OPEN
Compute Project

OCP U.S. SUMMIT 2017

Santa Clara, CA

Microsoft Project Olympus Hyperscale GPU Accelerator (HGX-1)

Siamak Tavallaei

- Principal Architect, Microsoft Azure Cloud Hardware Infrastructure

Robert Ober

- Tesla Chief Platform Architect, NVIDIA Corp.

Talk Outline

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR (HGX-1)

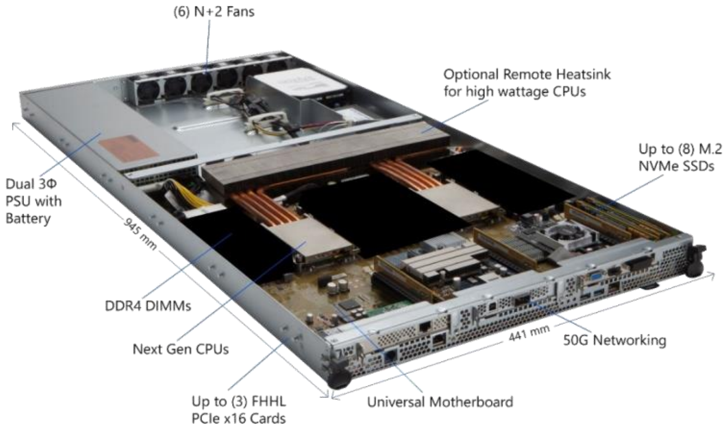
- Project Olympus Modular Architecture
- nVidia SXM2 with NVLink
- Collaborative Chassis Design with Ingrasys
- Enabling Components
- High-level Feature List
- Use cases
- Performance Advantages for various Workloads

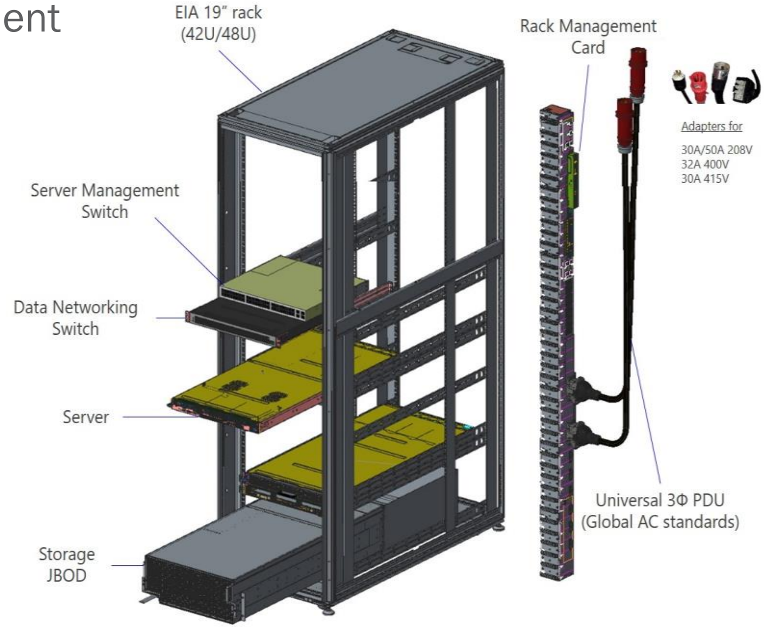
PROJECT OLYMPUS BASE

PROJECT OLYMPUS MODULAR ARCHITECTURE

Establishes a baseline for cloud-scale standard deployment

Datacenter management, power, cooling, performance

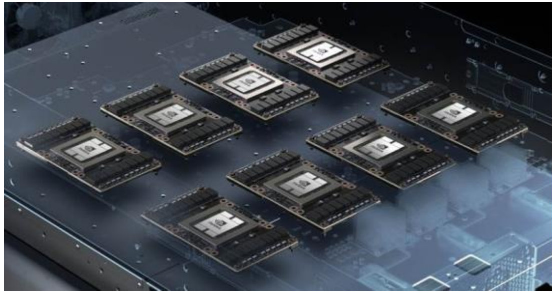




Industry-standard
Accelerated CLOUD COMPUTING

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR (HGX-1)

- Configurable and Flexible Accelerators
  - 8 x NVIDIA P100_SXM2 & NVLink
  - 8 x GPGPUs in PCIe Card Form Factor
- Expandable to Scale UP
  - From one to four Chassis
  - Internal PCIe Fabric Interconnect
- Scale Out via InfiniBand Fabric
- Host Head Node Options
  - 2S Project Olympus Server
  - 1S, 2S, 4S Server Head Nodes (eight x16 PCIe Links)
  - Up to 16 Head Nodes (sixteen x8 PCIe Links)



Industry-standard Accelerated CLOUD COMPUTING

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR (HGX-1) CHASSIS

- Configurable and Flexible Accelerators
  - 8 x NVIDIA P100_SXM2 & NVLink
  - 8 x GPGPUs in PCIe Card Form Factor
- Expandable to Scale UP
  - From one to four Chassis
  - Internal PCIe Fabric Interconnect
- Scale Out via InfiniBand Fabric
- Host Head Node Options
  - 2S Project Olympus Server
  - 1S, 2S, 4S Server Head Nodes (eight x16 PCIe Links)
  - Up to 16 Head Nodes (sixteen x8 PCIe Links)

ingrasys®

Microsoft

NVIDIA.

