Microsoft Project Olympus Hyperscale GPU Accelerator (HGX-1)

A Tech Blog

Author: Siamak Tavallaei, CSI, Azure Cloud
5/26/17

1 Introduction

We announced Microsoft Project Olympus Hyperscale GPU Accelerator (HGX-1) at Open Compute Summit (OCP) in March 2017 and presented it to HPC community at GPU Tech Con (GTC17) in May 2017. Its flexibility and expandability attracted much interest within the Artificial Intelligence (AI) community. In response to requests for detailed information, this article outlines HGX-1 architecture after a brief introduction to the state-of-the-art computing landscape.

There has been much interest in Deep Learning (DL) in the industry, and our technology partners are responding with various hardware accelerators and software framework innovations. While CPUs are suitable to run some Machine Learning applications, DL requires specialized processing elements such as GPGPUs, FPGAs, or ASICs to reduce the execution time and hardware cost to feasible levels.

At Microsoft, we have deployed Azure GPU and internal clusters for DL, and we have gained much insight into customer use cases and performance sensitivity of various workloads to CPU, GPGPU, PCIe Topology, Network throughput, and Storage. Based on this insight, HGX-1 provides a flexible and extensible platform for accelerated computing which scales up and scales out to provide differentiated value for Azure datacenters.

Based on Microsoft Project Olympus architecture, HGX-1 benefits from streamlined management suitable for Microsoft Datacenters at Cloud-scale. It is the building block for an *Elasticale*© architecture for late-binding choice of pooled or distributed, heterogenous resources at scale.

Providing one GPU at a time to VMs requires one PCIe Link per GPU; while, providing all GPUs in the server to one large Virtual Machine (VM) requires low-latency, high-bandwidth peer-to-peer interconnect between GPUs. With commercially available hardware, we must deploy different platforms to optimize for different customers. With a flexible IO-to-Host-Memory interconnect depicted in Figure 1, HGX-1 obviates the painful trade-off between good Host bandwidth and good peer-to-peer BW. HGX-1 provides our management software the flexibility to choose. Additionally, based on a Universal Building Block (UBB), HGX-1 architecture provides a multi-Chassis solution to scale up to four Chassis of eight high-powered Accelerators along with the related PCIe devices for storage and networking.

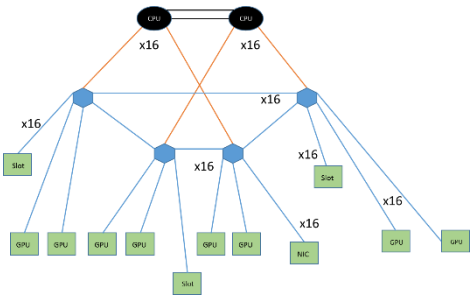


Figure 1 PCIe Topology of single-Chassis HGX-1

2 Landscape

Continued silicon advancements in compute and high-bandwidth memory along with the recent prevalence of labeled data have accelerated the growth of lucrative businesses benefiting from artificial Intelligence (AI), Deep Machine Learning (DL), HPC, and Video Transcoding.

Contemporary CPUs can run DL frameworks, tools, and libraries for *Inference* and small DL *Training* jobs; however, to accelerate *Training* jobs with large, deep neural networks (DNN) and datasets, many-core GPUs and sometimes many-GPU clusters are required. *Training* Jobs are typically throughput-computing and not latency-sensitive. With large batch sizes running on multi-GPU nodes in a datacenter, they

typically train once (several days), re-train based on new data (daily), and continue training based on additional fresh data.

Algorithms for Neural Networks (NN) include many parameters which can run independently. To accelerate computing, many algorithms take advantage of parallel and distributed computing. Various stages of the solution require different computational, network, and storage optimizations which expose a different bottleneck.

The interconnect bandwidth between GPUs limit massive parallelism. nVidia's P100 Pascal and the recently announced V100 GPUs provide a high-bandwidth private peer-to-peer network called NVLink. Such network along with large register files and high-BW memory place nVidia at an advantage. Increasing the GPGPU cluster size beyond the scope of NVLink requires high-bandwidth, low-latency networks such as InfiniBand or PCIe Switch Fabric.

While, the need for AI is growing, solutions are emerging to address different challenges. For *Training* Jobs, researchers and programmers deploy different processing elements, algorithms, toolchains, frameworks, and platforms to optimize model accuracy vs. execution time. For example, Microsoft Cognitive Toolkit—previously known as CNTK (http://www.cntk.ai/), is an open-source toolkit that trains DL algorithms to learn like the human brain.

