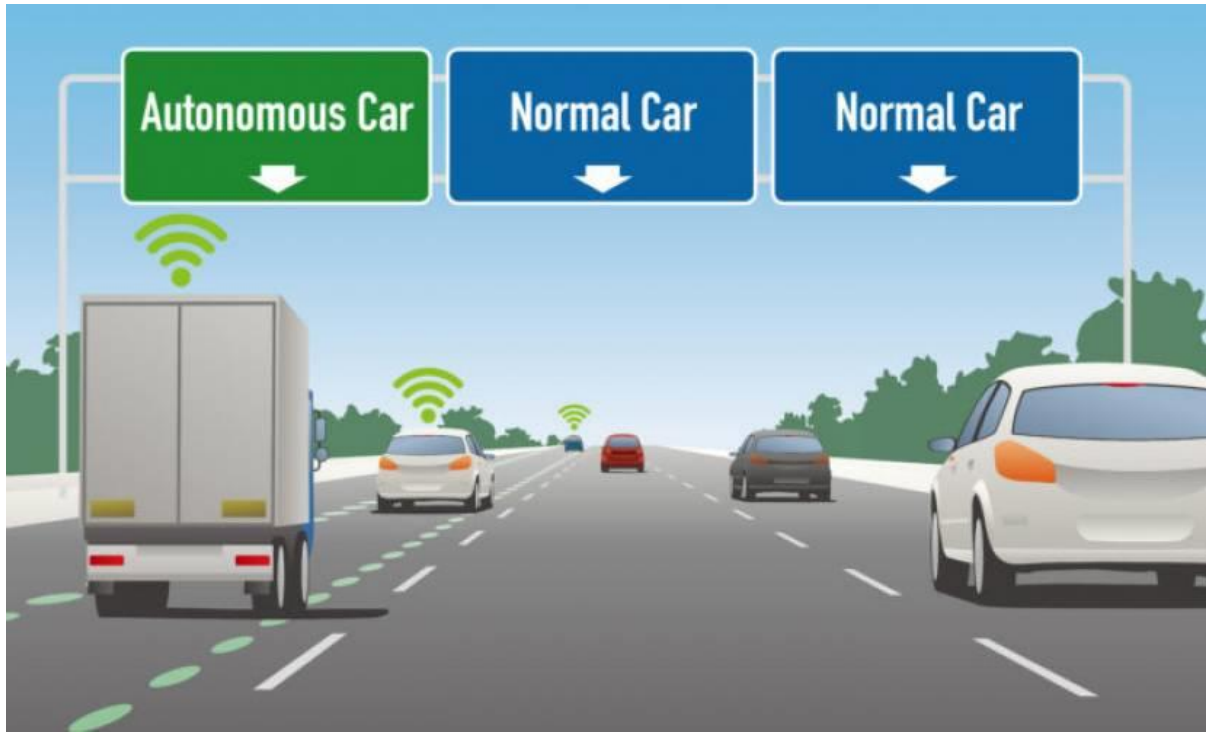
64 ports x 200 Gbps

ROSETTA HPC FEATURES

- Traffic classes
  - Class: Collection of buffers, queues and bandwidth
  - Intended to provide isolation between application groups via traffic shaping
- Aggressive adaptive routing
  - Has been very effective for Aries 5-hop dragonfly
  - Will be even more effective for Rosetta 3-hop dragonfly due to closer congestion information
- Multi-level congestion management
  - To minimize the impact of congested applications on others
- HPC enhancements to Ethernet protocol

ROSETTA TRAFFIC CLASSES

- Traffic Class controls
  - Priority
  - Assured minimum bandwidth
  - Maximum allowable bandwidth
  - Lossless or not
  - Switch buffer allocation/sharing