Industry-standard Accelerated CLOUD COMPUTING

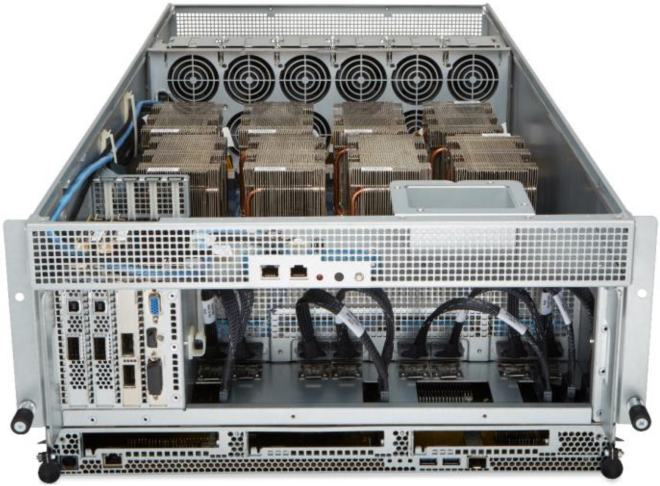
PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR (HGX-1)

- Configurable and Flexible Accelerators
  - 8 x NVIDIA P100_SXM2 & NVLink
  - 8 x GPGPUs in PCIe Card Form Factor
- Expandable to Scale UP (CNTK)
  - From one to four Chassis
  - Internal PCIe Fabric Interconnect
- Scale Out via InfiniBand Fabric (CNTK)
- Host Head Node Options
  - 2S Project Olympus Server
  - 1S, 2S, 4S Server Head Nodes (eight x16 PCIe Links)
  - Up to 16 Head Nodes (via sixteen x8 PCIe Links)

Industry-standard Accelerated CLOUD COMPUTING

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR (HGX-1)

- Configurable and Flexible Accelerators
  - 8 x NVIDIA P100_SXM2 & NVLink
  - 8 x GPGPUs in PCIe Card Form Factor
- Expandable to Scale UP (CNTK)
  - From one to four Chassis
  - Internal PCIe Fabric Interconnect
- Scale Out via InfiniBand Fabric (CNTK)
- Host Head Node Options
  - 2S Project Olympus Server
  - 1S, 2S, 4S Server Head Nodes (eight x16 PCIe Links)
  - Up to 16 Head Nodes (via sixteen x8 PCIe Links)



Industry-standard Accelerated CLOUD COMPUTING

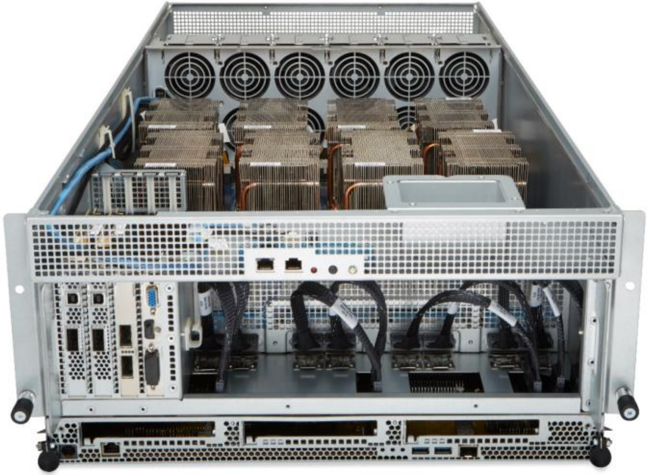
PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR (HGX-1)

- Configurable and Flexible Accelerators
  - 8 x NVIDIA P100_SXM2 & NVLink
  - 8 x GPGPUs in PCIe Card Form Factor
- Expandable to Scale UP (CNTK)
  - From one to four Chassis
  - Internal PCIe Fabric Interconnect
- Scale Out via InfiniBand Fabric (CNTK)
- Host Head Node Options
  - 2S Project Olympus Server
  - 1S, 2S, 4S Server Head Nodes (eight x16 PCIe Links)
  - Up to 16 Head Nodes (via sixteen x8 PCIe Links)