Different compute elements are also emerging to fulfil the need. General-purpose and many-core CPUs provide a uniform, straightforward software programing model. CPUs generally provide much memory capacity at reasonable bandwidth. GPGPUs are multi-core processing elements specialized for fused Matrix-Vector Multiplication-and-Add. They usually accompany CPUs to benefit from Host Memory and to augment CPU's computation capability for complex instructions, analysis, or visualization. This CPU/GPU coupling requires a heterogeneous software programming model such as CUDA, OpenGL, or OpenCL. AMD and nVidia actively produce GPGPUs to serve the marketplace. Since DL algorithms do not rely on high-precision variables with large dynamic range to converge, instead of using 64-bit double-precision floating point (FL64), specialized GPUs optimize 32-bit single-precision (FL32) and 16-bit half-precision floating point operations to increase throughput.

We have developed a large GPU cluster based on nVidia M40 for DL Training. It reduces training time to ~1 day vs. 10 days for a typical AlexNet training.

CPUs, Xeon-Phi, and GPUs apply the traditional processing elements (PE) based on Von Neumann Computing paradigm which Fetch Code, READ Data, Execute, Store Result, and Repeat. Von Neumann computing relies on efficient data movement from/to Memory. Large DNNs require careful memory and network bandwidth optimization to reduce data communication overhead. To optimize for DNNs on large datasets with high processing throughput, several startup companies are exploring non-Von Neumann Accelerator engines using ASICs or FPGAs to execute in-memory computation, flowgraph, or fabric-computing. They stream data through a successive set of Execution Engines to reduce communication overhead.

Intel/Altera and Xilinx continue to develop high-performing FPGA solutions which promise to provide cost- and power-optimized alternatives for DL.

While, *Training* jobs drive the current need for GPU clusters, FPGAs contend for *Inference*, and ASICs promise to be the long-term solution.

As customers shift their applications and workloads from Enterprise to the Cloud, they examine and benefit from Platform-as-a-Service (PaaS) offerings to rent hardware and Software-as-a-Service (SaaS) for turn-key solutions in the Cloud.

In the process of completing a DL job in a datacenter, Data travels from its collection/sampling point to the Cloud and into a DL Cluster. Then, it is Filtered and Transformed to fit the Format of the Framework or Model before it passes through several Execution and post-Processing steps to provide Insights. This process requires an efficient and comprehensive networking, storage, and compute hardware platform.

To address the promising GPGPU market, several OEMs have developed specialized servers. Most of them have focused on compute density and require more than 30kW per Rack.

In addition to Machine Learning, Azure customers have asked for solutions optimized for HPC applications using FL32 and FL64 and for solutions for remote visualization.

Azure customers are diverse; they need flexibility and various balance of CPU and GPU ratios. To allow the Cloud to scale out to meet ever-growing customer demand, we continue to identify and reduce performance bottlenecks. We need good datacenter-ready hardware to increase Cloud-scale flexibility and fungibility within the physical space, power, and cooling requirements of large datacenters. We need the flexibility to implement different Head-nodes and the freedom to choose the number of GPUs per Virtual Machine (VM).

The breadth of the need and the solutions to meet them illustrate why we needed to invest in a flexible platform. It is ideal to produce one hardware SKU which can handle various configurations mostly via software. While, homogeneity reduces complexity, heterogeneity provides flexibility of choice to combat uncertainty.

3 HGX-1: A Datacenter-Ready Flexible and Extensible Platform

Designed for 19" EIA Racks, Project Olympus Server is the baseline compute element which meets critical Cloud-scale datacenter standards of power, cooling, management, and mainstream performance. In collaboration with industry partners, to augment the performance and storage capabilities of Project Olympus Server, we have built a series of expansion boxes to serve bulk storage (DX-88), high-performing SSDs (FX-16), and various accelerators. In particular, we built a PCIe Expansion Box (EB) for Project Olympus Hyperscale GPU Accelerator (HGX-1) in collaboration with Ingrasys and nVidia.

With many cores optimized for 64-bit, 32-bit, and 16-bit floating point operations, nVidia's Pascal architecture is suitable for Deep Learning Neural Networks (DNN) and provides much advantage over CPU-only servers. HGX-1 accelerates HPC and Artificial Intelligence (AI) applications and frameworks using Pascal P100 and V100 SXM2 modules with NVLink as a high-bandwidth interconnect (20 and 25 GB/s) for efficient peer-to-peer communication.

