Alex Ishii and Denis Foley with
Eric Anderson, Bill Dally, Glenn Dearth
Larry Dennison, Mark Hummel and John Schafer
NVIDIA® DGX-2™ SERVER AND NVSWITCH™

16 Tesla™ V100 32 GB GPUs
FP64: 125 TFLOPS
FP32: 250 TFLOPS
Tensor: 2000 TFLOPS
512 GB of GPU HBM2

Single-Server Chassis
10U/19-Inch Rack Mount
10 kW Peak TDP
Dual 24-core Xeon CPUs
1.5 TB DDR4 DRAM
30 TB NVMe Storage

12 NVSwitch Network
Full-Bandwidth Fat-Tree Topology
2.4 TBps Bisection Bandwidth
Global Shared Memory
Repeater-less

New NVSwitch Chip
18 2nd Generation NVLink™ Ports
25 GBps per Port
900 GBps Total Bidirectional Bandwidth
450 GBps Total Throughput

©2018 NVIDIA CORPORATION
OUTLINE

The Case for “One Gigantic GPU” and NVLink Review

NVSwitch — Speeds and Feeds, Architecture and System Applications

DGX-2 Server — Speeds and Feeds, Architecture, and Packaging

Signal-Integrity Design Highlights

Achieved Performance
Users explicitly express parallel work in NVIDIA CUDA®

GPU Driver distributes work to available Graphics Processing Clusters (GPC)/Streaming Multiprocessor (SM) cores

GPC/SM cores can compute on data in any of the second-generation High Bandwidth Memories (HBM2s)

GPC/SM cores use shared HBM2s to exchange data
TWO GPUS WITH PCIE

Access to HBM2 of other GPU is at PCIe BW (32 GBps (bidirectional))

Interactions with CPU compete with GPU-to-GPU

PCIe is the “Wild West” (lots of performance bandits)
TWO GPUS WITH NVLINK

All GPCs can access all HBM2 memories

Access to HBM2 of other GPU is at multi-NVLink bandwidth (300 GBps bidirectional in V100 GPUs)

NVLinks are effectively a “bridge” between XBARs

No collisions with PCIe traffic
THE “ONE GIGANTIC GPU” IDEAL

Highest number of GPUs possible

Single GPU Driver process controls all work across all GPUs

From perspective of GPCs, all HBM2s can be accessed without intervention by other processes (LD/ST instructions, Copy Engine RDMA, everything “just works”)

Access to all HBM2s is independent of PCIe

Bandwidth across bridged XBARs is as high as possible (some NUMA is unavoidable)
Problem Size Capacity
Problem size is limited by aggregate HBM2 capacity of entire set of GPUs, rather than capacity of single GPU

Strong Scaling
NUMA-effects greatly reduced compared to existing solutions
Aggregate bandwidth to HBM2 grows with number of GPUs

Ease of Use
Apps written for small number of GPUs port more easily
Abundant resources enable rapid experimentation
INTRODUCING NVSWITCH

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bidirectional Bandwidth per NVLink</td>
<td>51.5 GBps</td>
</tr>
<tr>
<td>NRZ Lane Rate (x8 per NVLink)</td>
<td>25.78125 Gbps</td>
</tr>
<tr>
<td>Transistors</td>
<td>2 Billion</td>
</tr>
<tr>
<td>Process</td>
<td>TSMC 12FFN</td>
</tr>
<tr>
<td>Die Size</td>
<td>106 mm^2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bidirectional Aggregate Bandwidth</td>
<td>928 GBps</td>
</tr>
<tr>
<td>NVLink Ports</td>
<td>18</td>
</tr>
<tr>
<td>Mgmt Port (config, maintenance, errors)</td>
<td>PCIe</td>
</tr>
<tr>
<td>LD/ST BW Efficiency (128 B pkts)</td>
<td>80.0%</td>
</tr>
<tr>
<td>Copy Engine BW Efficiency (256 B pkts)</td>
<td>88.9%</td>
</tr>
</tbody>
</table>
NVSWITCH BLOCK DIAGRAM

