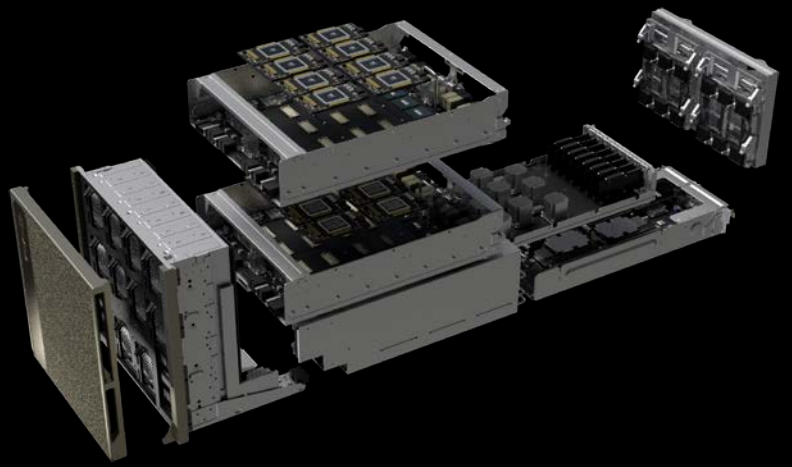
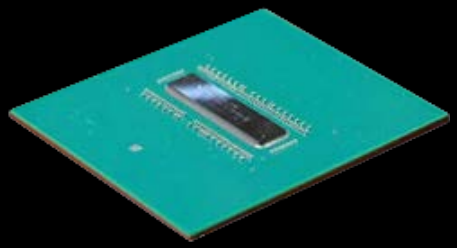
NVSWITCH AND DGX-2

NVLINK-SWITCHING CHIP AND SCALE-UP COMPUTE SERVER

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

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

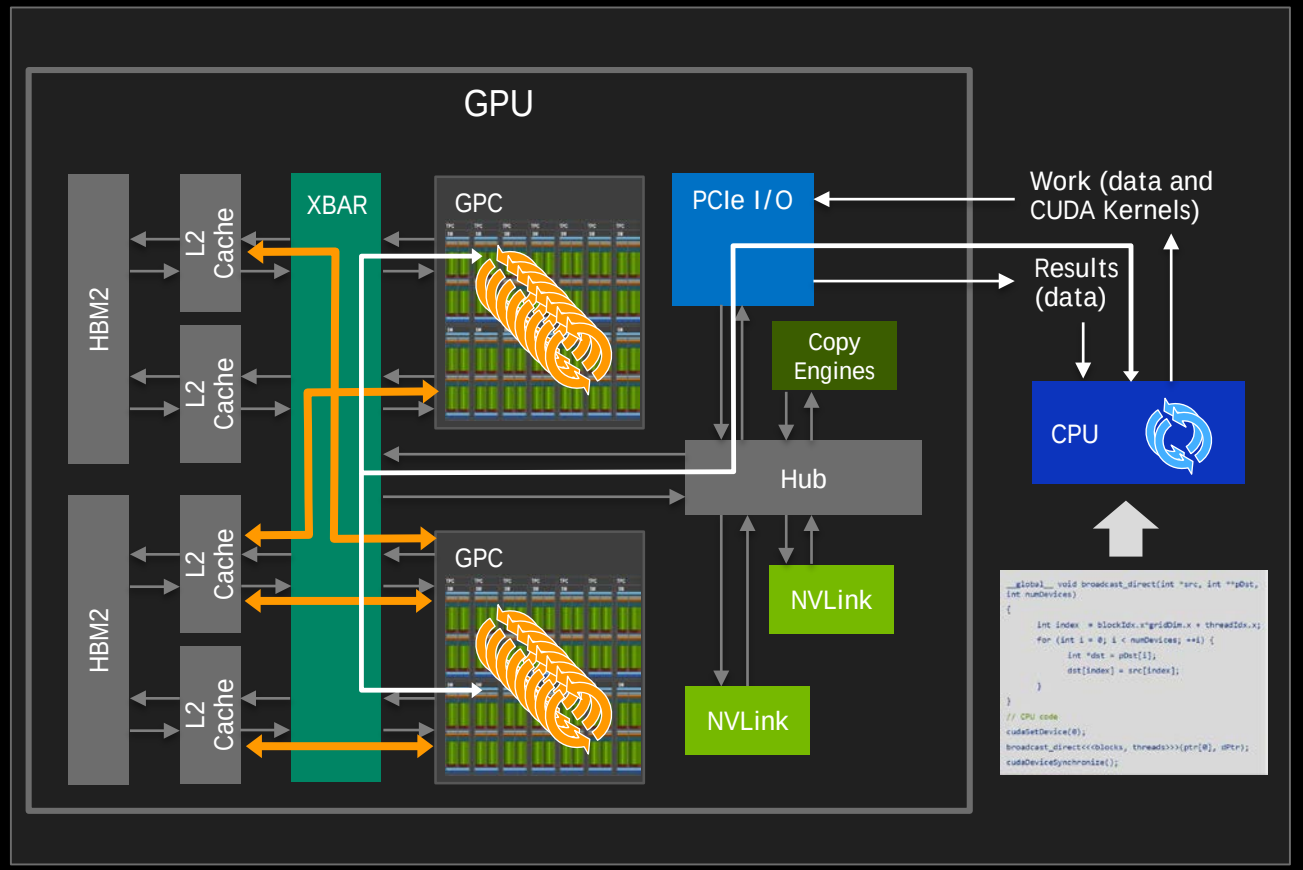
MULTI-CORE AND CUDA WITH ONE GPU

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

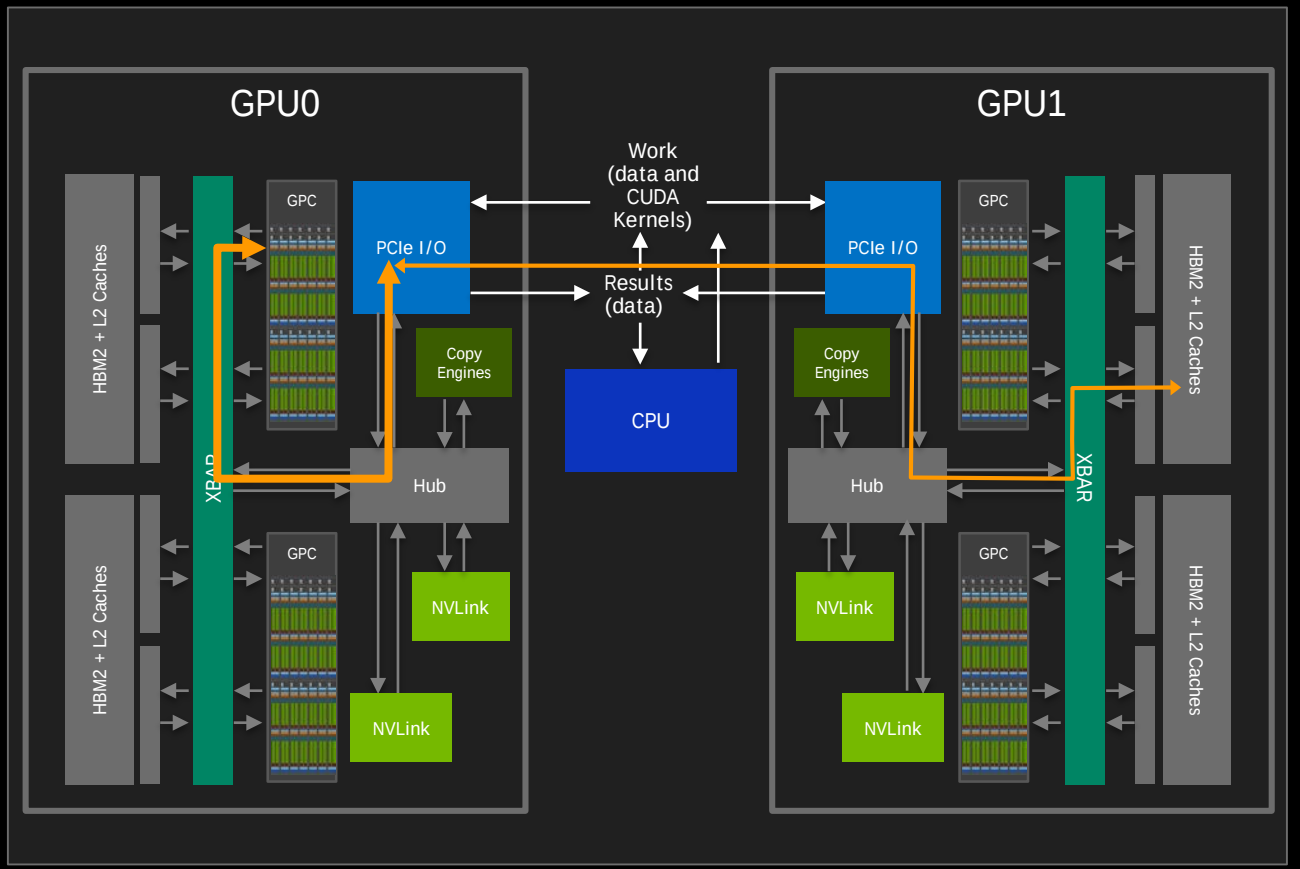


TWO GPUS WITH PCIE

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)