3.1 HGX-1 Attributes

The underlying attributes of HGX-1 architecture include:

- Purpose-built hardware platform: compute, fabric, fabric management, storage, network
- Disaggregated mainstream compute from accelerated compute to provide flexibility of choice
- Compute requirement: CPUs, GPGPUs, ASICs, …
- Bandwidth optimized (GPU-to-Host and GPU-to-GPU peer-to-peer)

- High Host-to-GPU bandwidth (oversubscription ratio can be 1:1)
- Designed for scale-up, scale-out based on a Universal Building Block (UBB)
- Optimized for highly parallel workloads (scale-up within one and up to four chassis, scale-out in clusters of Racks to include ~1000 GPGPUs)
- Reduce network bandwidth bottleneck via a private PCIe Switching Fabric (GPU_Direct, RDMA, GPU-to-IB per Switch)
- A distributed PCIe Switch Fabric: four six-ported Switch per Chassis (sixteen, 96-Lane PCIe Switches in total)
- CPU Root Complex as an extension of the PCIe Switch Fabric (Not traversing inter-CPU Link)
- No bandwidth oversubscription for peer-to-peer (reduced blocking, reduced hop count for inter- and intra-chassis traffic)
- Multi-tenant support via Virtualization and Physicalization (rightsizing hardware elements for the needs of Virtual Machines)
- Hardware Isolation through Switch Fabric
- Networks: front-end (Ethernet datacenter network to connect to public domains to bring data and produce results); back-end (performance domain using InfiniBand, PCIe Switch Fabric, and NVLink)
- Project Olympus Server as the baseline to benefit from datacenter-readiness: Node Manager, Rack Manager, balanced 3-phase dual-feed Power
- Ample power: 6x1600W (4800W of N+N power; 3200W of 3+2+1 power)
- Ample cooling: 12 Fans to keep 8 x 300W + 4 x 75W slots cool
- General-purpose slots: InfiniBand, SmartNIC, M.2 Farm, HBAs, etc.
- Multi-Chassis interconnect: low-latency PCIe Switch Fabric to scale up to 32 GPGPUs
- Mezzanine to PCIe Connector to allow other use cases: different Mezzanine, Host, GPGPU

3.2 HGX-1 Feature Set

The high-level feature set of HGX-1 include:

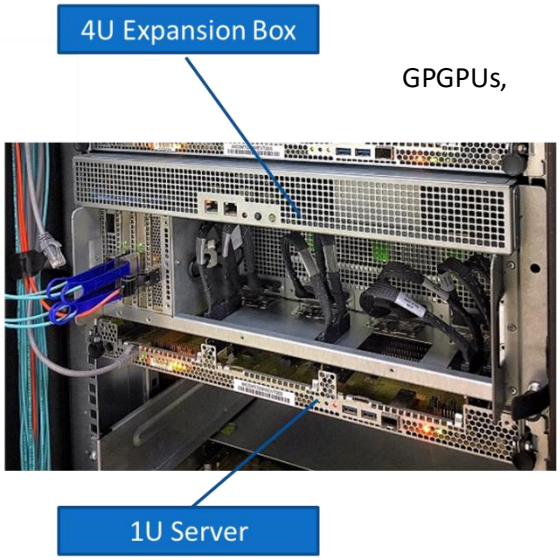
- 4U Chassis Form Factor
- Six 1600W PSUs (N+N)
- Twelve Fans (N+2)
- Four 96-Lane PCIe Switches from Broadcom (PEX9797)
- Eight bifurcatable x16 Links (Cables for External PCIe Interconnect: 8x16 or 16x8)
- 4 x FH¾L PCIe Cards + 8 x 300W GPGPUs (SXM2 or double-width FH¾L PCIe Form Factors)
- Node Management (AST2500/2400 BMC family, 1GbE Link to Rack Manager)
- Rack Management Sideband : 2x RJ45 Ports for out-ot-band Power Control
- PCIe Fabric Management for multi-Chassis Configurations, multi-Hosting, and IO-Sharing
- Four general-purpose PCIe Cards in addition to configurable and flexible Accelerators
  - Eight nVidia Pascal P100 and V100 SXM2_NVLink
- Flexible choice of GPGPUs in PCIe Card Form Factor
  - Various GPGPUs in double-width, 300W PCIe Card form factor
  - nVidia GPGPUs Such as V100, P100, P40, P4, M40, K80, M60 etc.
  - AMD Radeon
  - Intel Xeon-Phi