GPU-XBAR-bridging device; not a general networking device

Packet-Transforms make traffic to/from multiple GPUs look like they are to/from single GPU

XBAR is non-blocking

SRAM-based buffering

NVLink IP blocks and XBAR design/verification infrastructure reused from V100
NVLINK: PHYSICAL SHARED MEMORY

Virtual-to physical address translation is done in GPCs

NVLink packets carry physical addresses

NVSwitch and DGX-2 follow same model

©2018 NVIDIA CORPORATION
Hop-by-hop error checking deemed sufficient in intra-chassis operating environment

Low BER ($10^{-12}$ or better) on intra-chassis transmission media removes need for high-overhead protections like FEC

HW-based consistency checks guard against “escapes” from hop-to-hop

SW responsible for additional consistency checks and clean-up

**NVSwitch: Reliability**

- Hop-by-hop error checking deemed sufficient in intra-chassis operating environment
- Low BER ($10^{-12}$ or better) on intra-chassis transmission media removes need for high-overhead protections like FEC
- HW-based consistency checks guard against “escapes” from hop-to-hop
- SW responsible for additional consistency checks and clean-up
Packet-processing and switching (XBAR) logic is extremely compact.

With more than 50% of area taken-up by I/O blocks, maximizing “perimeter”-to-area ratio was key to an efficient design.

Having all NVLink ports exit die on parallel paths simplified package substrate routing.
No NVSwitch

Connect GPU ↔ directly
Aggregate NVLinks into gangs for higher bandwidth
Interleaved over the links to prevent camping
Max bandwidth between two GPUs limited to the bandwidth of the gang

With NVSwitch

Interleave traffic across all the links and to support full bandwidth between any pair of GPUs
Traffic to a single GPU is non-blocking, so long as aggregate bandwidth of six NVLinks is not exceeded
Add NVSwitches in parallel to support more GPUs

Eight GPU closed system can be built with three NVSwitches with two NVLinks from each GPU to each switch

Gang width is still six — traffic is interleaved across all the GPU links

GPUs can now communicate pairwise using the full 300 GBps bidirectional between any pair

NVSwitch XBAR provides unique paths from and source to any destination — non-blocking, non-interfering
Taking this to the limit —
connect one NVLink from each
GPU to each of six switches

No routing between different
switch planes required

Eight NVLinks of the 18 available
per switch are used to connect
to GPUs

Ten NVLinks available per switch
for communication outside the local
group (only eight are required to
support full bandwidth)

This is the GPU baseboard
configuration for DGX-2
Two of these building blocks together form a fully connected 16 GPU cluster

Non-blocking, non-interfering (unless same destination is involved)

Regular load, store, atomics just work

Left-right symmetry simplifies physical packaging, and manufacturability
# NVIDIA DGX-2: SPEEDS AND FEEDS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DGX-2 Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Tesla V100 GPUs</td>
<td>16</td>
</tr>
<tr>
<td>Aggregate FP64/FP32</td>
<td>125/250 TFLOPS</td>
</tr>
<tr>
<td>Aggregate Tensor (FP16)</td>
<td>2000 TFLOPS</td>
</tr>
<tr>
<td>Aggregate Shared HBM2</td>
<td>512 GB</td>
</tr>
<tr>
<td>Aggregate HBM2 Bandwidth</td>
<td>14.4 TBps</td>
</tr>
<tr>
<td>Per-GPU NVLink Bandwidth</td>
<td>300 GBps bidirectional</td>
</tr>
<tr>
<td>Chassis Bisection Bandwidth</td>
<td>2.4 TBps</td>
</tr>
<tr>
<td>InfiniBand NICs</td>
<td>8 Mellanox EDR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DGX-2 Spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPUs</td>
<td>Dual Xeon Platinum 8168</td>
</tr>
<tr>
<td>CPU Memory</td>
<td>1.5 TB DDR4</td>
</tr>
<tr>
<td>Aggregate Storage</td>
<td>30 TB (8 NVMes)</td>
</tr>
<tr>
<td>Peak Max TDP</td>
<td>10 kW</td>
</tr>
<tr>
<td>Dimensions (H/W/D)</td>
<td>17.3”(10U)/ 19.0”/32.8” (440.0mm/482.3mm/834.0mm)</td>
</tr>
<tr>
<td>Weight</td>
<td>340 lbs (154.2 kgs)</td>
</tr>
<tr>
<td>Cooling (forced air)</td>
<td>1,000 CFM</td>
</tr>
</tbody>
</table>

