HGX-2
Fusing HPC and AI Computing into One Unified Architecture

William Tsu
wtsu@nvidia.com
5/30/2018
WHAT IS HGX-2?

Highest Performance Cloud Server Platform for AI and HPC Workloads

In this talk:

- More details on the system architecture
- How we work with partners to deliver HGX-2 to cloud data center
- Improvements to current best platform
INTRODUCING NVSWITCH
NVLink Switch

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td># of NVLINK Ports</td>
<td>18</td>
</tr>
<tr>
<td>Per Port Bandwidth</td>
<td>50 GByte/s (both directions)</td>
</tr>
<tr>
<td>Architecture</td>
<td>18 x 18 Full Crossbar</td>
</tr>
<tr>
<td>Process</td>
<td>TSMC 12FFN</td>
</tr>
<tr>
<td>Transistor</td>
<td>2 Billion</td>
</tr>
</tbody>
</table>
HGX-2 GPU BASEBOARD

- Fundamental server building block
- 8 V100 32GB GPU + 6 NVSwitch
- For each GPU
  - PCIe x16 @ 32GB/s
  - 6 NVLINK @ 300GB/s
- 8 GPU fully connected by NVSwitch
16 GPU FULLY CONNECTED

➢ Every GPU can communicate with another GPU simultaneously at full 300GB/s
➢ Every GPU has direct read/write access to full 512GB GPU memory
PHYSICAL IMPLEMENTATION

- 16 NVIDIA Tesla V100 GPUs
- 0.5TB Memory

- 12 NVIDIA NVSwitches
  - Direct GPU-to-GPU Connection Between All 16 GPUs

- PCIe connectors
- NVLink connectors

- PCB based NVLink bridge
WORKING WITH ECOSYSTEM

- NVIDIA focus on delivering the most performance optimized GPU sub-system
- Server partners focus on rest of system design (power, thermal, chassis) & customize to cloud data center needs
- Save partners’ resource and together we bring the latest technology to market faster
ARCHITECTURE RECOMMENDATION

➢ Separate CPU and GPU enclosures
➢ Four PCIe x16 cables for sufficient bandwidth
➢ Shallow balance PCIe tree
➢ GPU RDMA capable NIC (network interface card)
➢ NIC & NVMe close to GPU
HGX-2 SETTING AI SPEED RECORD

- Fastest Single Node to date - 15,500 images/sec on Resnet-50 Training
- Replaces 300 CPU server node (Xeon Gold)
- 1/8 the Cost, 1/60 the Space, 1/18 the Power

Workload: ResNet50, 90 epochs to solution | CPU server: dual-socket Intel Xeon Gold 6140
COMPARING WITH HGX-1

➢ HGX-1 - Fastest architecture shipping today
➢ Hybrid cube mesh
➢ Each GPU has:
  ➢ 6 NVLink at 300GB/s for GPU peer to peer traffic
  ➢ PCIe x16 at 32GB/s for CPU<->GPU communication
### HGX-2 ADVANCES OVER HGX-1

<table>
<thead>
<tr>
<th></th>
<th>HGX-1</th>
<th>HGX-2</th>
</tr>
</thead>
<tbody>
<tr>
<td># of GPU per node</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Node spec</td>
<td>1 Petaflop, 256GB GPU Memory</td>
<td>2 Petaflops, 512GB GPU Memory</td>
</tr>
<tr>
<td>1GPU to 1GPU</td>
<td>1 or 2 NVLINK or PCIe</td>
<td>Always 6 NVLINK</td>
</tr>
<tr>
<td>Fully Connected by NVLink</td>
<td>4 GPU</td>
<td>16 GPU</td>
</tr>
<tr>
<td>Bisection Bandwidth</td>
<td>300GB/s</td>
<td>2,400GB/s (48 NVLink)</td>
</tr>
<tr>
<td>Multi-GPU Deep Learning</td>
<td>Data Parallel All-Reduce</td>
<td>Faster Data Parallel All-Reduce, Enable Model Parallel</td>
</tr>
</tbody>
</table>
2X FASTER THAN TWO HGX-1

<table>
<thead>
<tr>
<th>HPC</th>
<th>AI Training</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2X FASTER</strong></td>
<td><strong>2.7X FASTER</strong></td>
</tr>
<tr>
<td>Physics (MILC benchmark) 4D Grid</td>
<td>Language Model (Transformer with MoE) All-to-all</td>
</tr>
<tr>
<td><strong>2.4X FASTER</strong></td>
<td><strong>2X FASTER</strong></td>
</tr>
<tr>
<td>Weather (ECMWF benchmark) All-to-all</td>
<td>Recommender (Sparse Embedding) Reduce &amp; Broadcast</td>
</tr>
<tr>
<td><strong>2X FASTER</strong></td>
<td><strong>2X FASTER</strong></td>
</tr>
<tr>
<td>Reduce &amp; Broadcast</td>
<td>Recommendation (Sparse Embedding) Reduce &amp; Broadcast</td>
</tr>
</tbody>
</table>

2 HGX-1V servers have dual socket Xeon E5 2698v4 Processor. 8 x V100 GPUs. Servers connected via 4X 100Gb IB ports (run on DGX-1) | HGX-2 server has dual-socket Xeon Platinum 8168 Processor. 16 V100 GPUs (run on DGX-2)
HGX-2 PARTNERS
TAKE AWAY

➢ HGX-2: The Most Powerful Cloud Server Platform for AI and HPC workloads
   ➢ 16 GPU per node, new NVSwitch technology
➢ Enable faster time to solution, and open up new use cases
   ➢ 16 GPU fully connected, easier to program, new model parallel
   ➢ 512GB GPU memory for larger problems, e.g. deeper network, bigger recommender
➢ Work with server ecosystem partners to deliver the fastest platform to cloud data center
MORE RESOURCES

➢ HGX-2 product page and general info:

➢ NVSwitch technical overview:

➢ HGX-2 developer blog: