Power Systems AC922 Overview

Chris Mann
IBM Distinguished Engineer
Chief System Architect, Power HPC Systems
December 11, 2017

IBM POWER HPC Platform Strategy

- High-performance computer and high-performance analytics drive common platform design

- Servers will be predominately 2-socket designs

- Developing deeper relationships with technology partners – ref OpenPOWER

- Majority of floating-point performance will come from GPUs

- Utilize Industry-standard compliant 19" racks and electronics enclosures
    - Air and water cooling options

- Platforms will be based on a common enclosure form factor
    - Enclosure provides working envelope that we will continue to enhance with the latest technology from IBM, NVIDIA, Mellanox and other OpenPOWER partners
    - Enclosure provides a platform with sufficient power, cooling capability to support these enhancements

IBM POWER GPU Intensive Roadmap

POWER S822LC
- 2 POWER8 Processors
  - 190 Turismo module
- 2 x16 Gen 3 FHFL PCIe slots
  - Supports 2 NVidia K80 GPU's
  - Supports 2 PCIe adapters
- 1 x8 Gen 3 HHHL PCIe, CAPI
- 1 x16 Gen 3 HHHL PCIe, CAPI
- 1 x8 Gen 3 PCIe
- 32 DDR3 IS DIMM's
  - 4, 8, 16, 32GB DIMMs
  - 32 – 1024GB Memory Capacity
- 2 SATA SFF HDD / SSD
- 2 1300W Power Supplies
  - 200VAC Input
- BMC support structure
  - IPMI, USB, EN, VGA
- Air cooled

POWER S822LC for HPC
- 2 POWER8 w/ NVLink Processors
  - 190 module
- 1, 2, 4 NVidia "Pascal" GPU's
  - 300W, SXM2 Form Factor, NVLink 1.0
- 2 x16 Gen 3 HHHL PCIe, CAPI enabled
- 1 x8 Gen3 HHHL PCIe, CAPI enabled
- 32 DDR4 IS DIMM's
  - 4, 8, 16, 32GB DIMM's
- 2 SATA SFF HDD / SSD
- Pluggable NVMe storage adapter
  - 1.6, 3.2TB Capacity
- 2 1300W power supplies
  - 200VAC Input
- BMC Support Structure
  - IPMI, USB, EN, VGA
- Air and water cooled options
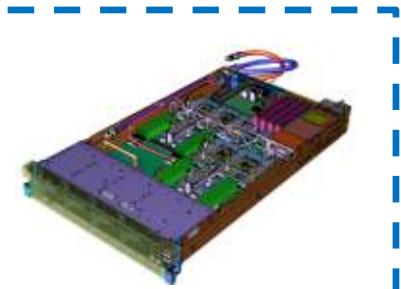
POWER AC922
- 2 POWER9 Processors
  - 190, 250W modules
- 4-6 NVidia "Volta" GPU's
  - 300W, SXM2 Form Factor, NVLink 2.0
- 6 GPU configuration, water cooled
- 4 GPU configuration, air or water cooled
- 2 Gen4 x16 HHHL PCIe, CAPI enabled
- 1 Gen4 x4 HHHL PCIe
- 1 Gen4 Shared x8 PCIe adapter
- 16 IS DIMM's
  - 8, 16, 32, 64, 128GB DIMMs
- 2 SATA SFF HDD / SSD
- 2 2200W power supplies
  - 200 VAC, 277VAC, 400VDC input
  - N+1 Redundant
- Second generation BMC Support Structure
- Pluggable NVMe storage adapter option

2015 **2016** **2017 - 2018**

Introducing Power AC922 2U HPC Server – Premier Accelerator Platform
Leveraging POWER9 Innovation
High-performance core, bandwidth, accelerator differentiation