Video Transcoding

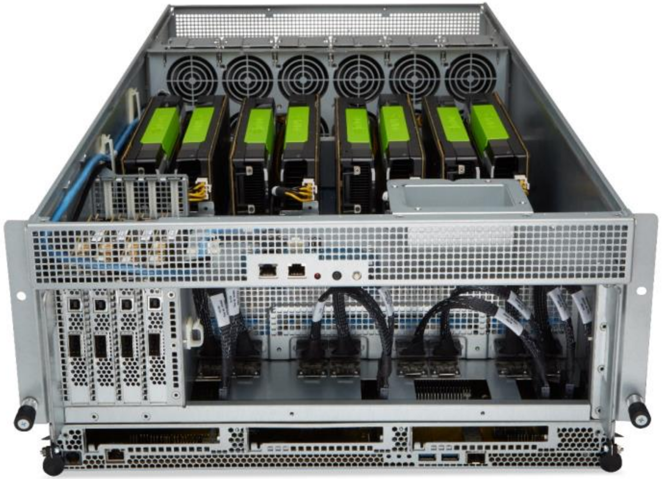
HPC

DNN



PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS

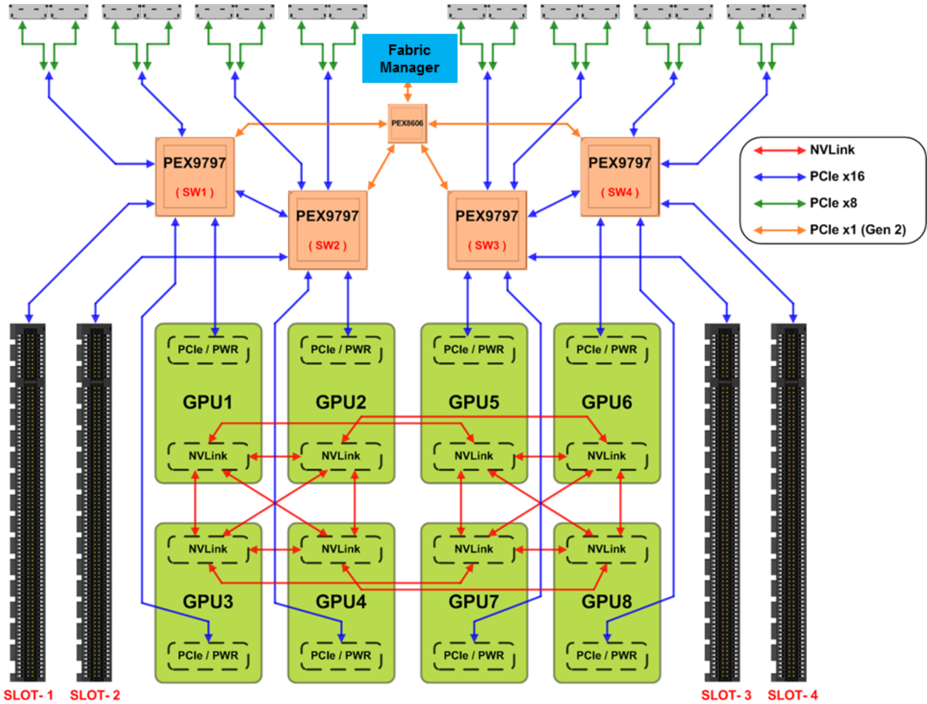
- Configurable and Flexible Accelerators
  - 8 x NVIDIA P100_SXM2 & NVLink
  - 8 x GPGPUs in PCIe Card Form Factor
- Expandable to Scale UP
  - From one to four Chassis
  - Internal PCIe Fabric Interconnect
- Scale Out via InfiniBand Fabric
- Host Head Node Options
  - 2S Project Olympus Server
  - 1S, 2S, 4S Server Head Nodes (eight x16 PCIe Links)
  - Up to 16 Head Nodes (sixteen x8 PCIe Links)



Enabling Components

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS

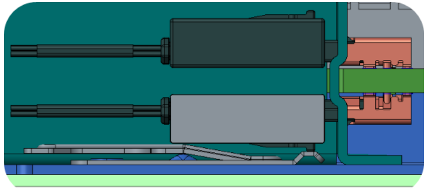
- Flexible PCIe Interconnect Topology
- GPGPU-to-Host via high-BW PCIe Links
- Peer-to-peer without Host interaction
  - GPGPU peer-to-peer via NVLink
  - GPGPU peer-to-peer to IB NICs via x16 PCIe

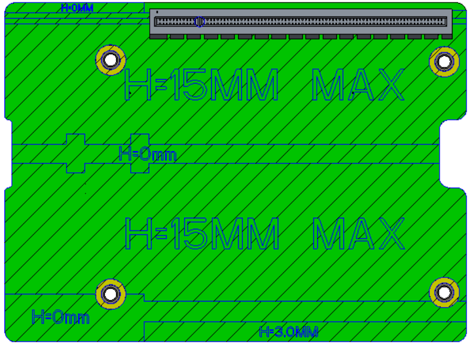


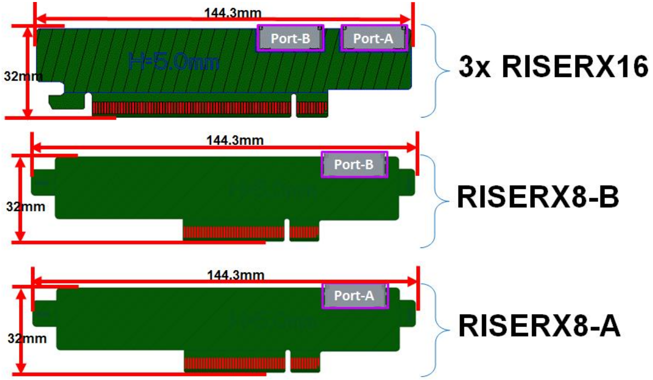
Enabling Components

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS

- Riser Boards
  - Plug into the Server Head Node
  - x16, x8 Type-A, x8 Type-B
- X8 OCuLink Cable/Connector
  - For Chassis-to-Chassis Interconnect
- Mezzanines
  - MEZZ1x16
  - Various PCIe Slot Configs.