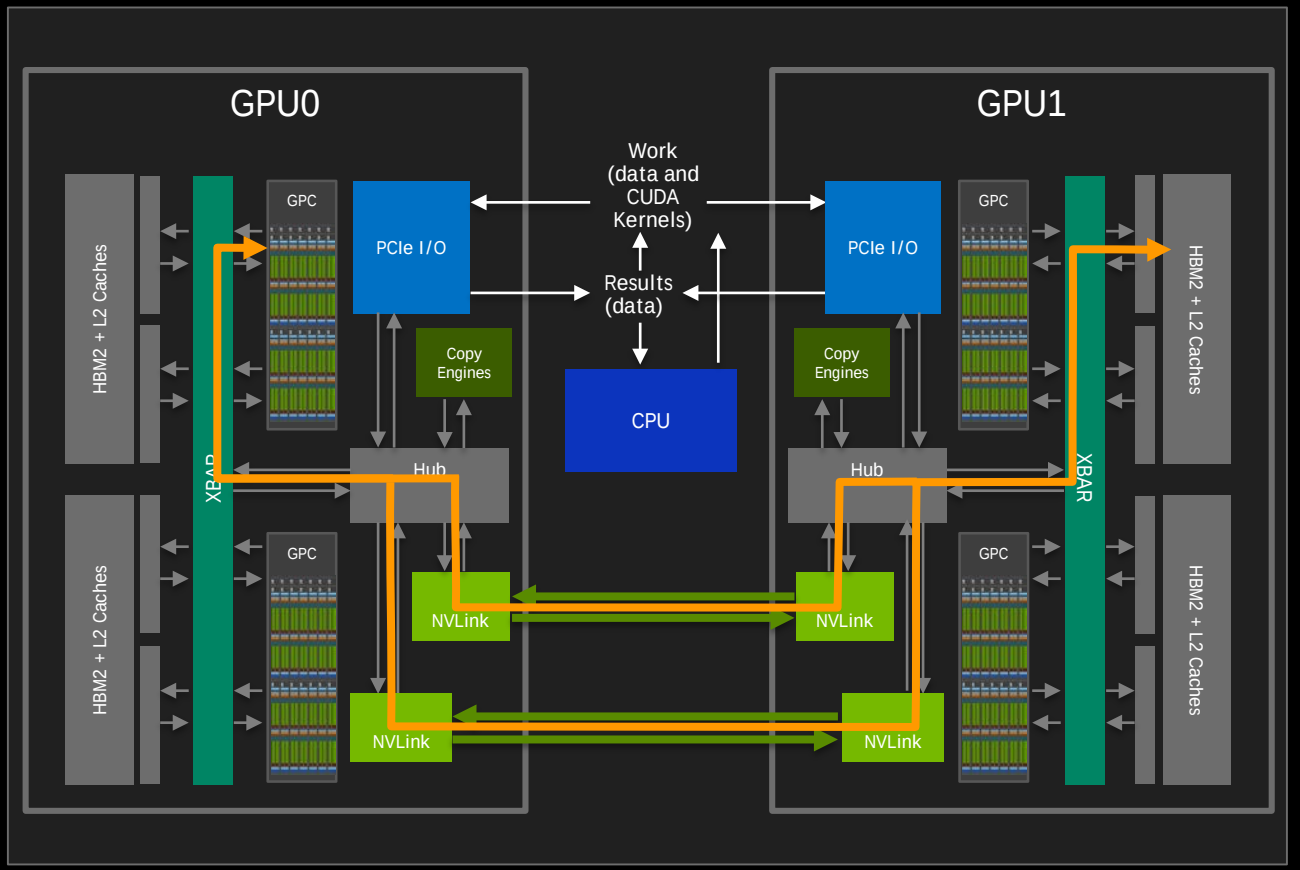
TWO GPUS WITH NVLINK

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

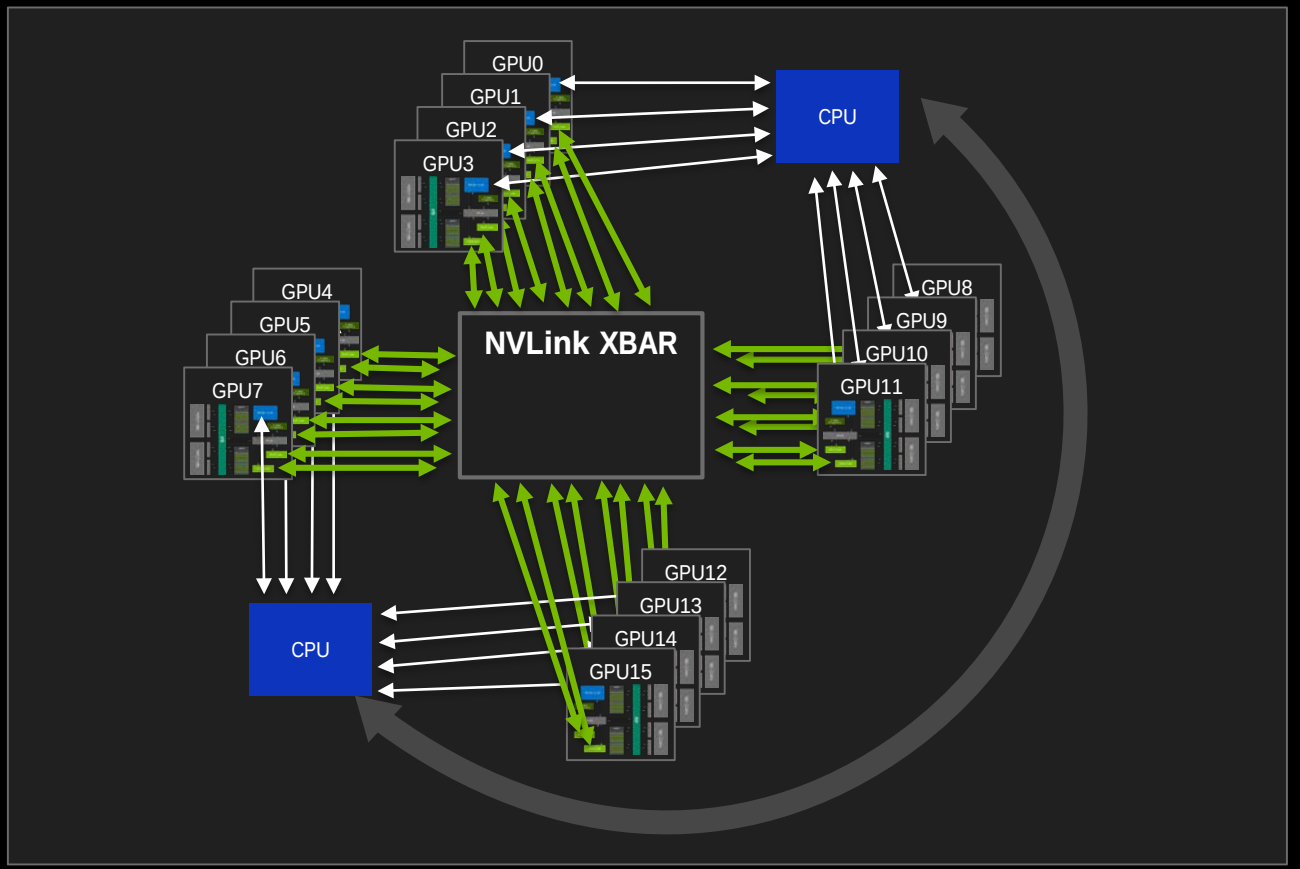
Highest number of GPUs possible

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)