ADAPTIVE ROUTING

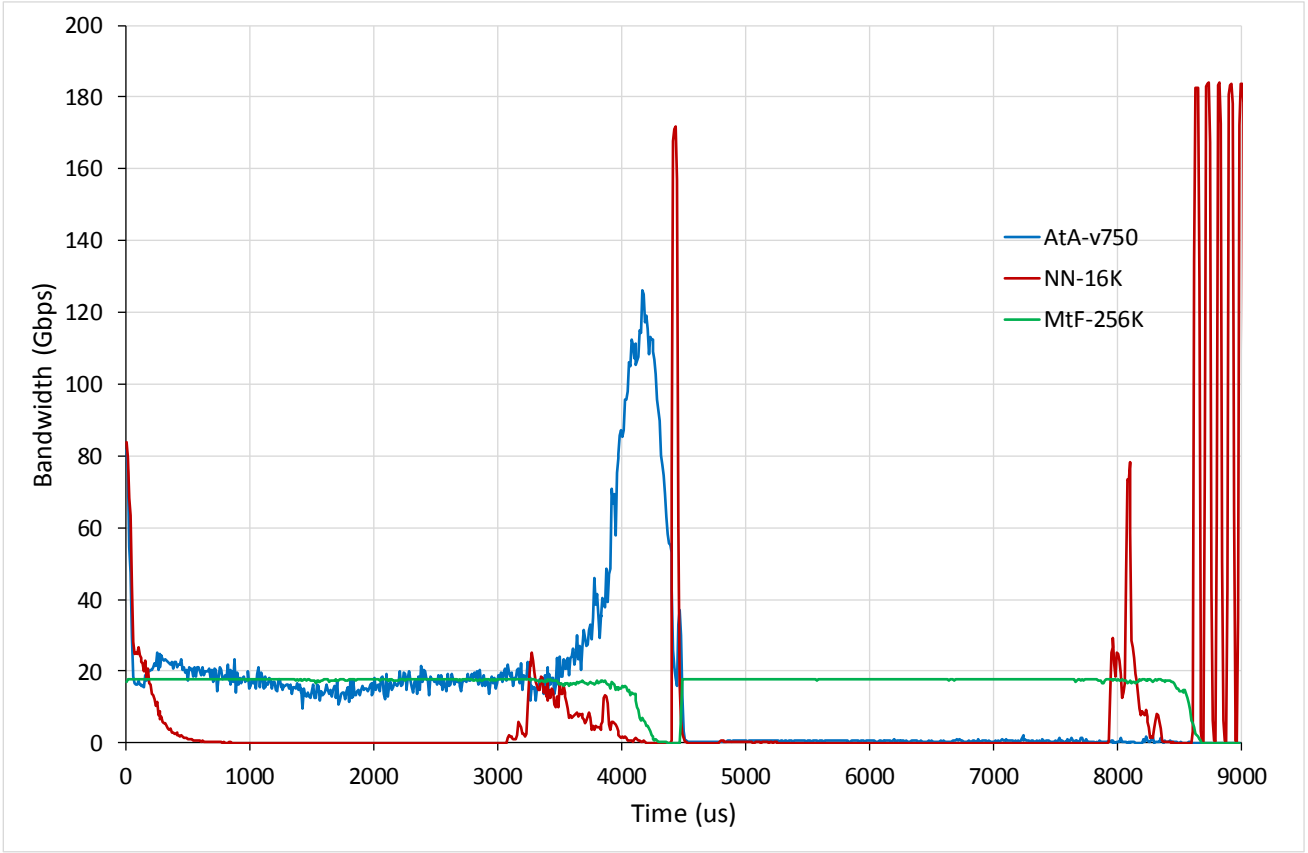
- Fine grain adaptive routing capability
  - Packet level routing which allows for out-of-order delivery of packets
  - Dynamic routes between any source/destination pair but packets arrive in-order at destination
  - Dynamic re-routing to avoid congested routes for source/destination pairs.
  - Improved route selection including more bias towards smaller hop count
- Minimal and non-minimal routing
  - Biases for direct and minimal paths
- Local and remote loading information
  - 3-hop networks allow closer congestion information

CONGESTION MANAGEMENT

- Multi-level congestion management protocol
  - Limit # of packets (amount of data) any node can inject into the network
  - Control network buffering
  - Mitigate incast