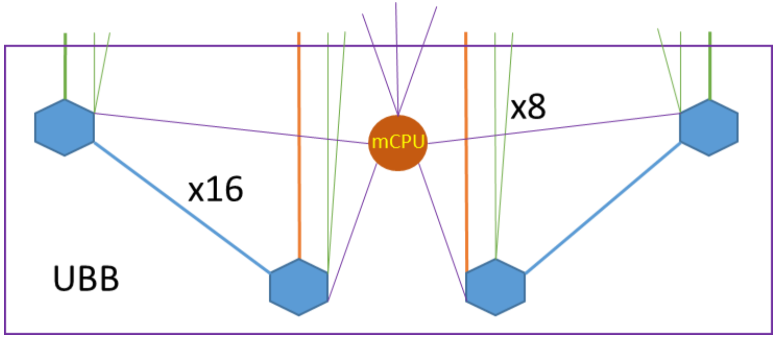


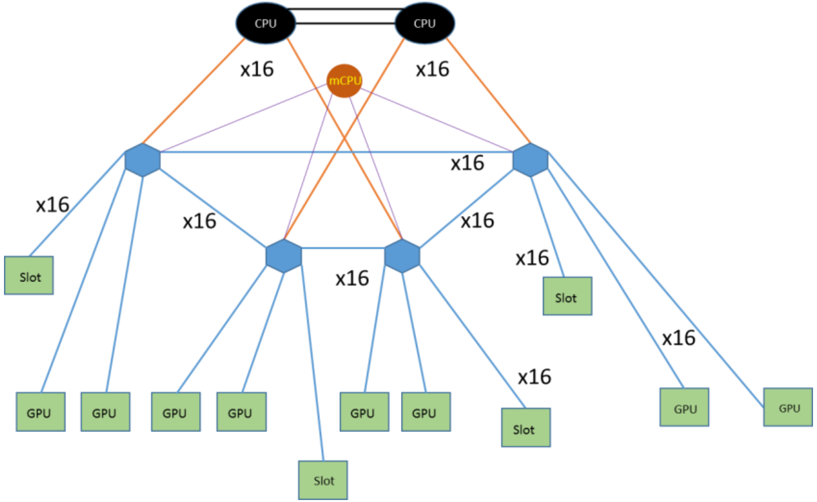


Enabling Components

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS

- Flexible PCIe Interconnect Topology
- Great peer-to-peer bandwidth
- Extensible as Chassis-to-Chassis Interconnect

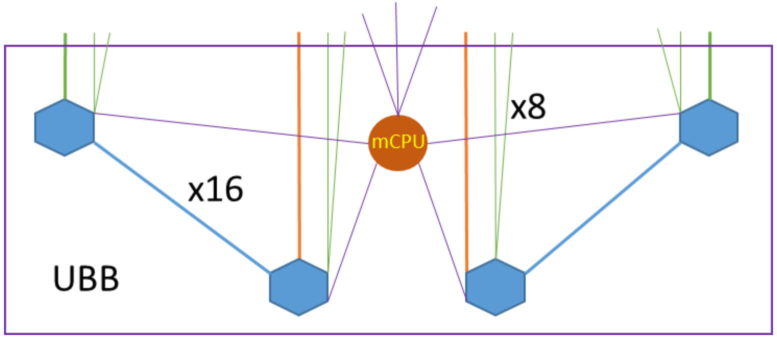


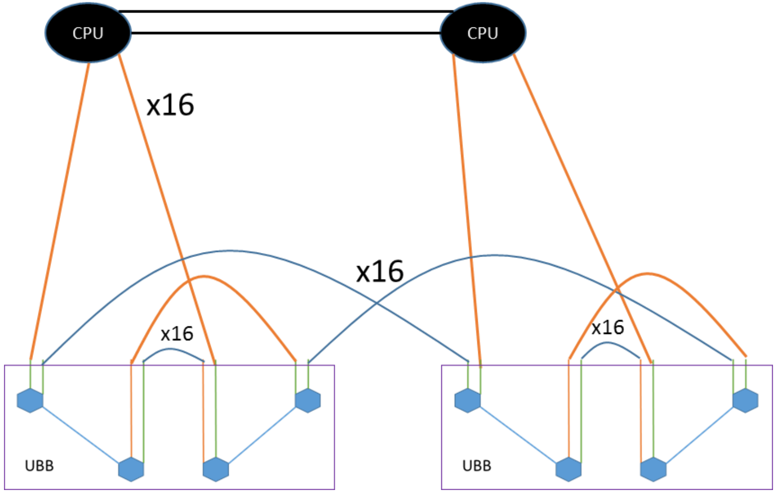


Enabling Components

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS

- Flexible PCIe Interconnect Topology
- Great peer-to-peer bandwidth
- Extensible as Chassis-to-Chassis Interconnect