“ONE GIGANTIC GPU” BENEFITS

Problem Size Capacity

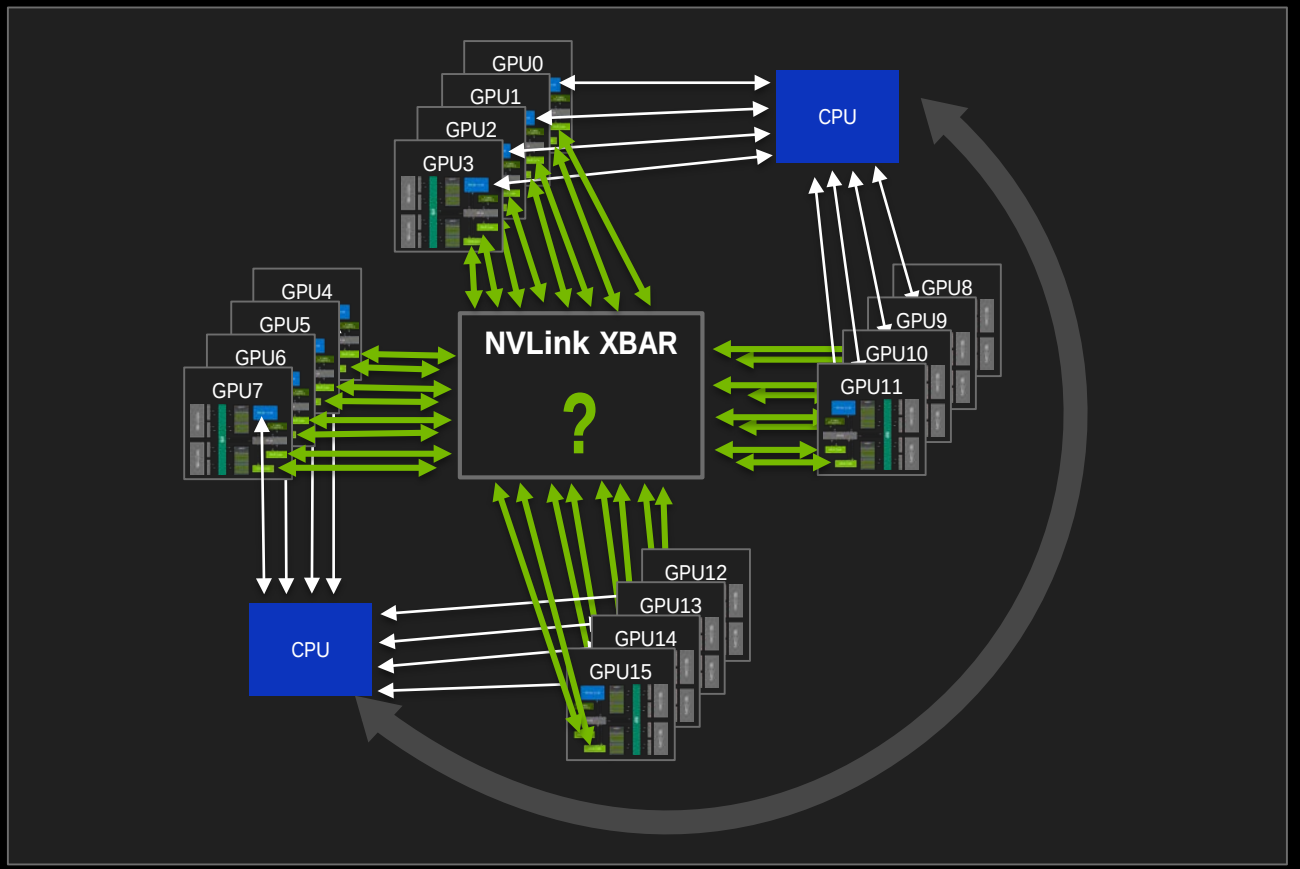
- Problem size is limited by aggregate HBM2 capacity of entire set of GPUs, rather than capacity of single GPU

Strong Scaling

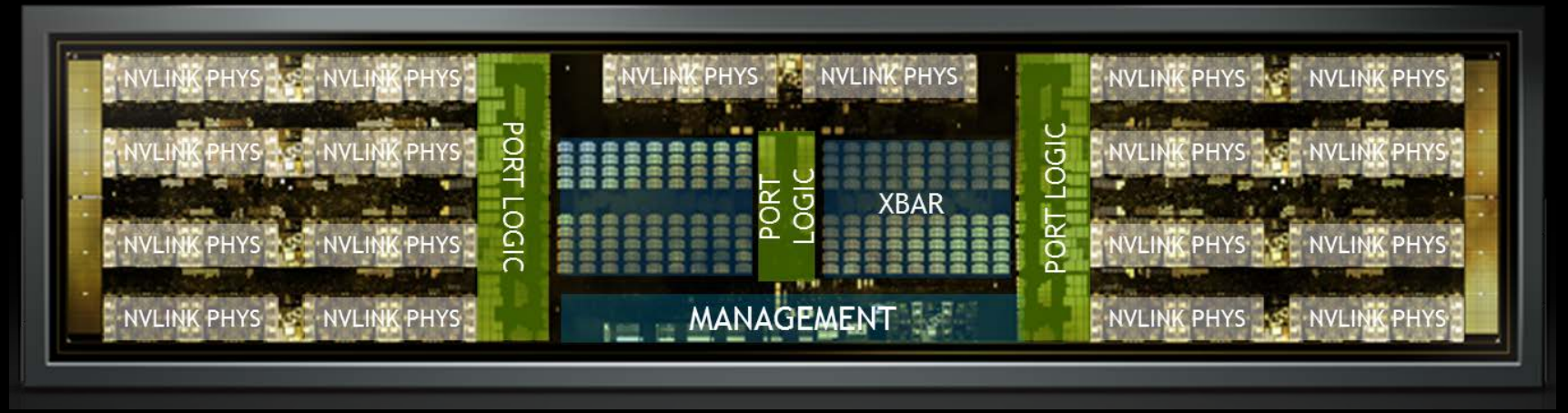
- NUMA-effects greatly reduced compared to existing solutions
- Aggregate bandwidth to HBM2 grows with number of GPUs

Ease of Use

- Apps written for small number of GPUs port more easily
- Abundant resources enable rapid experimentation



INTRODUCING NVSWITCH



Parameter	Spec
Bidirectional Bandwidth per NVLink	51.5 GBps
NRZ Lane Rate (x8 per NVLink)	25.78125 Gbps
Transistors	2 Billion
Process	TSMC 12FFN
Die Size	106 mm^2

Parameter	Spec
Bidirectional Aggregate Bandwidth	928 GBps
NVLink Ports	18
Mgmt Port (config, maintenance, errors)	PCIe
LD/ST BW Efficiency (128 B pkts)	80.0%
Copy Engine BW Efficiency (256 B pkts)	88.9%

NVSWITCH BLOCK DIAGRAM

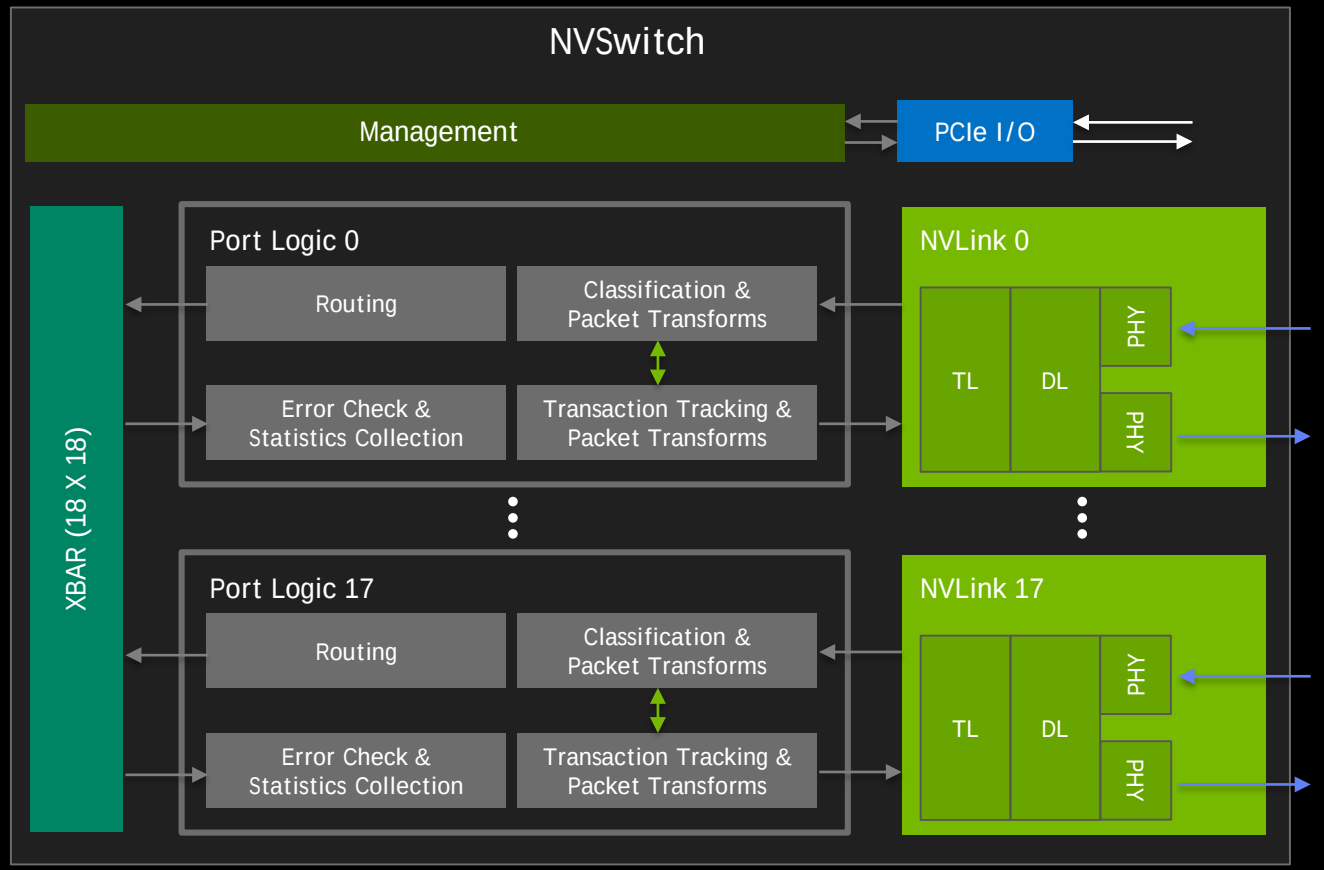
GPU-XBAR-bridging device;
not a general networking device