  - Other FPGA or ASIC Accelerators
- High PCIe Bandwidth to Host Memory and for peer-to-peer
- Up to 4 PCIe-interconnected Chassis (with a dedicated PCIe Fabric Management Network)
- Expandable to Scale UP
  - From one to four Chassis
  - Internal PCIe Fabric Interconnect
- Scale Out via InfiniBand or Ethernet RoCE-v2 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)

3.3 HGX-1 Implementation Notes

To integrate high-performing compute engines such as FPGAs, and emerging ASIC-based accelerators, we architected HGX-1 as a PCIe-connected Expansion Box (EB). As Figure 2 shows, Project Olympus Universal Server cables to HGX-1 Chassis as a Head Node. To interconnect HGX-1 Chassis to the Server, various Riser Boards plug into the Server (x16, x8), and a set of Cables and Connectors provide the Chassis-to-Chassis Interconnect. This disaggregated approach provides the flexibility to independently choose different CPUs in the Head Node or Accelerators in the Expansion Box.

Using an Upper Baseboard as a Universal Building Block (UBB), HGX-1 supports 8 nVidia Tesla P100 or V100 SXM2 GPGPUs.



GPGPUs,

Figure 2 HGX-1 Chassis cabled to Head Node

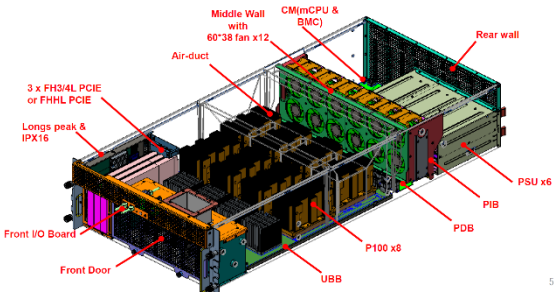


Figure 4 HGX-1 Expansion Chassis with 8 SXM2 Modules

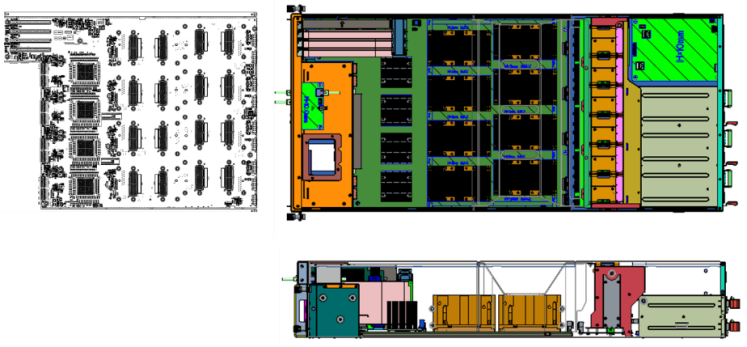


Figure 3 HGX-1 Expansion Chassis Top-view, Side-view, and Upper Baseboard

3.4 Flexible Platform

HGX-1 provides a flexible platform for intercepting future solutions. It offers eight x16 PCI Links (up to sixteen x8) to interconnect to Host Node(s) and other HGX-1 Chassis. These Links provide for a flexible

inter-Switch Link (ISL) topology for optimized Peer-to-Peer traffic. Different cabling topologies provide higher IO bandwidth to Host Memory or better peer-to-peer interconnect of up to 4 Expansion Chassis.

Figure 5 shows how eight nVidia's Pascal P100 SXM2 modules use a private NVLink for efficient peer-to-peer traffic to accelerate HPC applications and frameworks such as Artificial Intelligence (AI).