For compute-intensive workloads, accelerators are key to Cognitive/AI economics

POWER9 PowerAccel

P8

CAPI 1.0
PCIe Gen3
NVLink 1.0

P9

CAPI 2.0
PCIe Gen4
NVLink 2.0
OpenCAPI

Fast PowerAccel Interconnects for Accelerators

NVLink 1.0

Faster PowerAccel Interconnect for Accelerators

OpenCAPI / NVLink 2.0
PCIe Gen 4
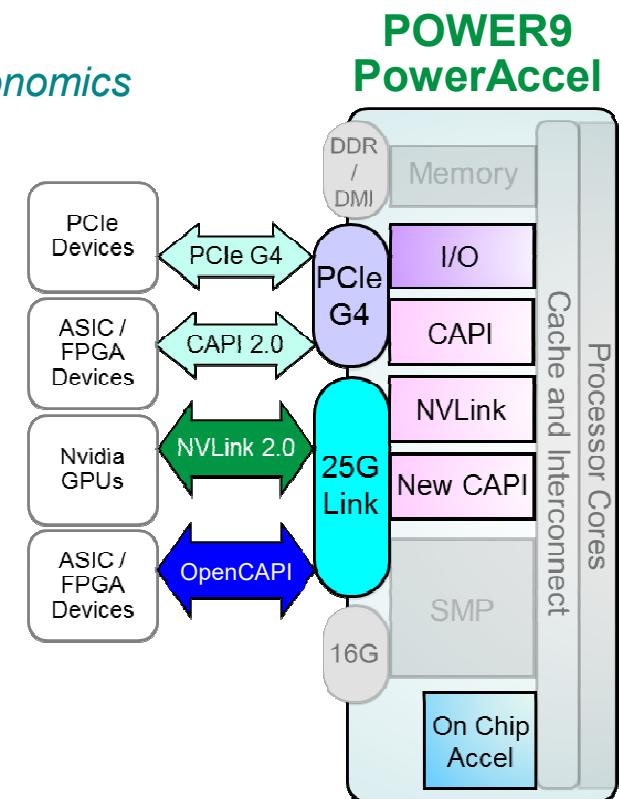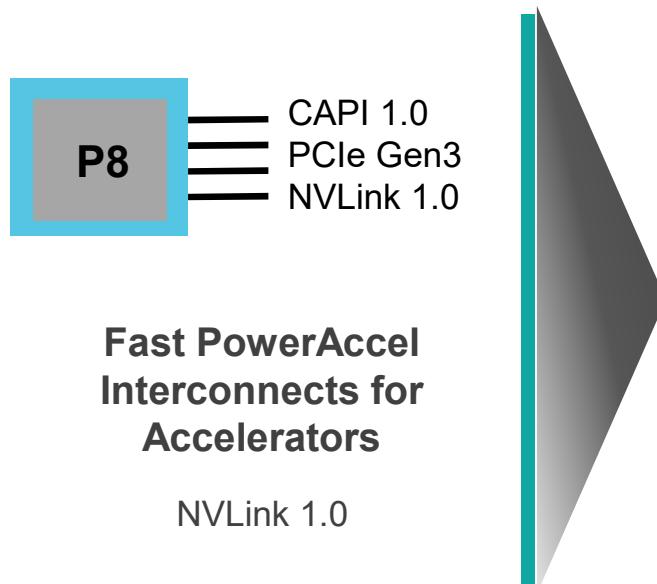
IBM internal and Business Partners only

Power AC922 - *POWER9 with increased GPU and IO bandwidth for differentiation*

Realize unprecedented performance and application gains with POWER9 and NVLink 2.0

• 2 POWER9 CPUs and up to 4 "Volta" NVLink 2.0 GPUs in a versatile 2U Linux server

• PCIe Gen4 bus has double I/O Bandwidth vs. PCIe Gen3

• CPU (Turbo)/GPU (Boost) enabled for improved data center efficiency and performance to be maintained at high levels

High level System Overview
- 2-Socket, 2U Packaging
- 40 P9 Processor cores
- 4 NVIDIA Volta 2.0 GPUs
- 1 TB Memory (16x - 64GB DIMMs)
- 4 PCIe Gen4 Slots
- 2x SFF (HDD/SSD), SATA, Up to 7.7 TB storage
- Supports 1.6TB and 3.2TB NVMe Adapters
- Default 3 year 9x5 warranty, 100% CRU

POWER9 HPC Product Description

POWER9 Processor – Common Features

- **14HP Technology**

- **24 POWER9 Cores with new SMC uArchitecture**
  - 4 threads per core
  - 2 cores per chiplet
  - LINUX Radix Page Table support

- **Large, Low-latency Cache**
  - 512k private L2, 10MB NUCA L3 per chiplet

- **Direct Attach Memory Suppor**t
  - 8 DDR4 channels
  - 2666 MHz DIMM support
  - 140 GB/s streaming bandwidth

- **SMP 2 Socket Support via 4B, 16 Gb/s X-Bus**

- **Leadership Hardware Acceleration Platform**
  - Enhanced on-chip acceleration
  - NVIDIA NVLink 2.0
  - CAPI 2.0, Coherent accelerator and storage attach vie PCIe G4
  - OpenCAPI 3.0, Improved latency and bandwidth, open interface

- **Network Interconnect**
  - CAPI 2.0 attached Mellanox CX-5 support
  - Address translation support
  - Atomics
  - NVDIA GPU direct
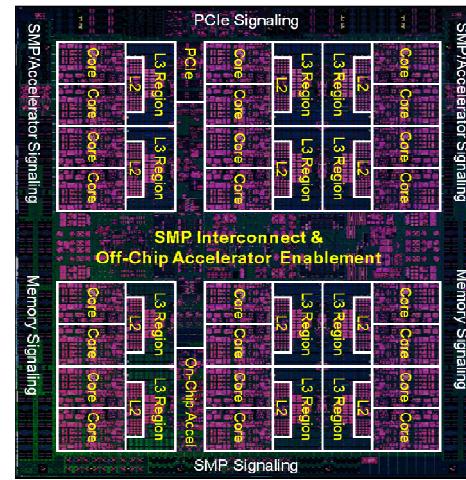
- **Gen4 PCIe, 48 Lanes @ 16Gb/s**
  - 6 PHBs
  - PCIe Bus 0 turbo-charged to support next generation HDR IB

- **Self Boot Capability**

- **Instant ON/OFF**

- **Cloud Management QoS**

POWER AC922 Design – 4 GPU

Power Supplies (2x)
- 2200W
- 200VAC, 277VAC, 400VDC input

NVidia Volta GPU
- 3 per socket
- SXM2 form factor
- 300W
- NVLink 2.0
- Air/Water Cooled

PCIe slot (4x)
- Gen4 PCIe
- 2, x16 HHHL Adapter
- 1, Shared slot
- 1 x8 HHHL Adapter