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

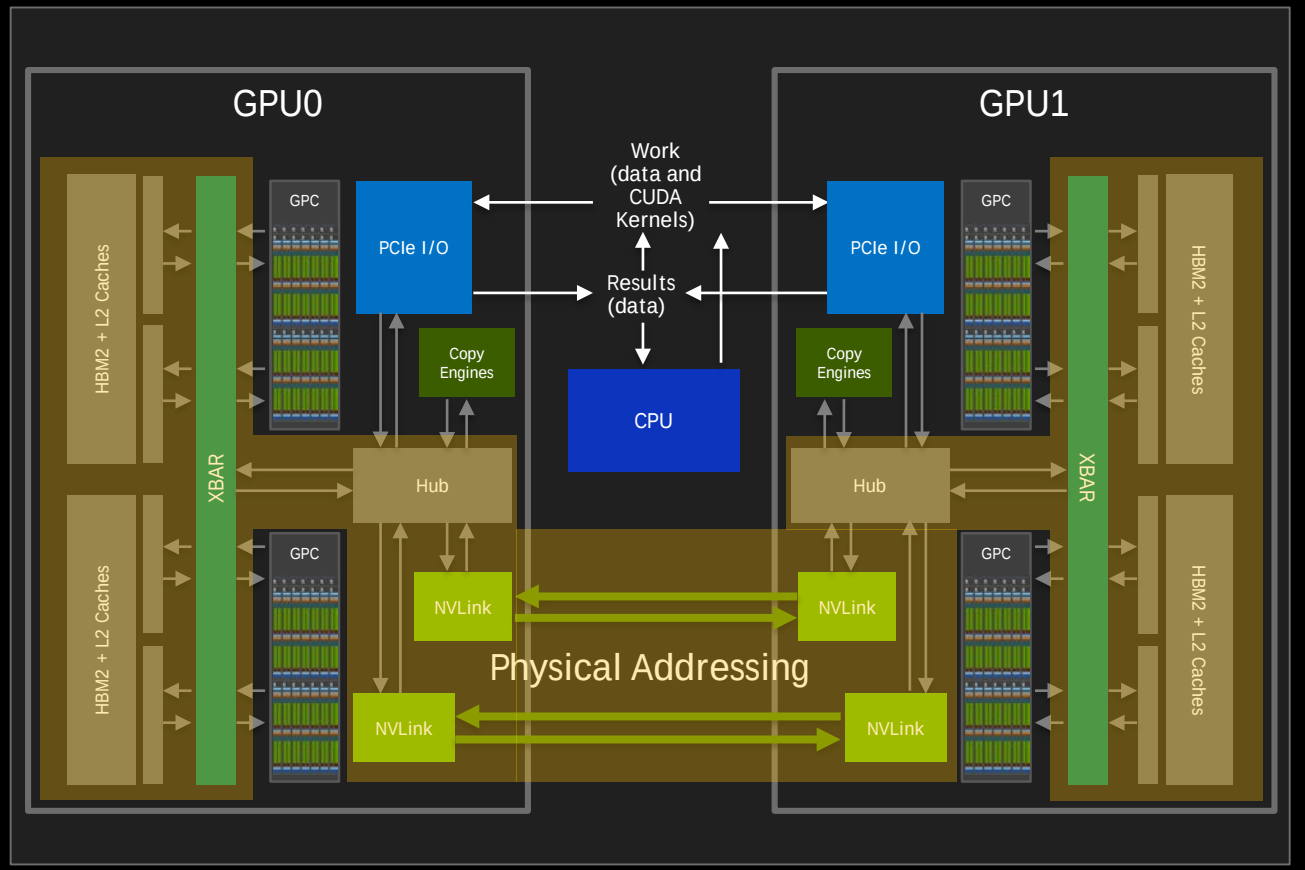


NVLINK: PHYSICAL SHARED MEMORY

Virtual-to physical address translation is done in GPCs

NVLink packets carry physical addresses

NVSwitch and DGX-2 follow same model



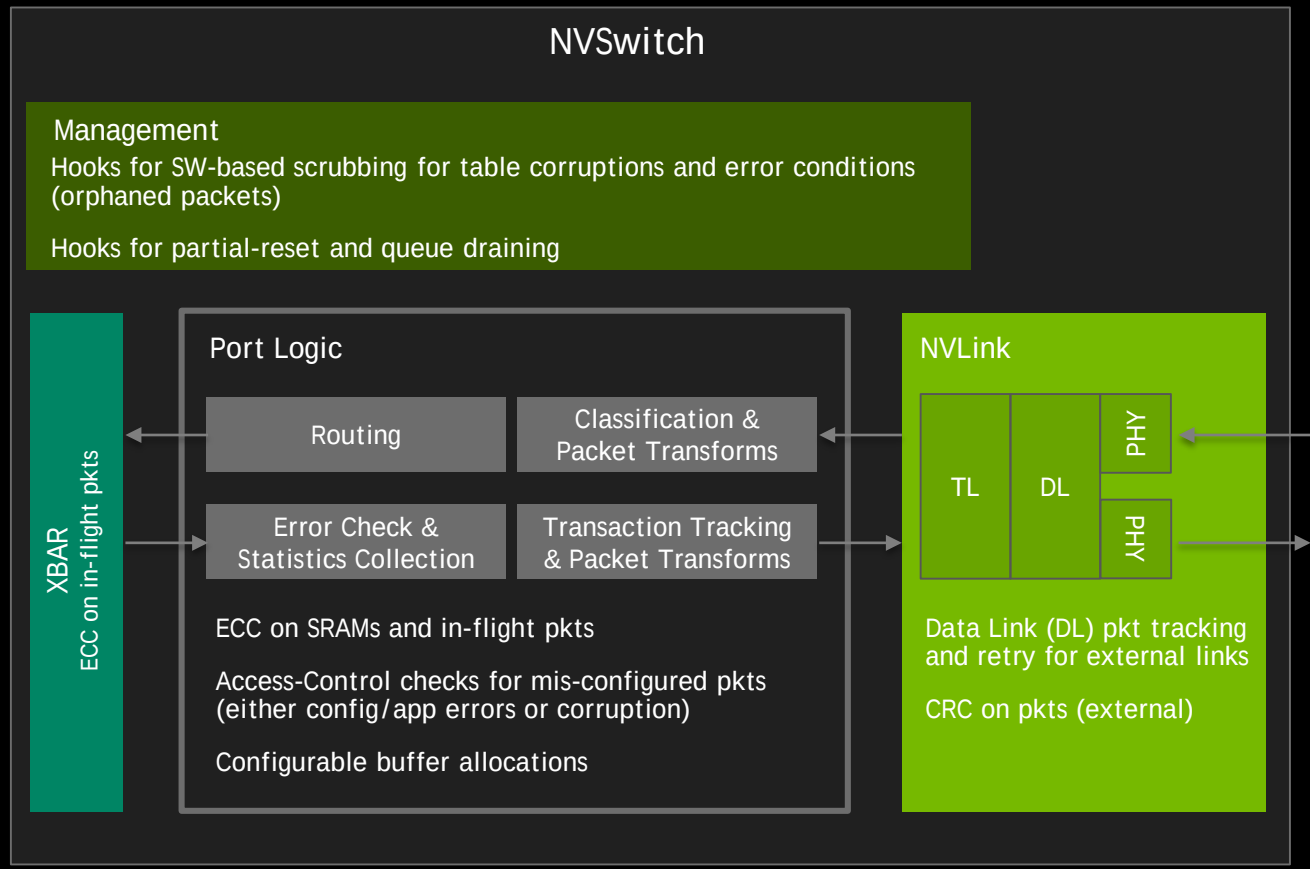
NVSWITCH: RELIABILITY