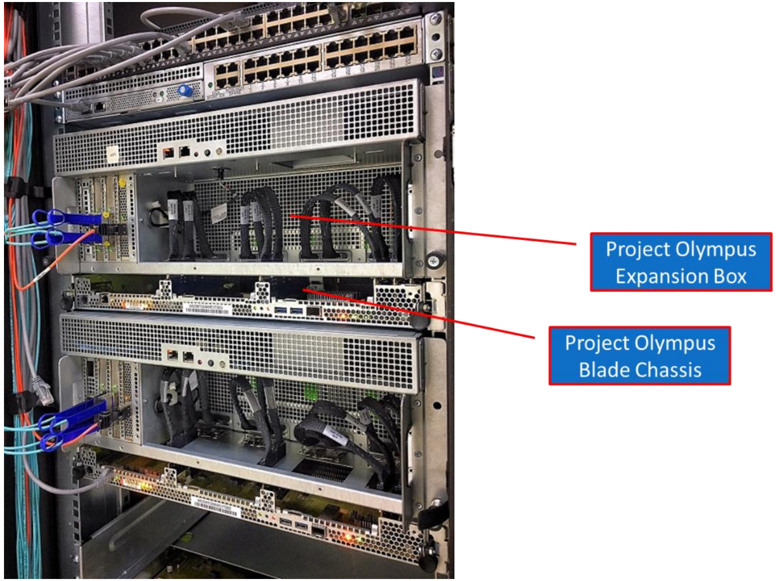




Enabling Components

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS

- Flexible Inter-Chassis PCIe Interconnect Topology



Specification Highlights

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS HIGHLIGHTS

- 4U Chassis Form Factor
- Six 1600W PSUs (N+N)
- Twelve Fans (N+2)
- Sixteen x8 OCuLink Cables for External PCIe Interconnect (8 x16)
- 4 x FH¾L PCIe Cards + 8 x 300W GPGPUs (SXM2 or double-width FH¾L PCIe Form Factors)
- Node Management (AST2500/2400 BMC family, 1GbE Link to Rack Manager)
- Rack Management Sideband: 2x RJ45 Ports for OoB Power Management
- PCIe Fabric Management for multi-Chassis Configurations, multi-Hosting, and IO-Sharing

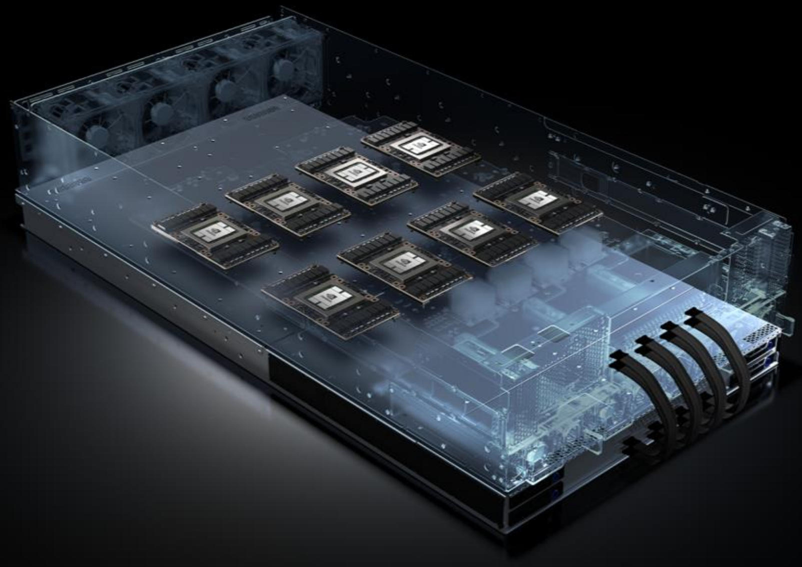
Specification Highlights

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS HIGHLIGHTS

- Flexible choice of GPGPUs
  - Eight Pascal P100 SXM2_NVLink
  - Various GPGPUs in double-width, 300W PCIe Card form factor
    - Such as P100, P40, P4, M40, K80, M60 etc.
- High PCIe Bandwidth to Host Memory and for peer-to-peer
- Up to 4 PCIe-interconnected Chassis (with a dedicated PCIe Fabric Management Network)

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR CHASSIS

Use Cases & Performance Advantage For Various Workloads



PROJECT OLYMPUS HGX-1 HYPERSCALE GPU ACCELERATOR

PARTNERSHIP + INTEROPERABILITY

CLOUD CHALLENGES

- 1 SKU, Multiple Instances
- Integration into Existing Datacenter

INSTANCES

- Granular, Latency Sensitive
- High Throughput Batch
- HPC: different CPU:GPU ratios
- DevOps / Development
- Production Deployment

Microsoft NVIDIA. ingrasys®

Project Olympus HGX-1

Hyperscale GPU Accelerator

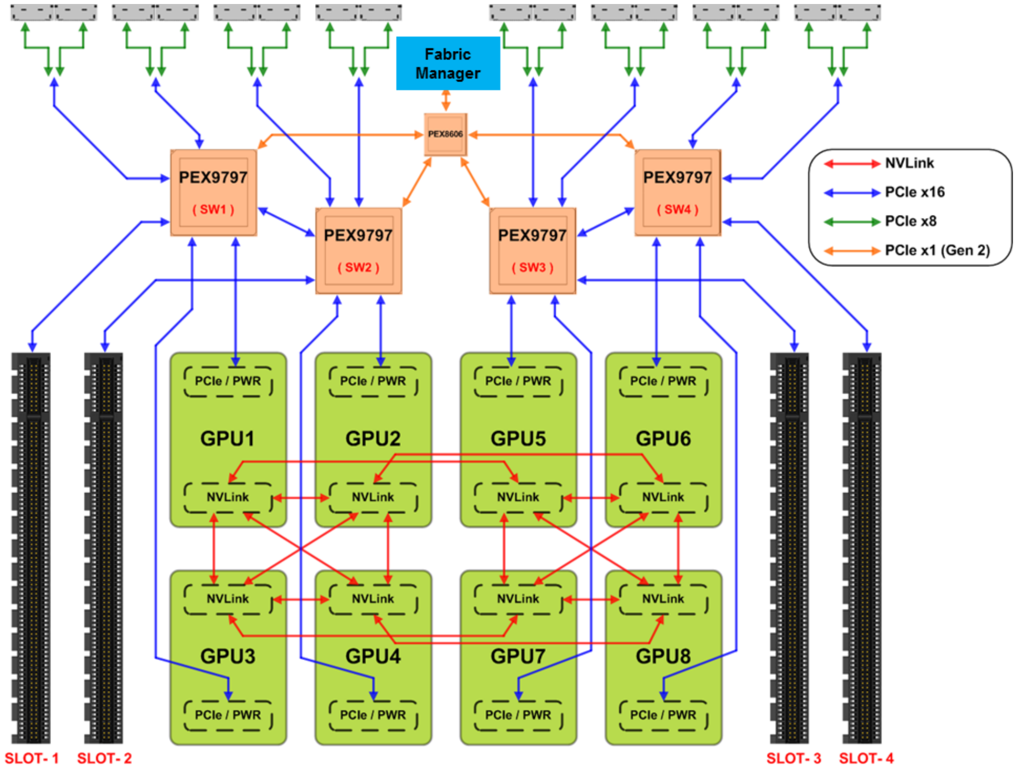
Configurable PCIe Cable to host + Expansion slots