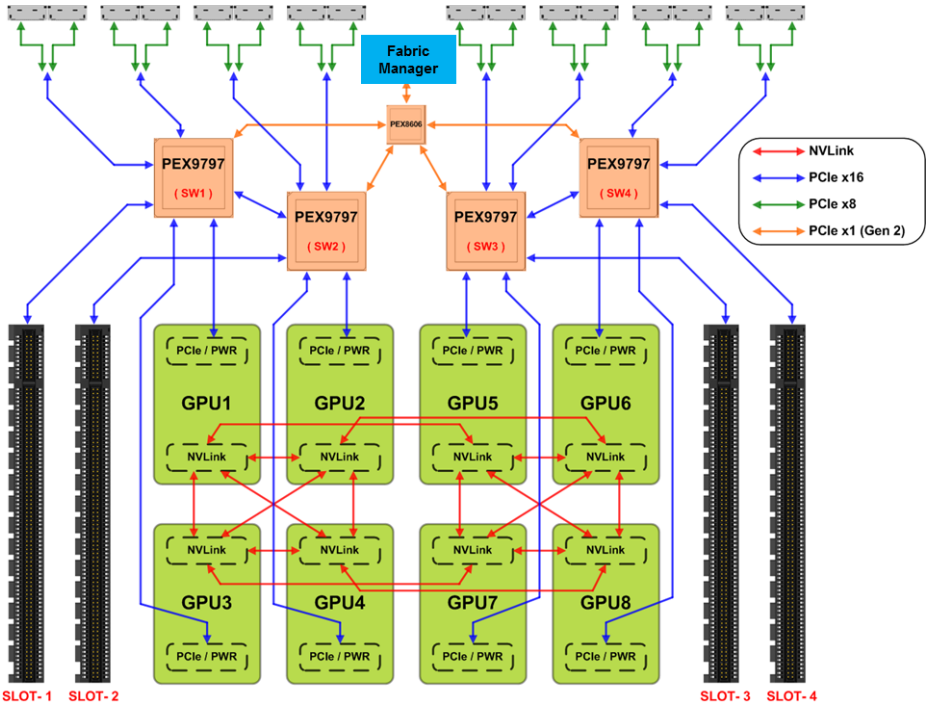


Figure 5 NVLink in Hyper-cube and cable-connected, flexible PCIe Topology

Along with the private NVLink fabric, Figure 5 illustrates a flexible PCIe Interconnect Topology for GPGPU-to-Host via high-bandwidth PCIe Links. Figure 6 depicts one HGX-1 expansion box as a Universal Building Block (UBB) with an extensible Chassis-to-Chassis Interconnect.

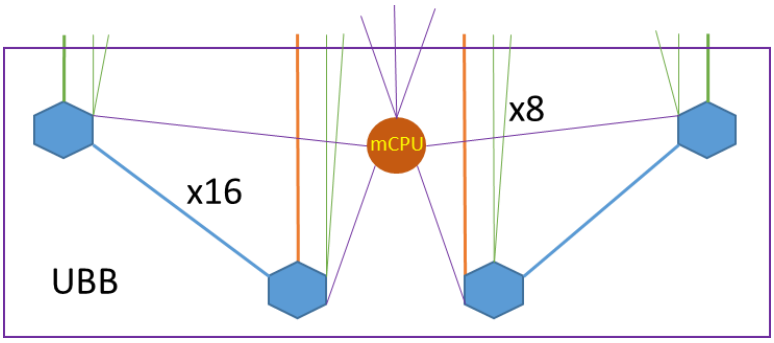


Figure 6 PCIe Topology of HGX-1 Universal Building Block (UBB)

3.5 Scale Up

There are several methods to scale up HGX-1 such as increasing the CPU Host capability, increasing the number of interconnected Chassis, or increasing the Accelerator compute capability. Figure 7 shows a typical 8-GPU configuration where NVLINK carries all inter-GPU communications while providing four x16

PCIe Links to Host CPUs. In this configuration, HGX-1 is a high-performing supercomputer in a box for DL Training jobs.