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



NVSWITCH: DIE ASPECT RATIO

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

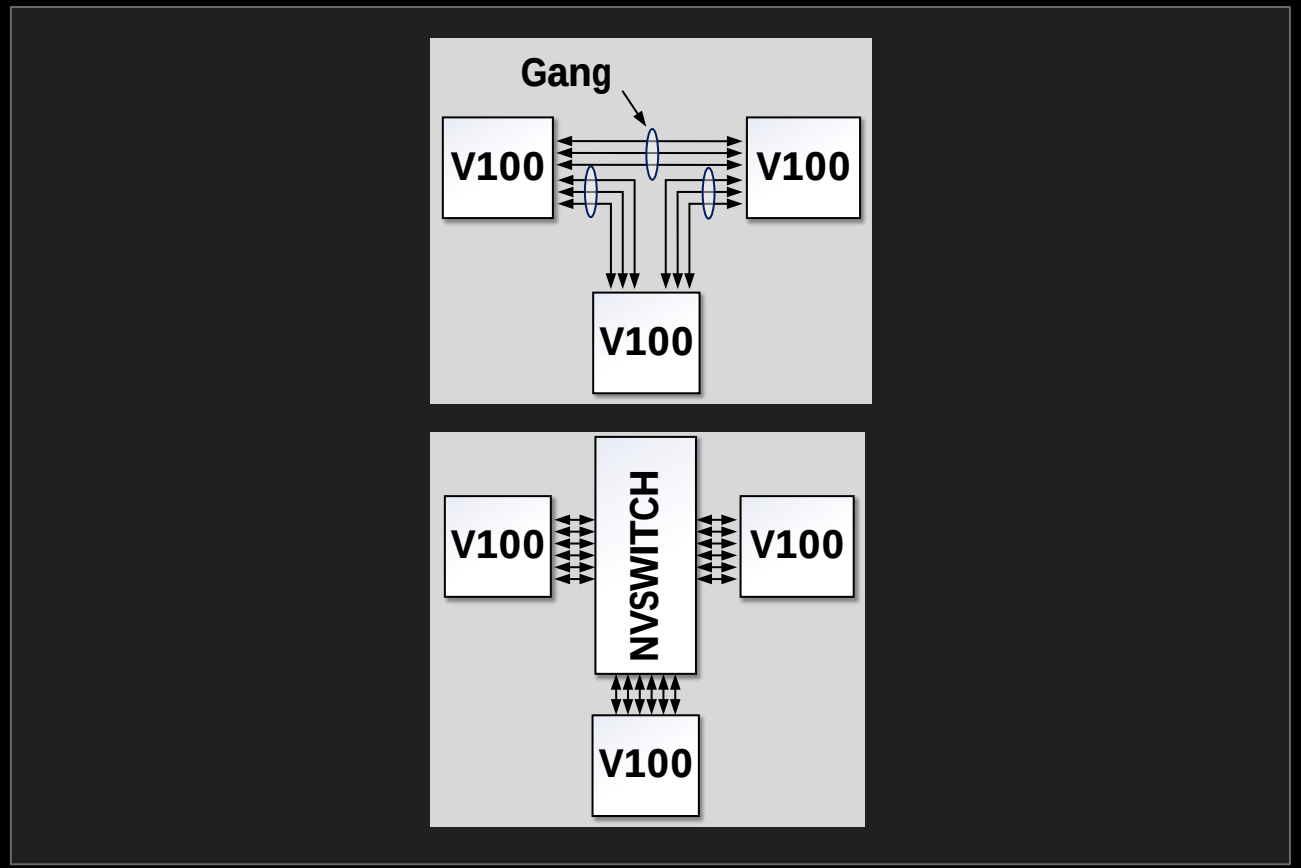
SIMPLE SWITCH SYSTEM

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



SWITCH CONNECTED SYSTEMS

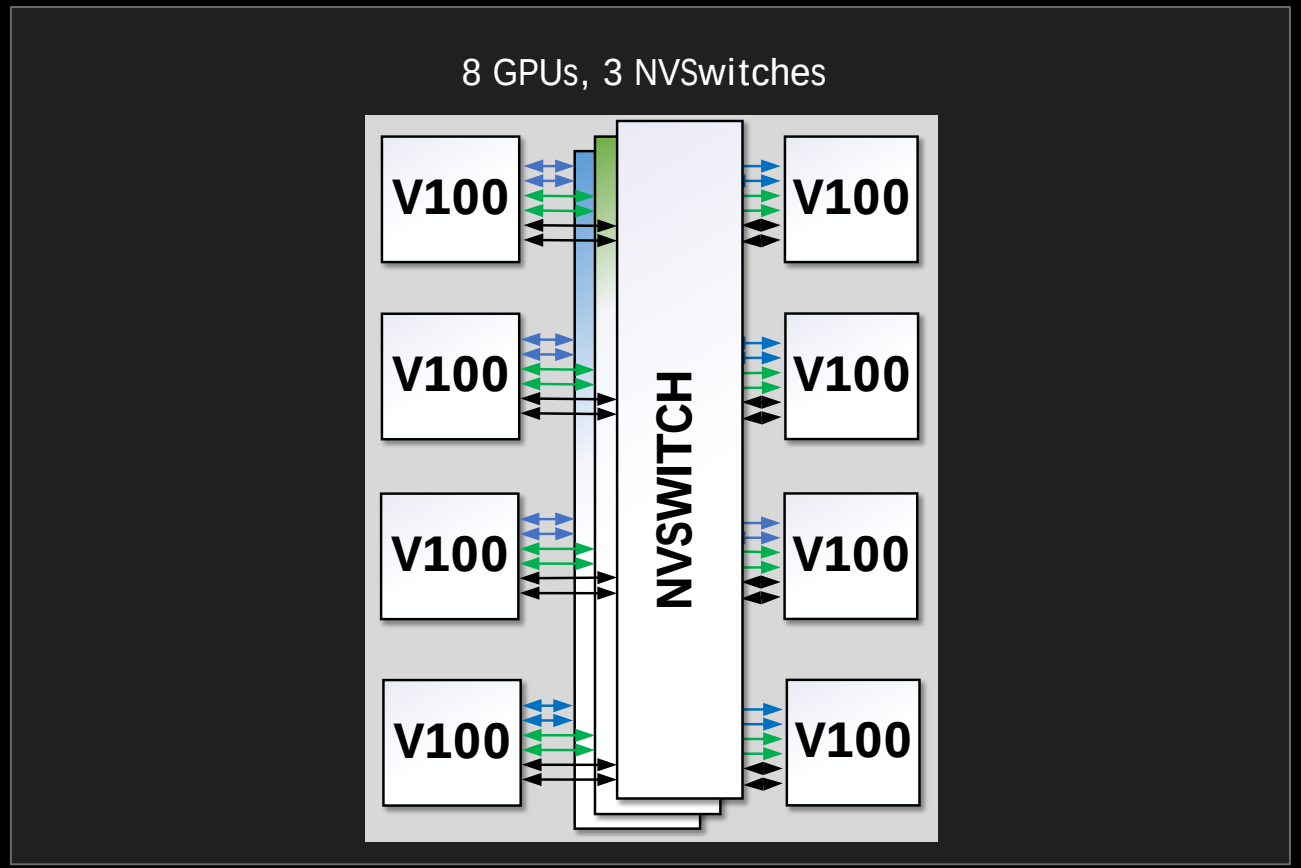
Add NVSwitches in parallel to support more GPUs