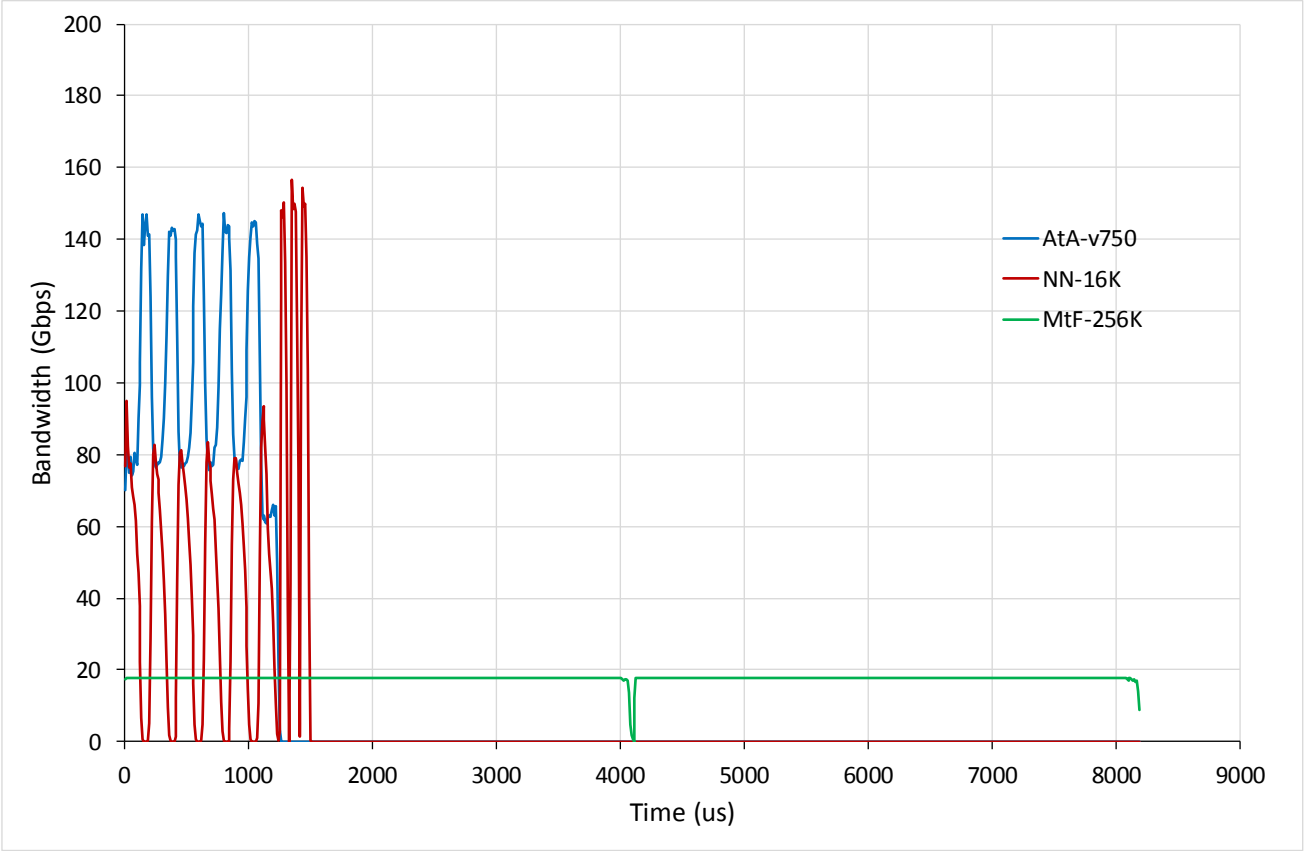
NO CONGESTION MANAGEMENT

- Simulated performance
- No Slingshot congestion control
- 1/3 of Nodes executing All to All Blue
- 1/3 of Nodes executing Nearest Neighbor Red
- 1/3 of Nodes executing Many to Few Green
  - Creating congestion
- All to All and Nearest Neighbor performance suffers



SLINGSHOT CONGESTION MANAGEMENT

- Simulated performance
- Slingshot congestion control enabled
- Many to Few congestion is controlled
  - Still gets maximum performance of egress limit
- Greatly improves time to solution for well behaved All to All and Nearest Neighbor