©2018 NVIDIA CORPORATION
Xeon sockets are QPI connected, but affinity-binding keeps GPU-related traffic off QPI

PCIe tree has NICs connected to pairs of GPUs to facilitate GPUDirect™ RDMAs over IB network

Configuration and control of the NVSwitches is via driver process running on CPUs
Two GPU Baseboards with eight V100 GPUs and six NVSwitches on each

Two Plane Cards carry 24 NVLinks each

No repeaters or redrivers on any of the NVLinks conserves board space and power
GPU Baseboards get power and PCIe from front midplane

Front midplane connects to I/O-Expander PCB (with PCIe switches and NICs) and CPU Motherboard

48V PSUs reduce current load on distribution paths

Internal “bus bars” bring current from each PSU to front-mid plane
Forced-air cooling of Baseboards, I/O Expander, and CPU provided by ten 92 mm fans

Four supplemental 60 mm internal fans to cool NVMe drives and PSUs

Air to NVSwitches is pre-heated by GPUs, requiring “full height” heatsinks
DGX-2: SIGNAL INTEGRITY OPTIMIZATIONS

NVLink repeaterless topology trades-off longer traces to GPUs for shorter traces to Plane Cards

Custom SXM and Plane Card connectors for reduced loss and crosstalk

TX and RX grouped in pin-fields to reduce crosstalk
Test has each GPU reading data from another GPU across bisection (from GPU on different Baseboard)

Raw bisection bandwidth is 2.472 TBps

1.98 TBps achieved Read bisection bandwidth matches theoretical 80% bidirectional NVLink efficiency

“All-to-all” (each GPU reads from eight GPUs on other PCB) results are similar
Results are “iso-problem instance” (more GFLOPS means shorter running time)

As problem is split over more GPUs, it takes longer to transfer data than to calculate locally

Aggregate nature of HBM2 capacity enables larger problem sizes

DGX-2: CUFFT (16K X 16K)

![Graph showing GFLOPS vs. Number of GPUs for DGX-2 and DGX-1™]

- Green line: ½ DGX-2 (All-to-All NVLink Fabric)
- Blue line: DGX-1™ (V100 and Hybrid Cube Mesh)
Important communication primitive in Machine-Learning apps

DGX-2 provides increased bandwidth and lower latency compared to two 8-GPU servers (connected with InfiniBand)

NVLink network has efficiency superior to InfiniBand on smaller message sizes
DGX-2: >2X SPEED-UP ON TARGET APPS
Two DGX-1™ Servers (V100) Compared to DGX-2 – Same Total GPU Count

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

Two DGX-1 servers have dual socket Xeon E5 2698v4 Processor. Eight Tesla V100 32 GB GPUs. Servers connected via four EDR IB/GbE ports | DGX-2 server has dual-socket Xeon Platinum 8168 Processor. 16 Tesla V100 32GB GPUs
Eight GPU baseboard with six NVSwitches

Two HGX-2 boards can be passively connected to realize 16-GPU systems

ODM/OEM partners build servers utilizing NVIDIA HGX-2 GPU baseboards

Design guide provides recommendations
New NVSwitch
- 18-NVLink-Port, Non-Blocking NVLink Switch
- 51.5 GBps-per-Port
- 928 GBps Aggregate Bidirectional Bandwidth

DGX-2
- “One Gigantic GPU” Scale-Up Server with 16 Tesla V100 GPUs
- 125 TF (FP64) to 2000 TF (Tensor)
- 512 GB Aggregate HBM2 Capacity
- NVSwitch Fabric Enables >2X Speed-Up on Target Apps