Memory DIMM's (16x)
- 8 DDR4 IS DIMMs per socket
- 8, 16, 32,64, 128GB DIMMs

BMC Card
- IPMI
- 1 Gb Ethernet
- VGA
- 1 USB 3.0

Power 9 Processor (2x)
- 18, 22C water cooled
- 16, 20C air cooled
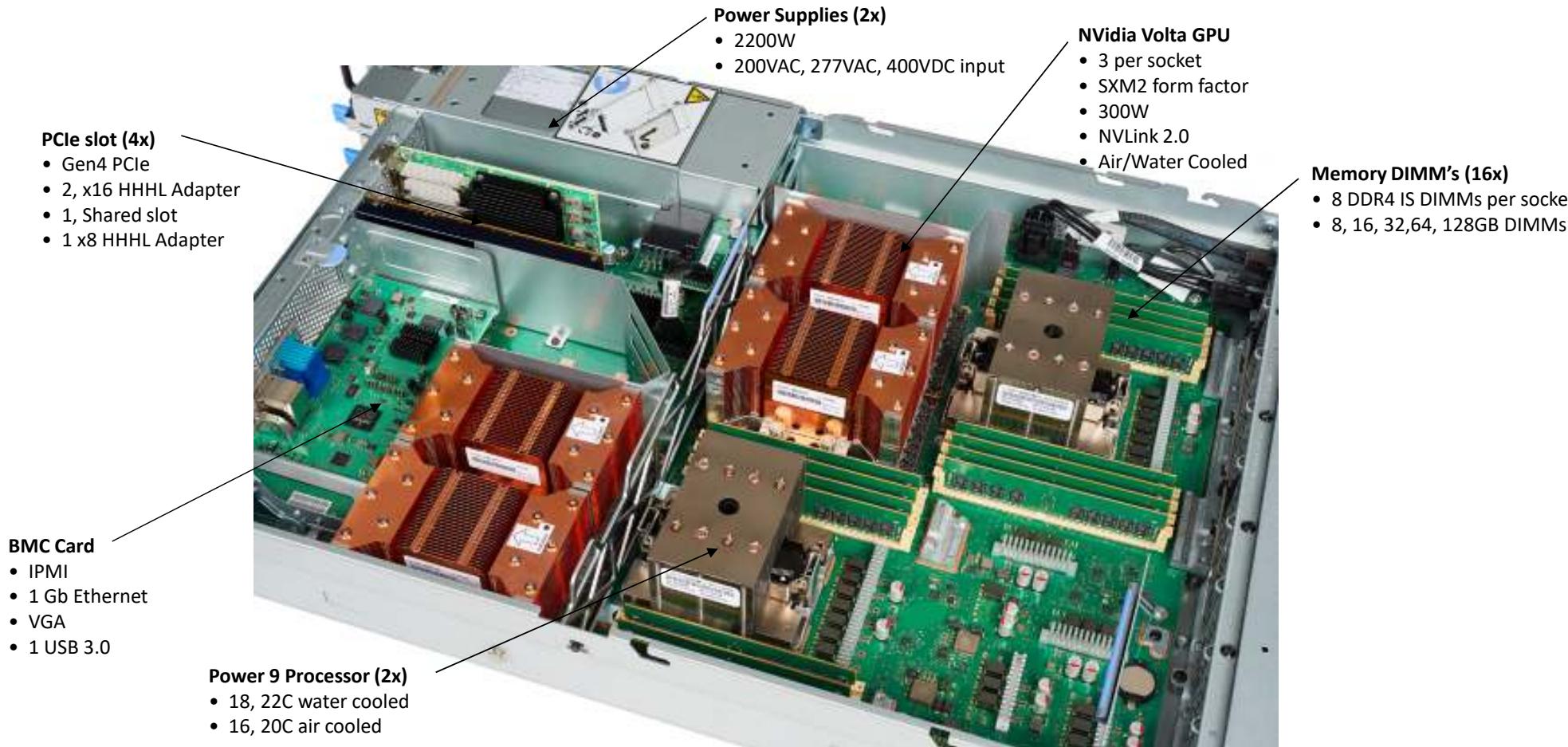
NVIDIA Volta Specifications

NVIDIA Volta GPU Features	
Peak double precision floating point performance	7.8 TFLOPS
Memory bandwidth	900 GB/sec
GPU Memory Size	16 GB
NVLink "Bricks" (8 lane interface)	6
NVLink Interconnect Bi-Directional	300GB/s
Maximum Power	300W