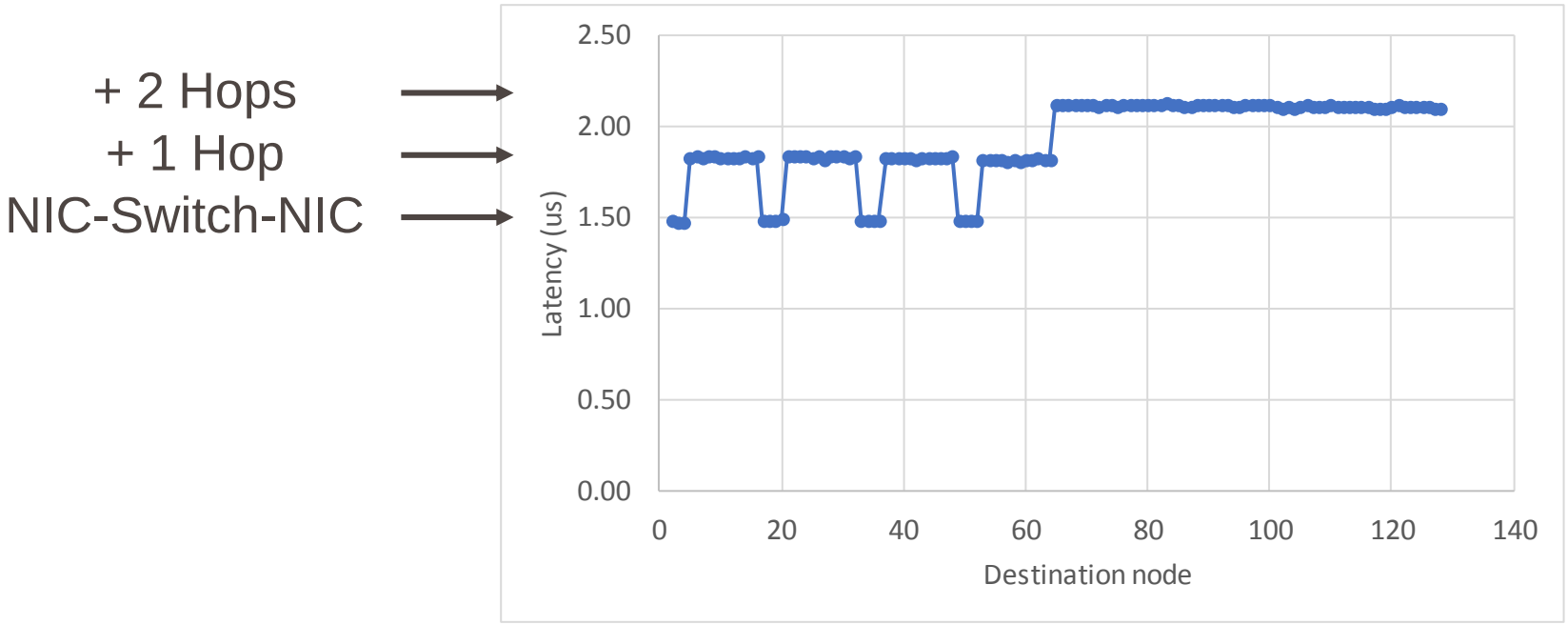


ENHANCEMENTS TO ETHERNET PROTOCOL

- Co-designed with Broadcom

- Reduced inter packet gap
- Reduced minimum frame size
- Optimized header
  - Based on IPV4 and compressed IPV6 formats
- Higher link speeds
  - Leverage emerging SerDes standards in Ethernet
- Low latency FEC
- Link level retry
  - Faster error recovery (potentially save latency by avoiding FEC correction delays)
- Reduced link width support
  - Energy reduction without wakeup penalty; failure tolerance