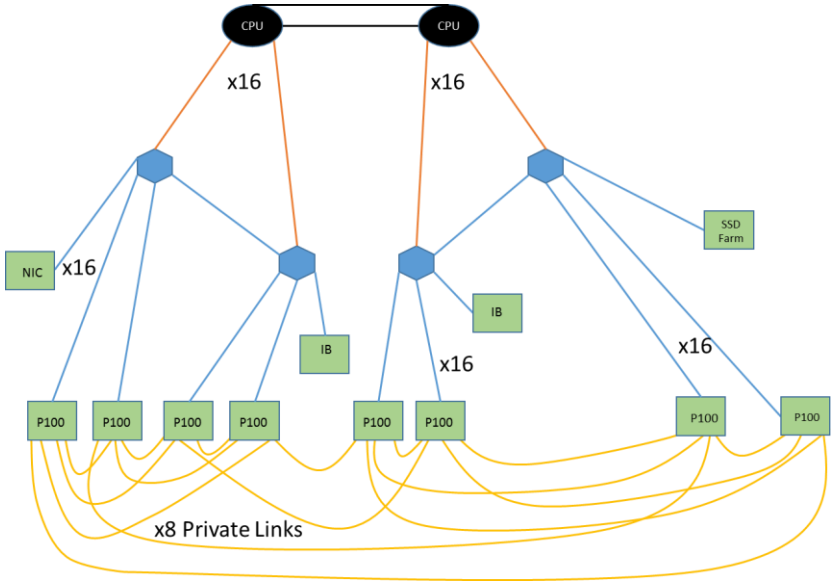


Figure 7 Typical 8-GPU Tesla P100 SXM2 Configuration

Figure 8 depicts a re-arranged cabling for two HGX-1 Chassis to provide a larger pool of GPUs dedicated to a single Training session using NVLink-connected Tesla P100 or V100 SXM2 GPGPUs.

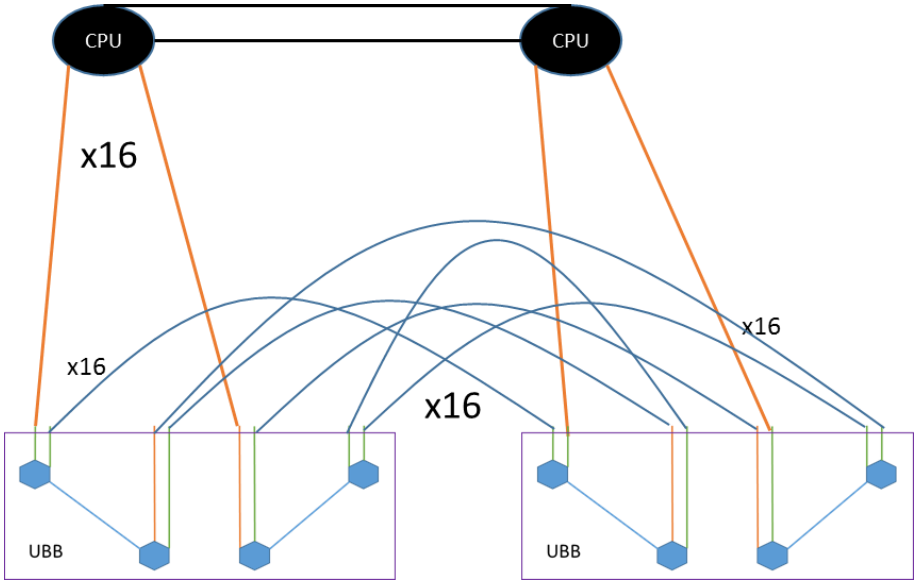


Figure 8 A tightly-coupled, two-Chassis PCIe Interconnect (NVLinks interconnect GPUs within each UBB)

Using Mezzanines such as MEZZ1x16 shown in Figure 9, HGX-1 provides additional flexibility for various PCIe Slot Configurations. To allow the flexibility of accelerator choice, HGX-1 supports eight double-width, 300W PCIe add-in Cards. Without involving inter-CPU Links, HGX-1 provides efficient GPGPU peer-to-peer via NVLink and GPGPU peer-to-peer to InfiniBand NICs via x16 PCIe.

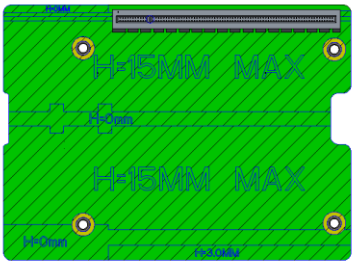


Figure 9 MEZZ1x16

To scale up the computation power and Host Memory bandwidth, HGX-1 of Figure 10 uses eight x16 PCIe Links to allows a four-CPU (4S) Head Node to interconnect eight P100/V100s so that each GPGPU has a full x16 PCIe Link to CPU and Host Memory without sharing the bandwidth with any other GPGPU. This provides for much bisection bandwidth.