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



NON-BLOCKING BUILDING BLOCK

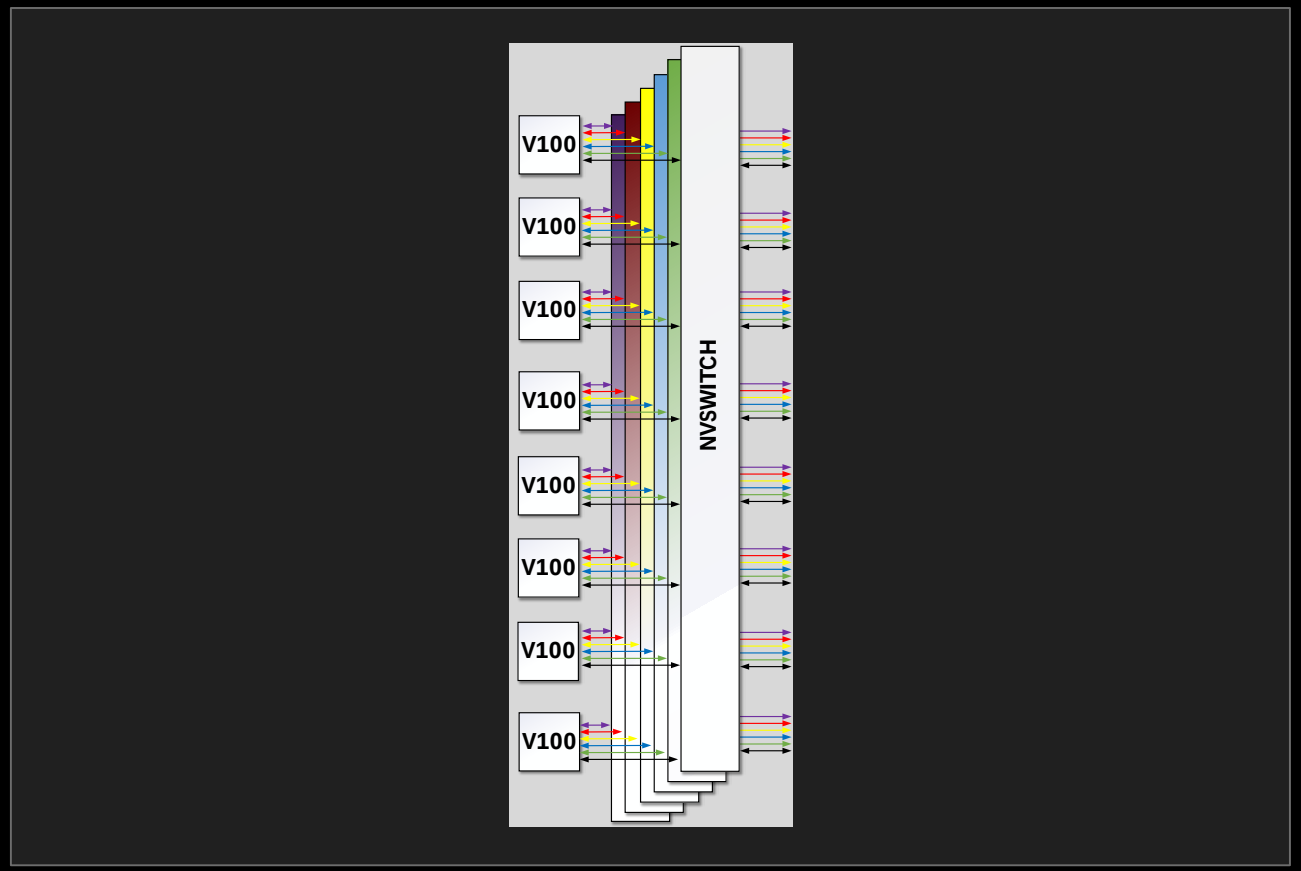
Taking this to the limit — connect one NVLink from each GPU to each of six switches