https://www.nvidia.com/en-us/data-center/tesla-v100/

NVLink Evolution in POWER HPC

NVIDIA P100 GPU with NVLink 1.0

Graphics Memory

40+40 GB/s

Graphics Memory

40+40 GB/s

40+40 GB/s

System Memory

POWER8

NVIDIA Volta GPU with NVLink 2.0

Graphics Memory

75+75 GB/s

Graphics Memory

75+75 GB/s

75+75 GB/s

System Memory

POWER9
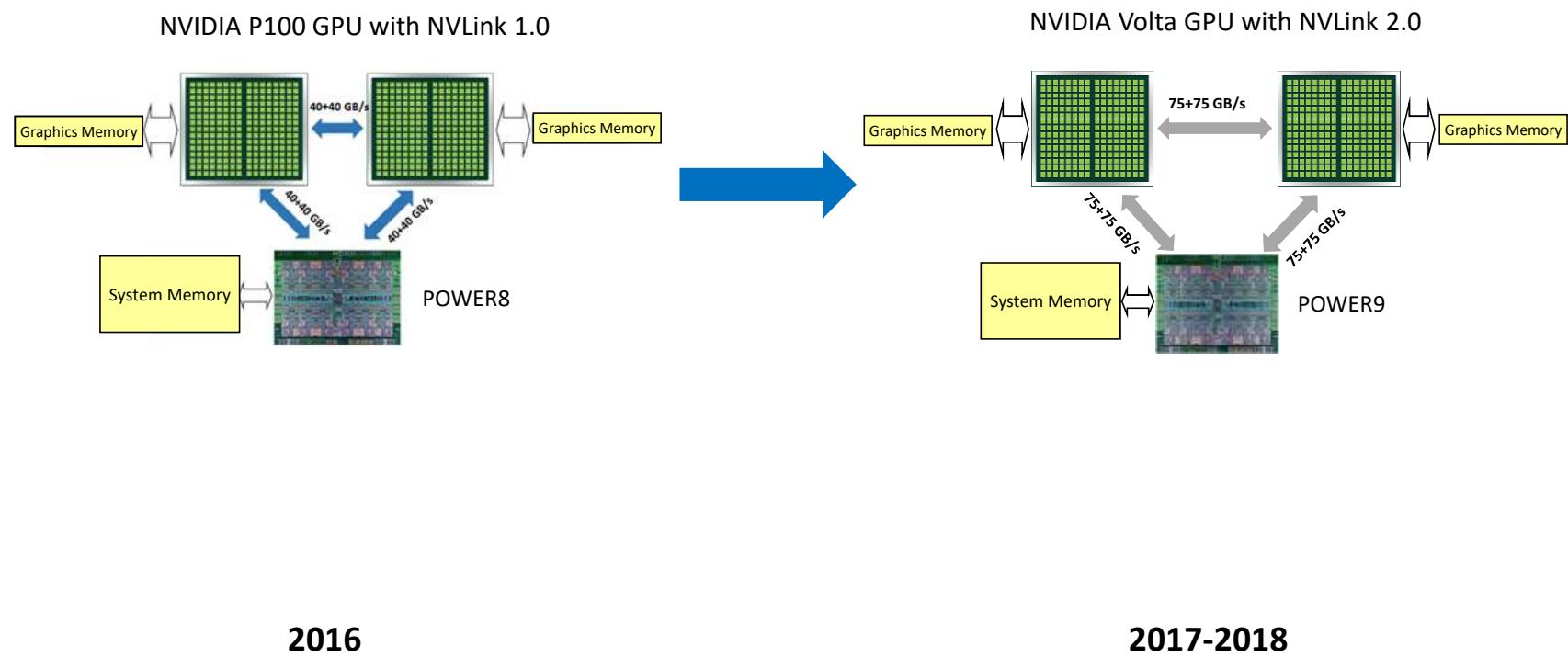
2016

2017-2018

GPU Attach Bandwith Comparison, PCIe Gen3 verses NVLink

IBM

All values are <u>MEASURED</u> Single Direction Bandwidths
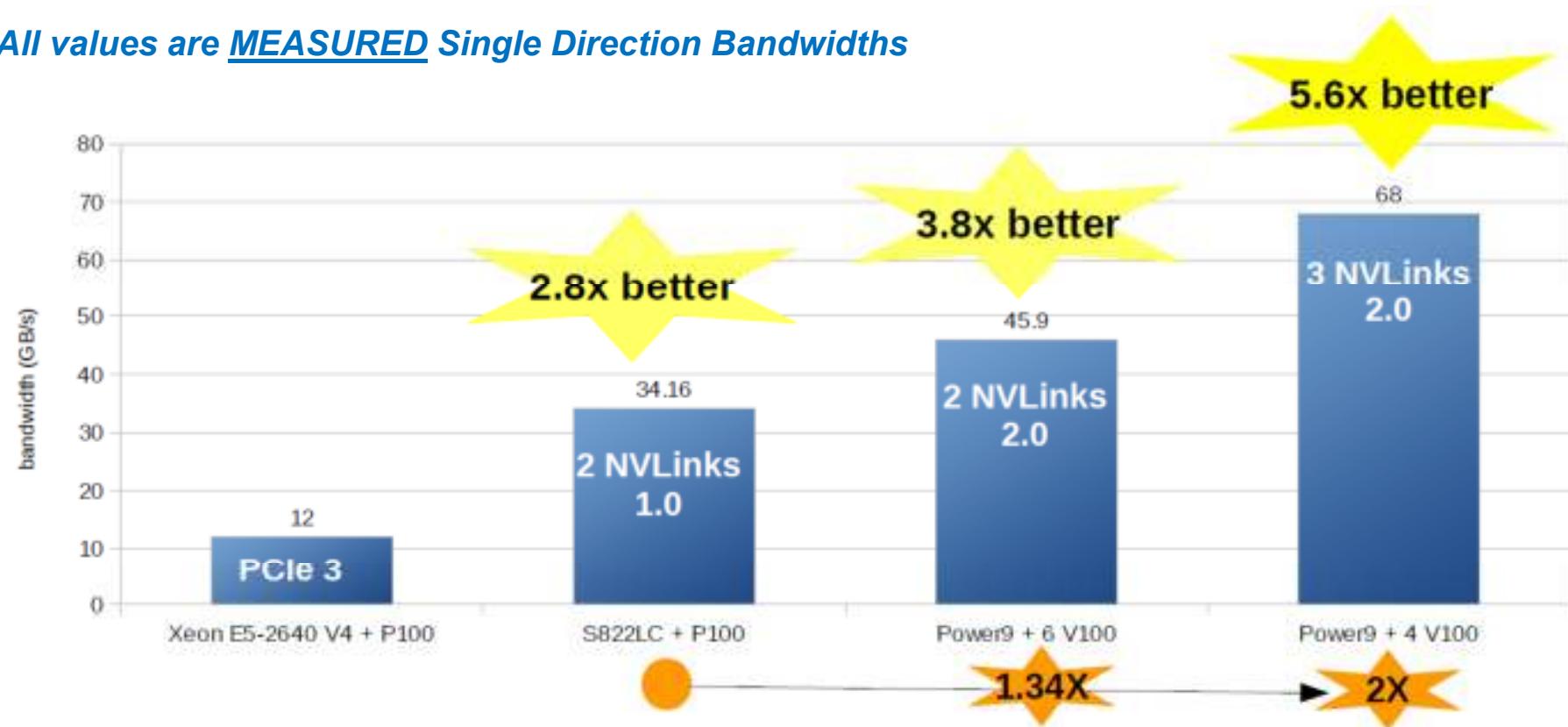
I/O Attachment Evolution in POWER HPC