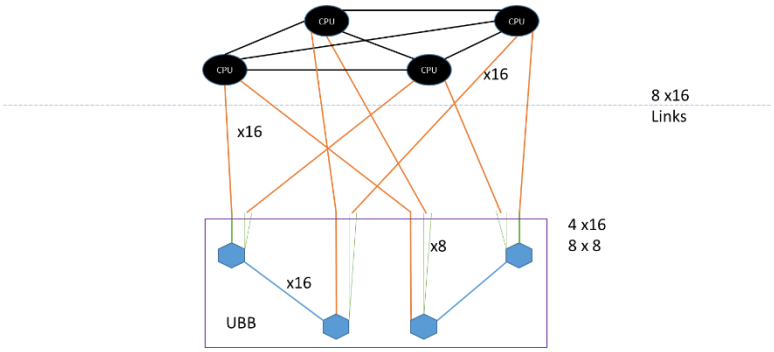


Figure 10 PCIe Interconnect between a 4S Head Node and a HGX-1 PCIe Expansion Box (UBB)

3.6 Extensible Platform

With six 1600W Power Supplies and twelve 60mm Fans, each HGX-1 Expansion Box supports eight 300W double-width PCIe Cards. Up to four HGX-1 Chassis may interconnect to provide for a large PCIe Device Cluster. In this Fabric Mode, the PCIe Fabric may offer multi-Host, Disaggregated/Remote IO, and Dynamic Device Allocation.

4 Industry-wide Partnership to Meet the Requirements

HGX-1 demonstrates industry-wide partnership and fosters interoperability within one hardware platform for multiple configurations to serve various Cloud-scale customer requirements. Based on the high-volume Project Olympus Server, HGX-1 assures seamless integration into existing datacenters. For workload-tuned implementations, HGX-1 allows flexible choice of Compute, Accelerator, Networking, and Storage components optimized for different workloads.

While, HGX-1 scales up to four Chassis using PCIe cables internal to the Chassis to avoid cable management challenges, Microsoft Cognitive Toolkit—previously known as CNTK (http://www.cntk.ai/) increases scale-out performance of HGX-1 using InfiniBand NICs.

Using a common set of building blocks which meet Cloud-scale datacenter requirements, HGX-1 provides a flexible, reconfigurable, adaptable hardware SKU to provide different accelerator configurations.

Project Olympus Hyperscale GPU Accelerator (HGX-1) improves performance of frameworks running DNN, HPC, and Video Transcoding. Refer to OCP Summit Technical Session (http://sched.co/9RU6) and OCP Summit HGX-1 Presentation for performance results as summarized below.

- Optimize HGX-1 Expansion Box configuration to further increase capability of a single Chassis
- Enable more computation per Head Node for HPC-optimized CPU-to-GPU ratio per Job Instance
- Combine multiple HGX-1 Chassis to increase Deep Learning Training throughput
- Use Mezzanine adapters to provide different PCIe add-in Card Accelerators

5 Adjacent Opportunities

Project Olympus base specification defines a modular architecture with clear internal and external interfaces. Hardware modules include Rack, Universal PDU, Rack Manager, 1U&2U Server and mechanical Enclosure, Power Supply (PSU), Universal Motherboard, and Expansion Modules for storage and compute acceleration; while, Software and Firmware modules include, RESTful API, Rack Manager, Software/Firmware/Interface, BMC Firmware, and System Firmware (BIOS/UEFI code). Visit Project Olympus OCP Contributions on GitHub and at Project Olympus OCP wiki pages to receive periodic updates.

The extensible architecture of Microsoft Project Olympus HGX-1 fosters an open-hardware and open-software environment for open contributions. Opportunities to contribute include:

- Fabric Manager, Mezzanine, Cable/Connector, Host, Storage (NVMe, RDMA)
- *Elasticale©* Computing (Elastic-scale: Elastic configuration at Scale)
    - Software-controlled late-binding of CPU, GPU, Storage, and NIC (small and large VMs) as composable compute partitions: drawing resources from different Chassis to form a union of CPU/Memory, GPU, storage, and NIC
    - Hybrid computing: GPGPU, FPGA, CPU, ASIC Silicon: Late-binding choice of pooled or distributed, heterogenous resources at scale
    - Peer-to-peer computing: bring GPU closer to Data (on GPU, in Host Memory, peer-to-peer RDMA in SSD NAND Flash: one hop away)
    - Multi-host access to a pool of PCIe devices
    - Shared IO (PCIe Resource-sharing such as NIC, Storage, or GPUs)
    - Storage resource pooling (within a Chassis or across Chassis)
- Socket and Interconnect standardization of GPGPU, ASIC, etc.