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



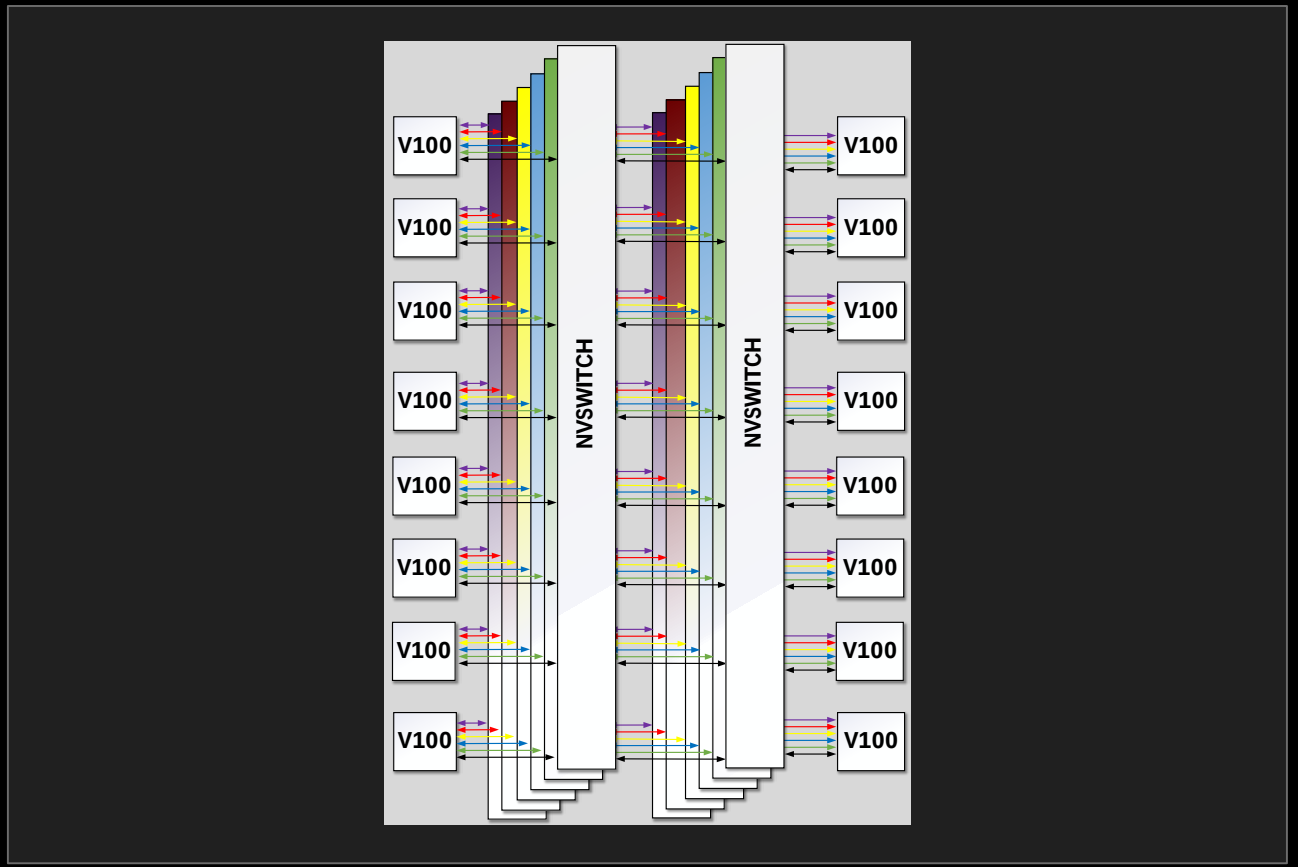
DGX-2 NVLINK FABRIC

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

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

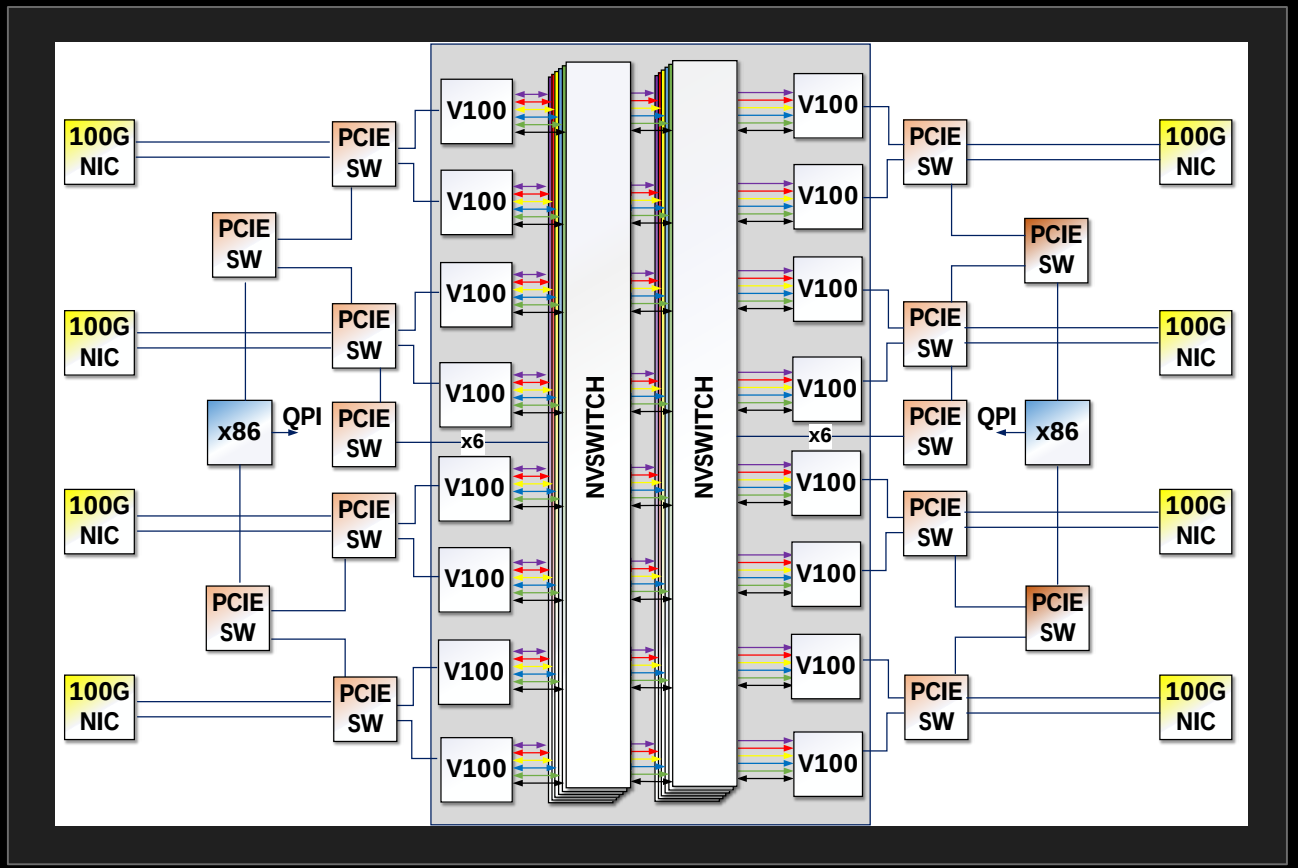
Parameter	DGX-2 Spec
CPUs	Dual Xeon Platinum 8168
CPU Memory	1.5 TB DDR4
Aggregate Storage	30 TB (8 NVMes)
Peak Max TDP	10 kW
Dimensions (H/W/D)	17.3”(10U)/ 19.0”/32.8” (440.0mm/482.3mm/834.0mm)
Weight	340 lbs (154.2 kgs)
Cooling (forced air)	1,000 CFM