X-Bus 8B @ 4.8GHz

X-Bus 4B @ 16GHz

PCIe Gen3 x8   PCIe Gen3 x16   PCIe Gen3 x8   PCIe Gen3 x16

PCIe Gen4 x16   PCIe Gen4 x8   PCIe Gen4 x16   PCIe Gen4 x4

PCIe Gen4 x8

IB-EDR NIC
Shared slot
CAPI
Mellanox "*Multi-Host Socket Direct*".
First industry implementation of Gen4 PCIe
Multi-host attachment of POWER9 and the Mellanox EDR-IB adapter
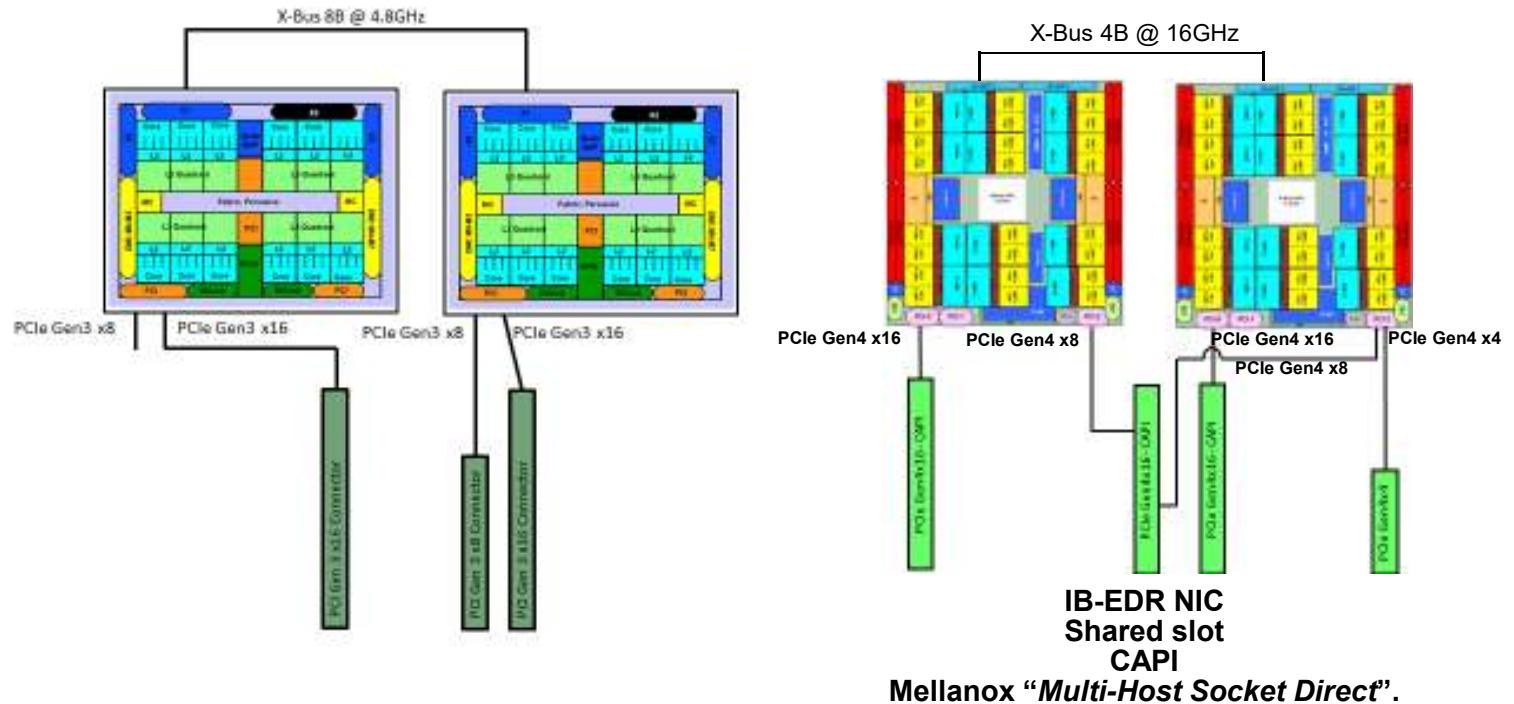
2016   **2017-2018**