NVIDIA P100 GPU

NVLink Hybrid Cube Mesh Fabric

20 Gbyte/sec per link Duplex

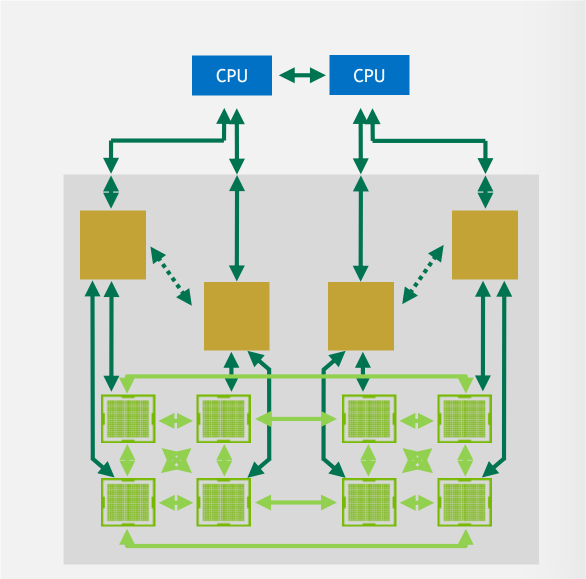
Adapters for other GPUs



DEEP LEARNING

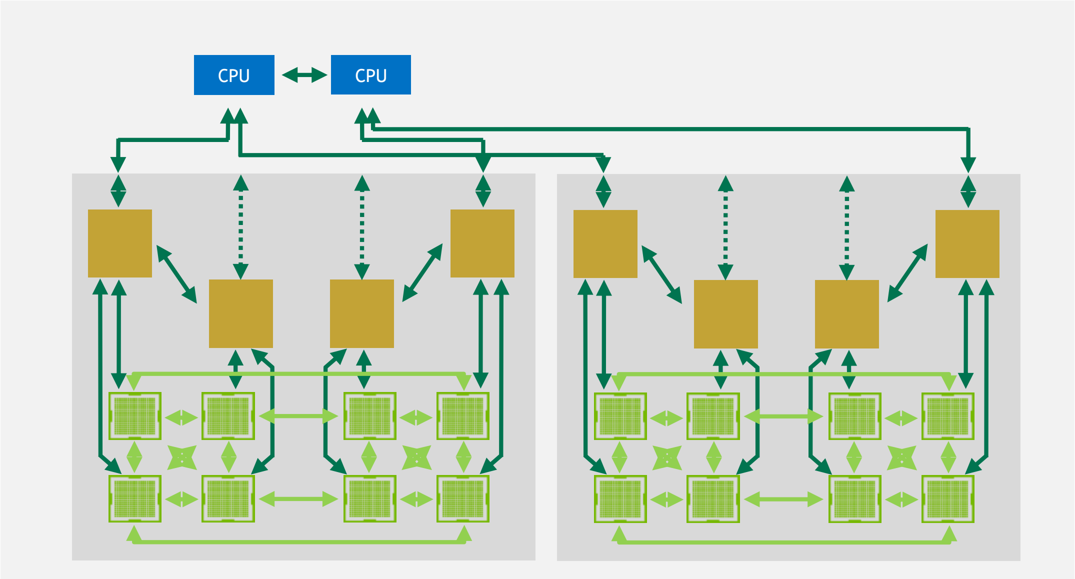
2 CPU : 8 GPU

8x P100 SXM2 | 4x x16 PCIe



2 CPU : 16 GPU

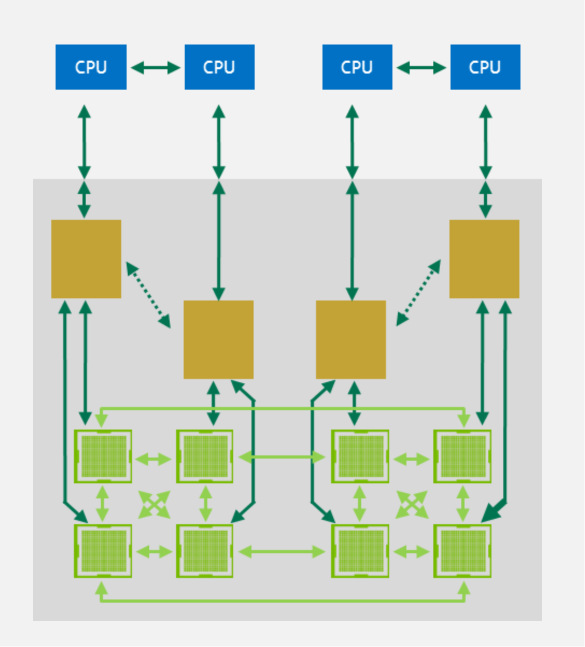
16x P100 SXM2 | 4x x16 PCIe



HPC

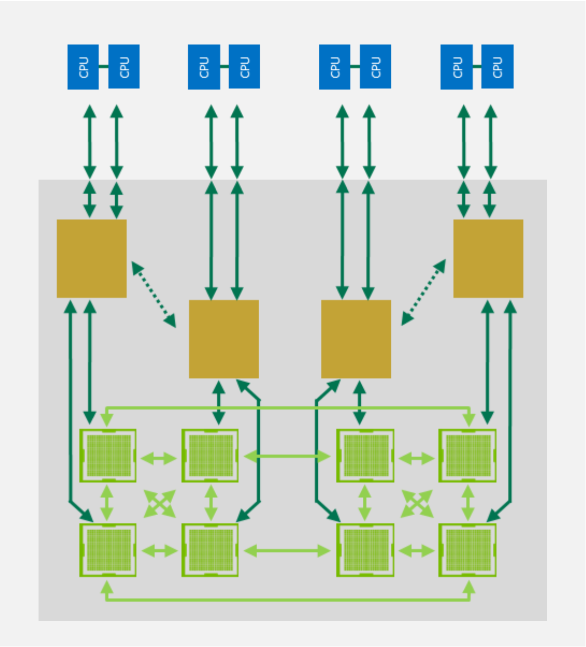
4 CPU : 8 GPU

8x P100 SXM2 | 4x x16 PCIe

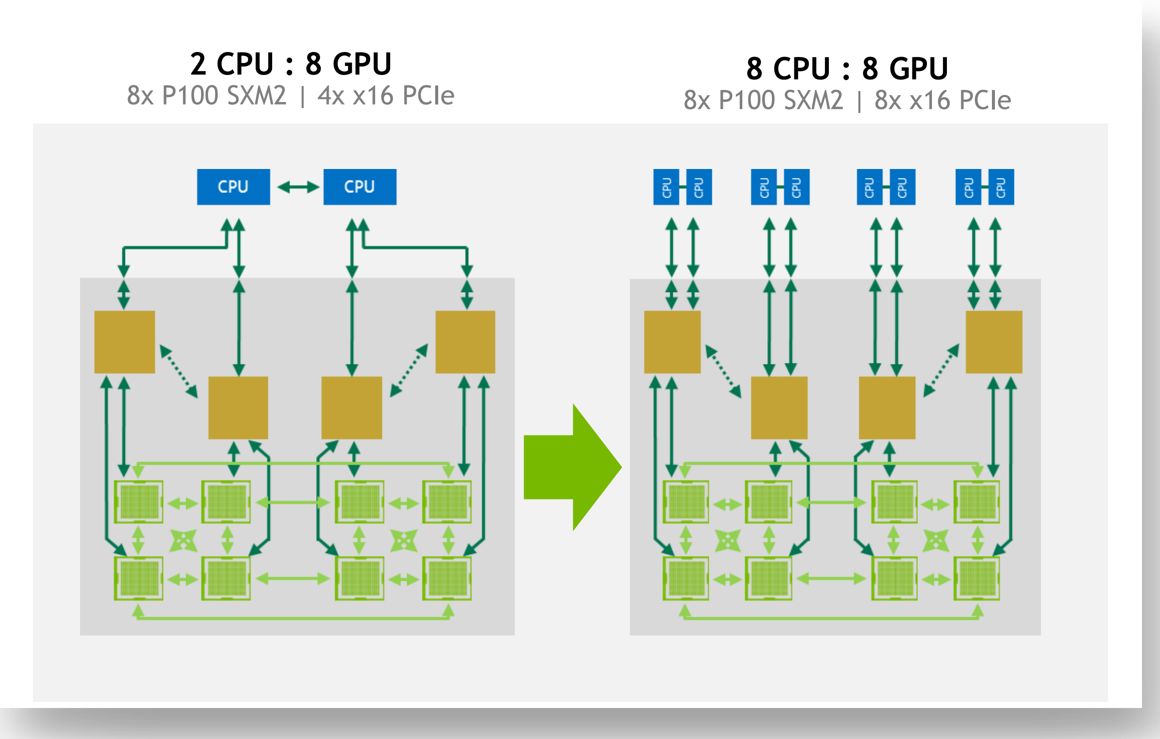


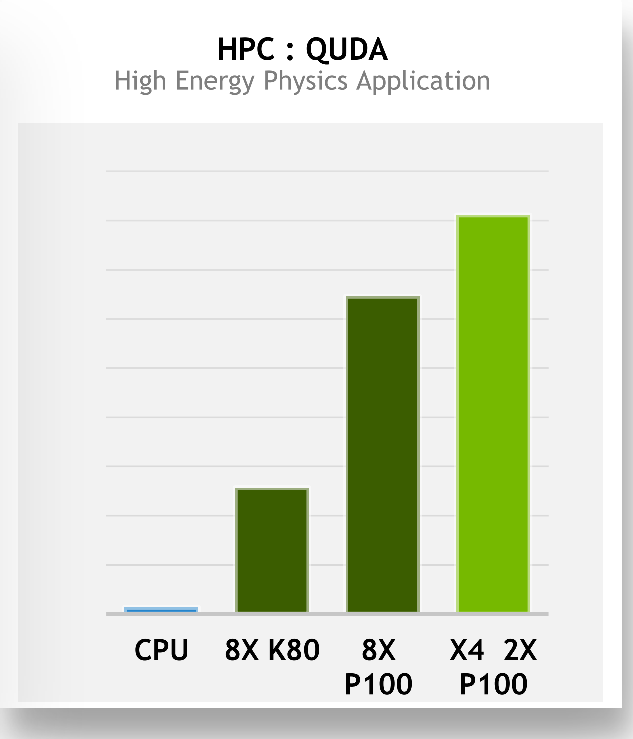
8 CPU : 8 GPU

8x P100 SXM2 | 8x x16 PCIe



WORKLOAD OPTIMIZED PERFORMANCE





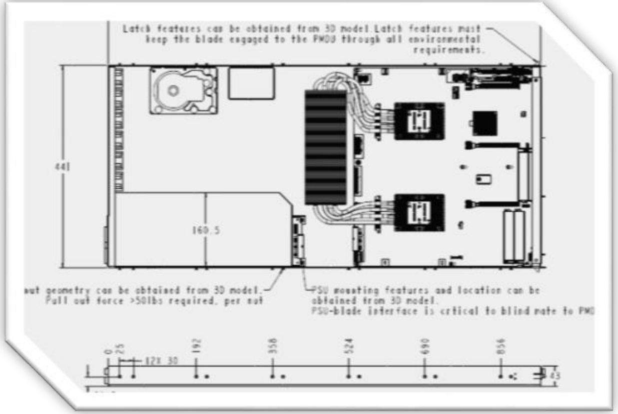
Summary

PROJECT OLYMPUS HYPERSCALE GPU ACCELERATOR (HGX-1)

- To augment the performance of Project Olympus Servers, we have collaborated with Ingrasys and nVidia on a PCIe Expansion Box we call:
  - Project Olympus Hyperscale GPU Accelerator (HGX-1)
- We are contributing this specification and its associated product/design to OCP

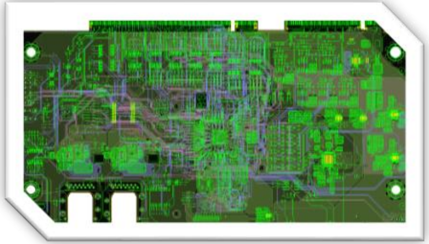
OCP Contributions

Mechanical CAD

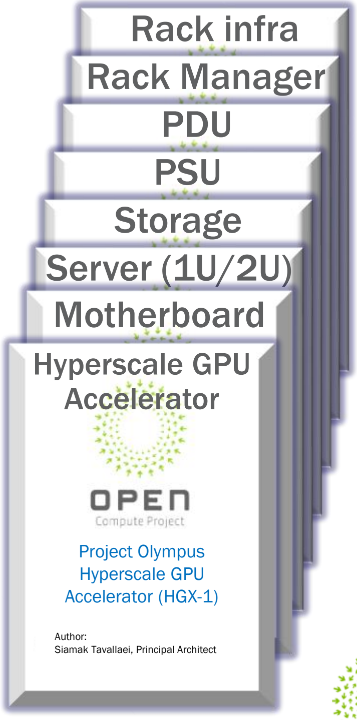


Latch features can be obtained from 3D model.Latch features must keep the blade engaged to the PMDU through all environmental requirements.