MPI POINT TO POINT LATENCY

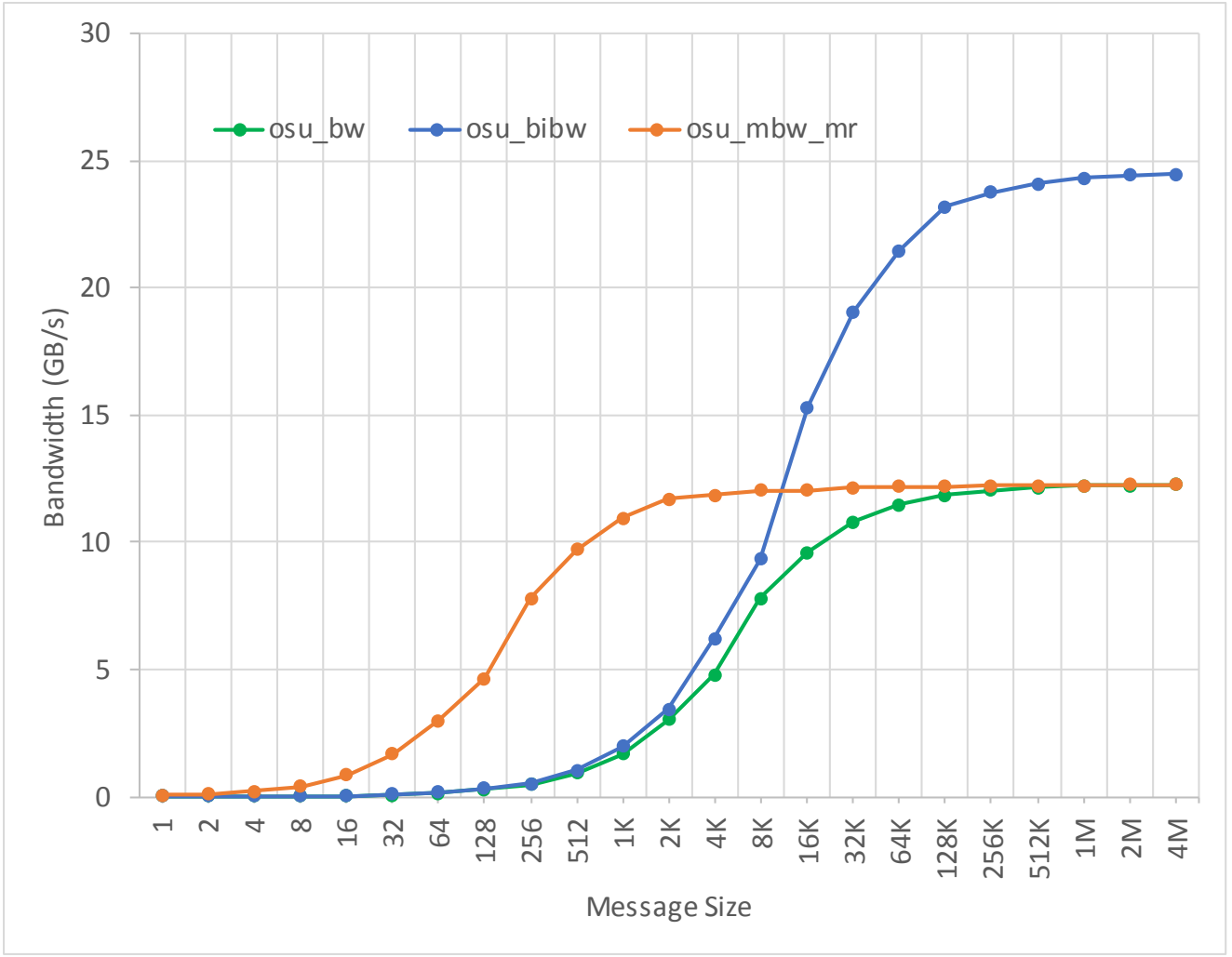


- Preliminary results
- 128 Nodes
- 2 Groups
- 4 Rosetta per Group
- 100G Std. NICs

MPI BANDWIDTH

- Preliminary results
- 128 Nodes
- 2 Groups
- 4 Rosetta per Group
- 100G Std. NICs

- Point to Point single node bandwidth



SAFE HARBOR STATEMENT

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts.

These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

THANK YOU

QUESTIONS?

cray.com

@cray_inc

linkedin.com/company/cray-inc-/