InfiniBand EDR 100Gb/s – PCIe Gen 4 verses PCIe Gen 3

Comparing PCI Express Gen3 to Gen4
EDR Infiniband IB Dual Port Bidirectional Bandwidth Test
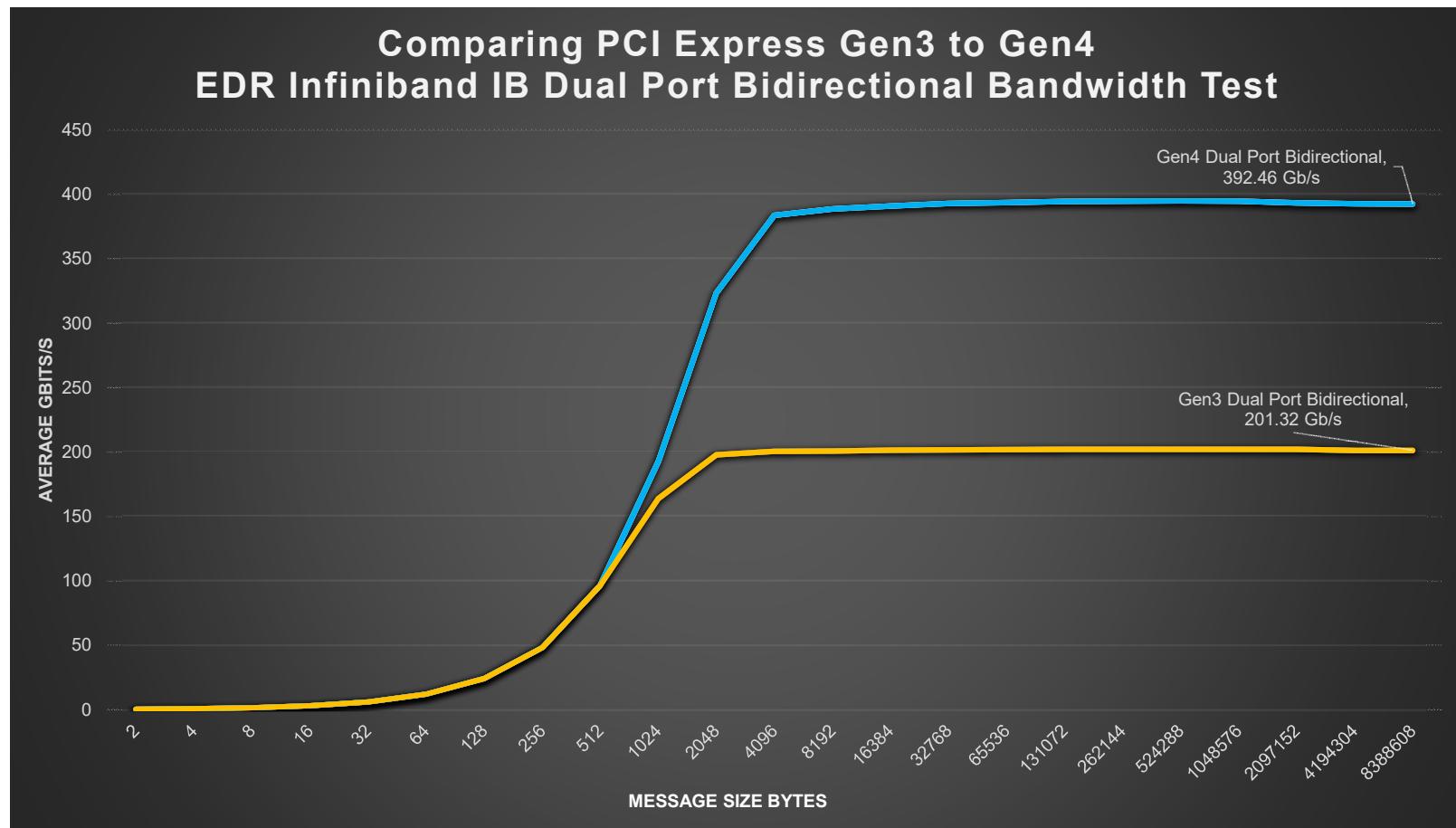
Front and Rear Views

Front

80mm CR Cooling Fans (4x)

Note: Front bezel is removed in this illustration

Service Indicators

Power Button

USB 3.0

SFF-4 Carrier (2X)
• SFF SATA HDD or SSD

PCIe Slot 1
• Gen4 x4 (x8 Connector)
• HHHL Slot

Rear

Power Supplies (2X)

PCIe Slot 2
• Gen4 Shared x8,x8
• HHHL Slot
• CAPI Enabled

PCIe Slot 3 & 4
• Gen4 x16
• HHHL Slot
• CAPI Enabled

Service Indicators

IPMI

VGA

1Gb Eth (2x)