441
160.5
ut geometry can be obtained from 3D model. Pull out force >50lbs required, per nut
PSU mounting features and location can be obtained from 3D model. PSU-blade interface is critical to blind mate to PMD
0
25
12X 30
192
358
524
690
856
43

Schematics & Board Files



Specifications



Rack infra

Rack Manager

PDU

PSU

Storage

Server (1U/2U)

Motherboard

Hyperscale GPU Accelerator

OPEN
Compute Project

Project Olympus
Hyperscale GPU
Accelerator (HGX-1)

Author:
Siamak Tavallaei, Principal Architect

https://github.com/opencomputeproject/Project_Olympus

Siamak Tavallaei

Principal Architect, Microsoft

Siamak Tavallaei is a Principal Architect at Microsoft's Azure division. Collaborating with industry partners, he drives a number of initiatives in research, design, and deployment of hardware for Microsoft's cloud-scale services such as Azure, Bing, Office 365, Exchange, and SQL across a global datacenter footprint. With over 30 patents and 27 years of computer industry experience, he has been instrumental in development and evolution of innovative multi-processor servers and technology initiatives in areas of storage and memory hierarchy as well as heterogeneous, distributed computing. He held the rank of Principal Member Technical Staff at Compaq and was a Distinguished Technologist at Hewlett-Packard before joining Microsoft. He is interested in Big Compute, Big Data, and Artificial Intelligence solutions based on distributed, heterogeneous, accelerated, and energy-efficient computing. His current focus is the optimization of large-scale, mega-datacenters for general-purpose computing and accelerated, tightly-connected, problem-solving machines built on collaborative designs of hardware, software, and management.

Rob Ober

Chief Platform Architect, Tesla Datacenter Products

At NVIDIA Rob works with hyperscales like Microsoft to define the Tesla GPU platforms. Previously Rob was Senior Fellow at SanDisk, FusionIO, LSI, AMD and Chief Architect at Infineon. Rob has more than 30 years experience in computer architecture, has more than 40 international patents in processors and systems, and has a degree in Systems Design Engineering from the University of Waterloo in Canada.

OPEN
Compute Project