USB 3.0

Water lines
(Option)
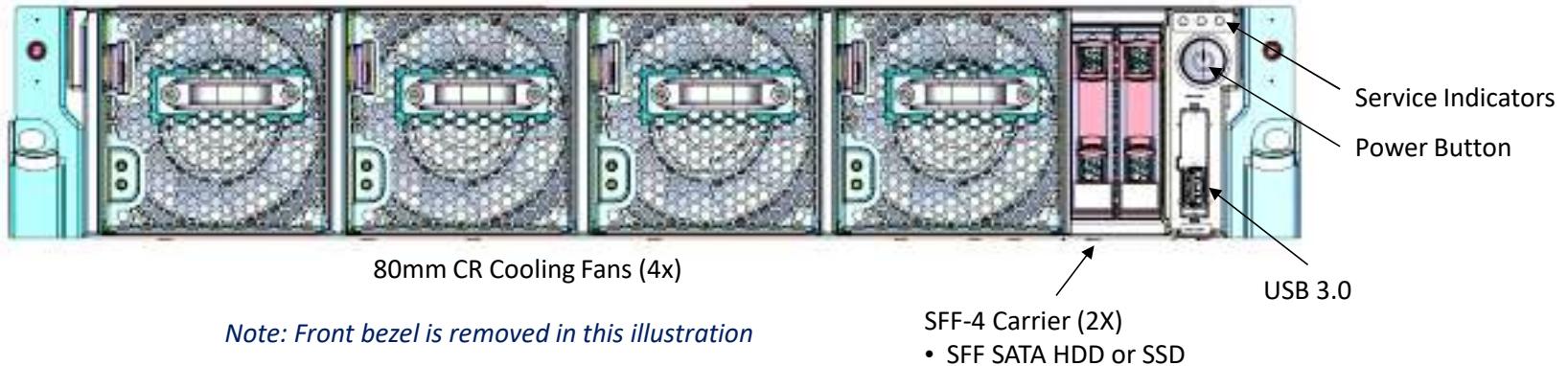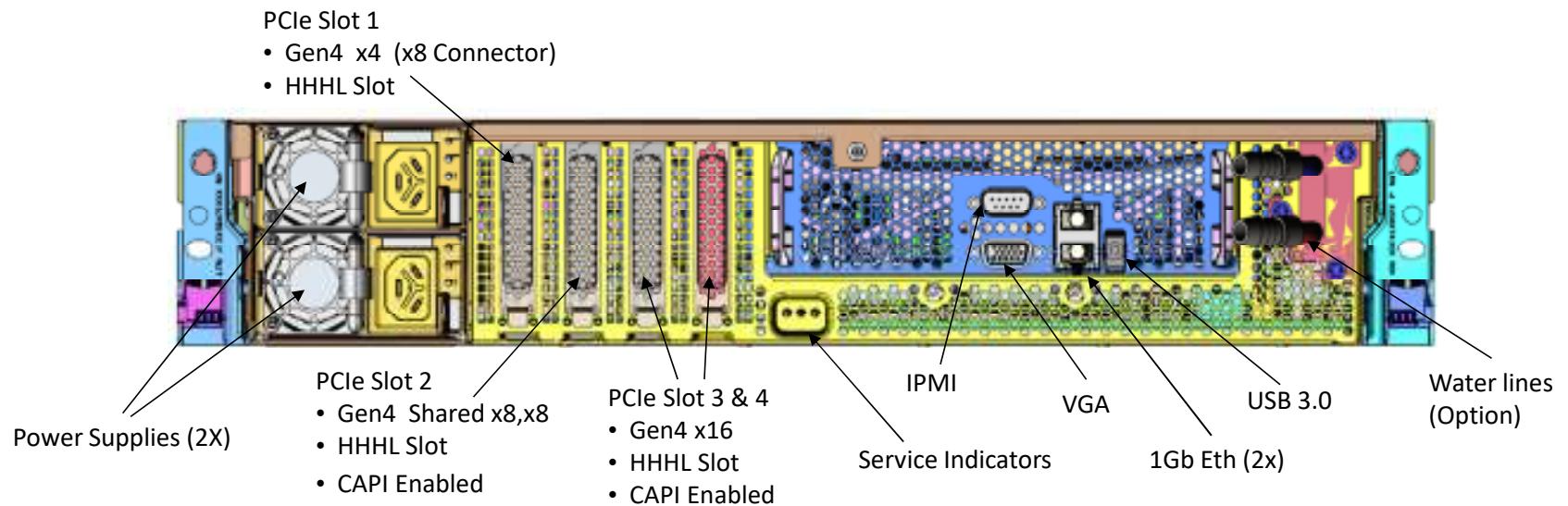
Rear View – AC Input

Unique AC input connection

- Requires unique AC cord to PDU

- Enables high voltage DC input to power supply

- Enables 480V Input voltage to the rack (future direction)
  - 277VAC distributed to power supplies

PDU

IEC C20

Rong Feng 203P-HP

AC922 Power Supply
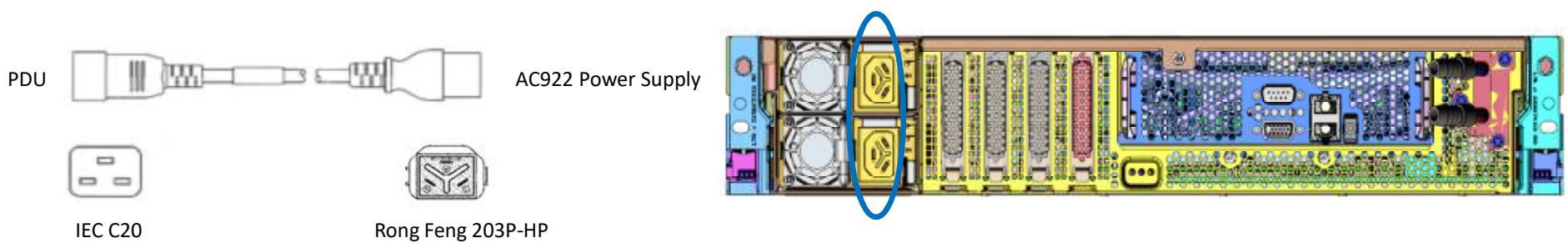
POWER AC922 Memory

- 16 direct attach industry standard DDR4 DIMMs are supported in Witherspoon
  - 8 DIMMs per P9
  - Maximum memory capacity increased to 2TB
  - Maximum memory bandwidth of 170GB/s per socket, 340GB/s for the enclosure
    - Measured Stream is 140GB/s per socket
- All memory DIMMs are installed in order to maximize memory bandwidth

DIMM Size	Maximum System Memory	Notes
16GB	256GB	DOE CORAL contract configuration LLNL
32GB	512GB	DOE CORAL contract configuration ORNL
64GB	1024GB	

CORAL

IBM, Mellanox, and NVIDIA awarded $325M U.S. Department of Energy's CORAL Supercomputers

CORAL: Leadership Class Supercomputers

5x – 10x HIGHER APP PERF THAN CURRENT SYSTEMS

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

June 2018 System Acceptance

IBM

IBM Systems

POWER AC922 Design – 6 GPU

Power Supplies (2x)
- 2200W
- 200VAC, 277VAC, 400VDC input

NVidia Volta GPU
- 3 per socket
- SXM2 form factor
- 300W
- NVLink 2.0
- Air/Water Cooled

PCIe slot (4x)
- Gen4 PCIe
- 2, x16 HHHL Adapter
- 1, Shared slot
- 1 x8 HHHL Adapter

Memory DIMM's (16x)
- 8 DDR4 IS DIMMs per socket
- 8, 16, 32,64, 128GB DIMMs

BMC Card
- IPMI
- 1 Gb Ethernet
- VGA
- 1 USB 3.0

Power 9 Processor (2x)
- 18, 22C water cooled
- 16, 20C air cooled
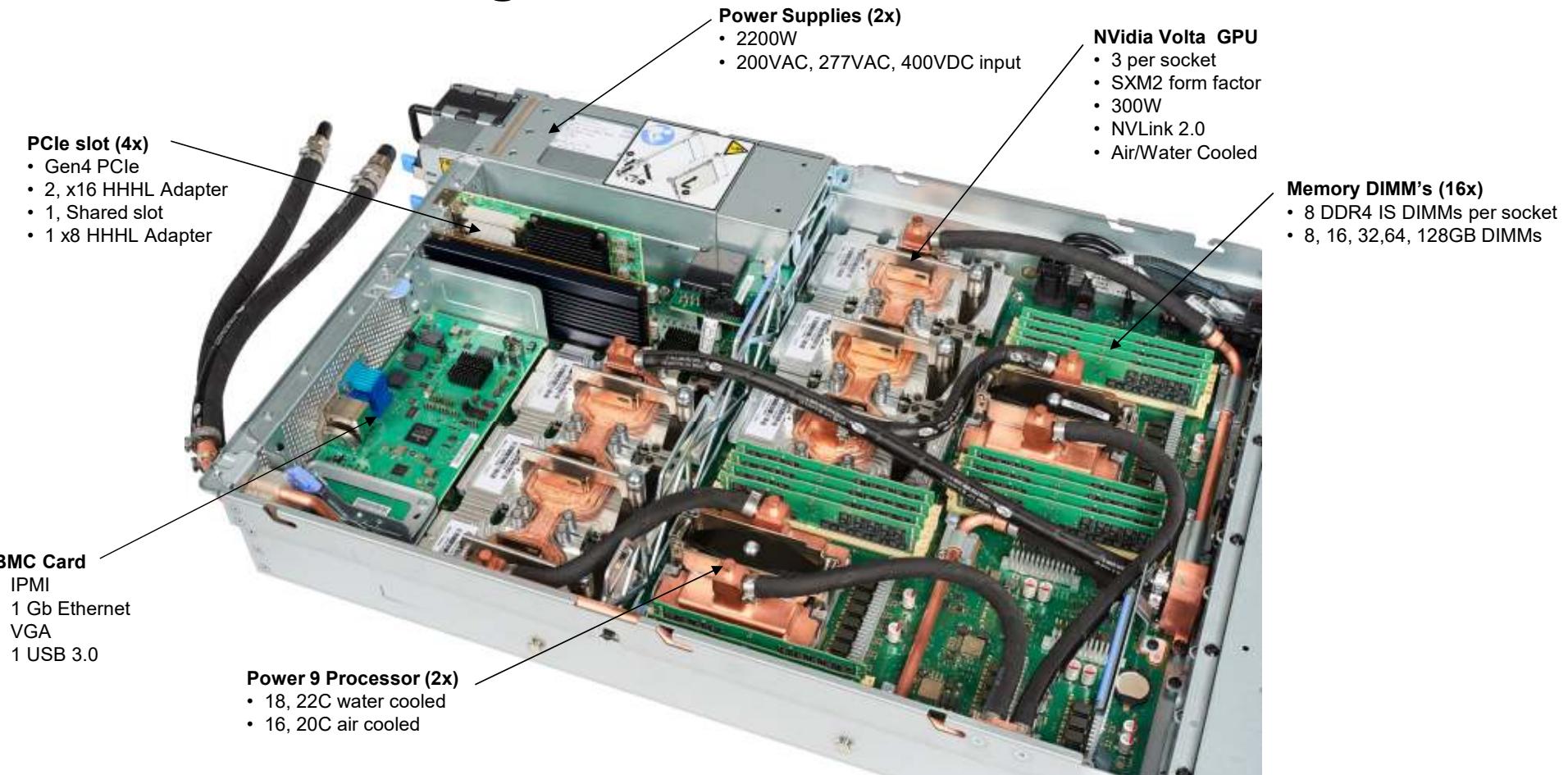
CORAL Installation at LLNL

CORAL Installation at ORNL

Thank you!

IBM

ibm.com/systems/hpc