DGX-2 PCIE NETWORK

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

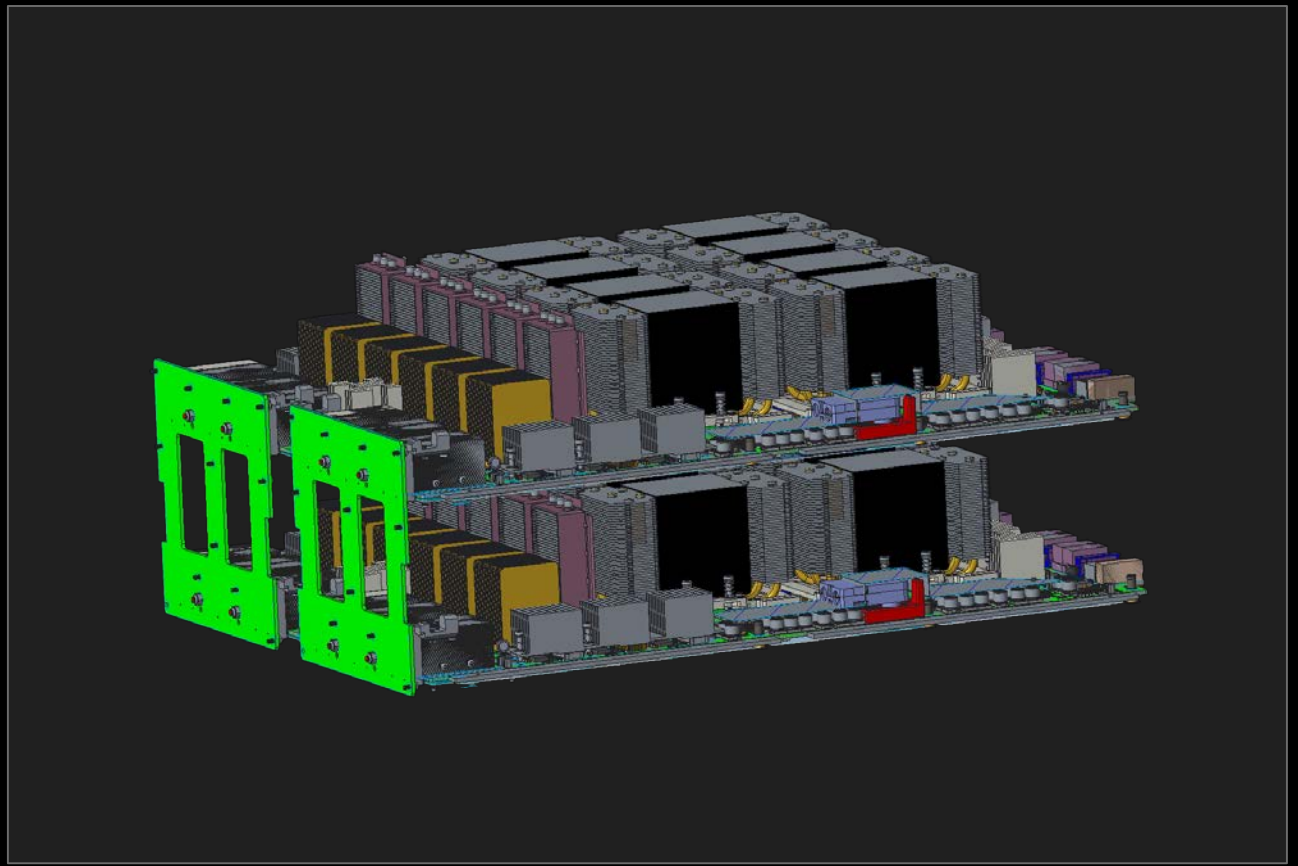


NVIDIA DGX-2: GPUS + NVSWITCH COMPLEX

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



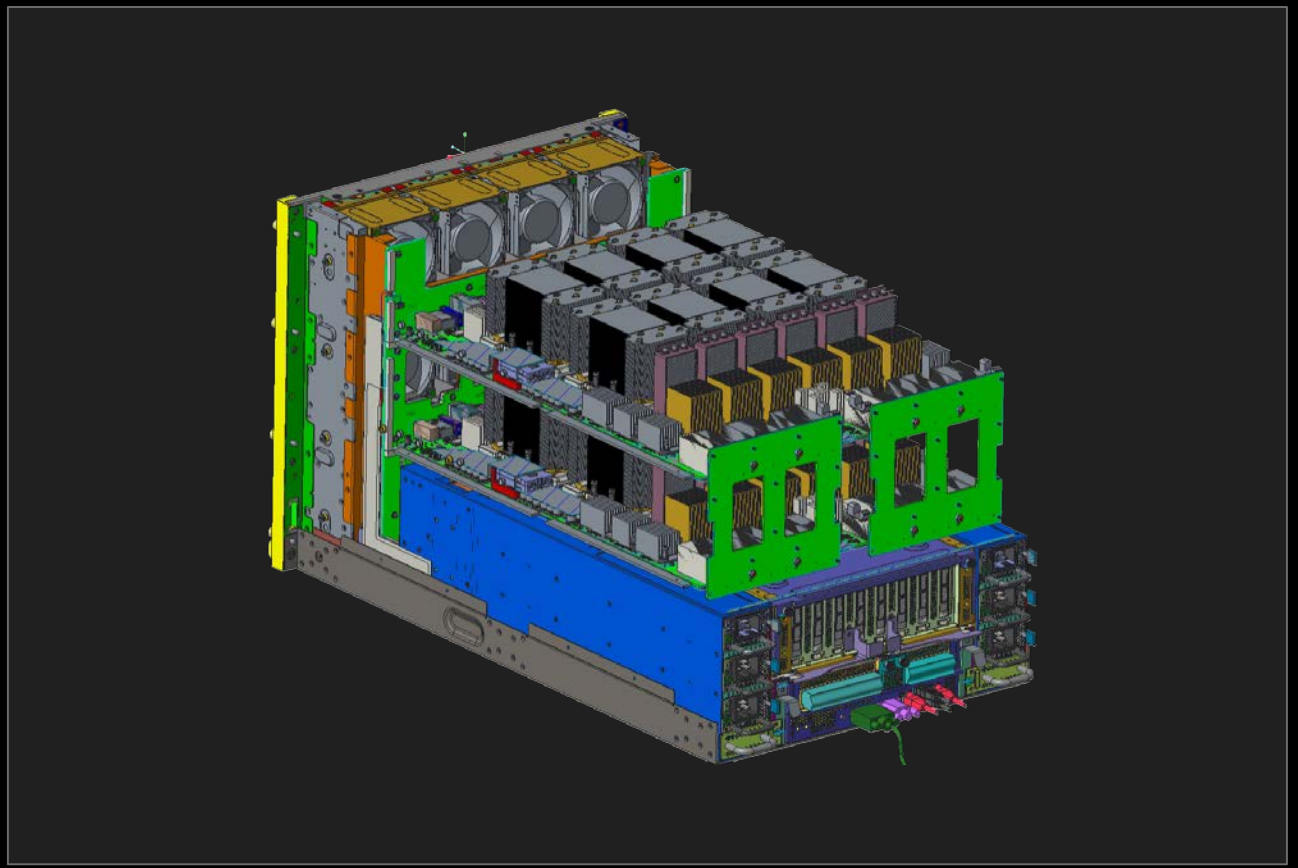
NVIDIA DGX-2: SYSTEM PACKAGING

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

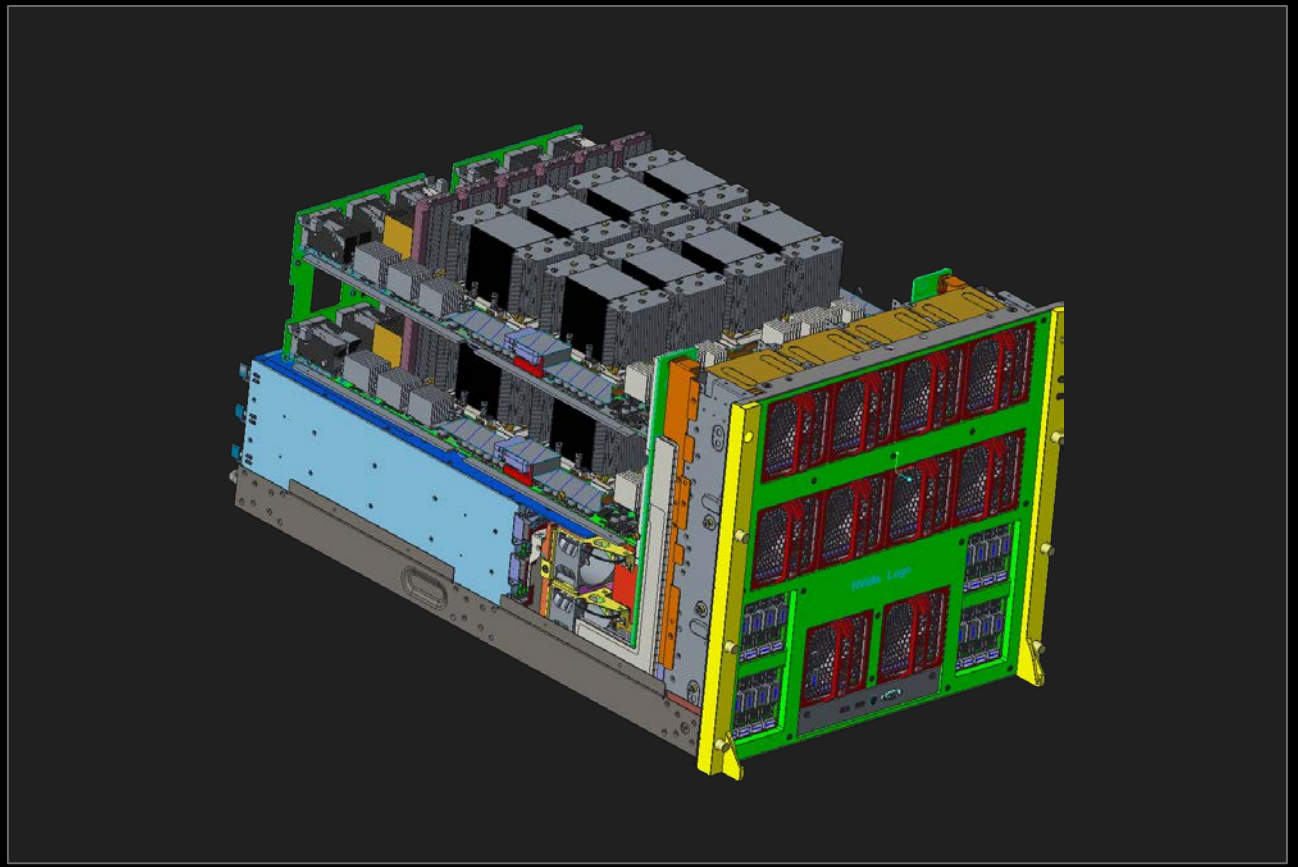


NVIDIA DGX-2: SYSTEM COOLING

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

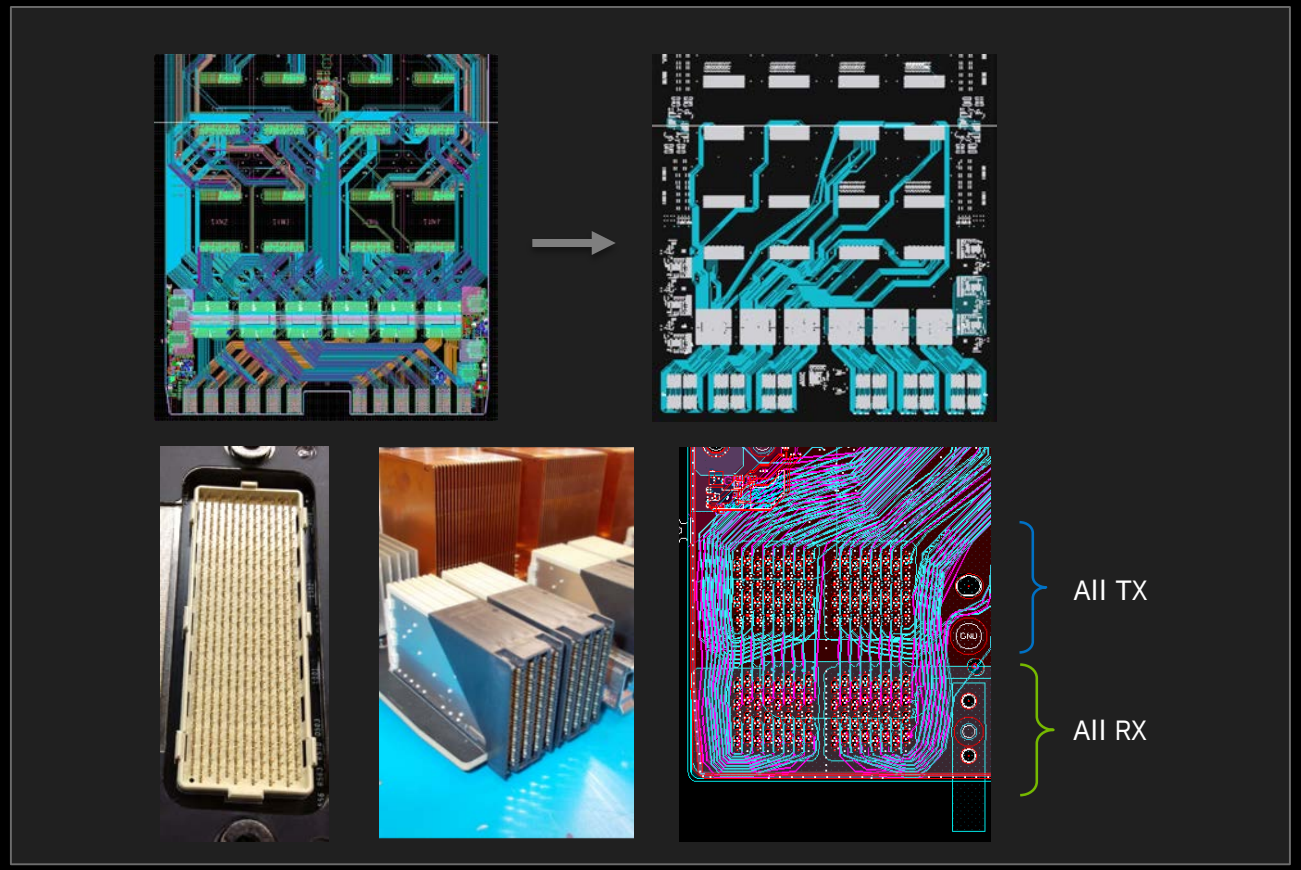


DGX-2: SIGNAL INTEGRITY OPTIMIZATIONS

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



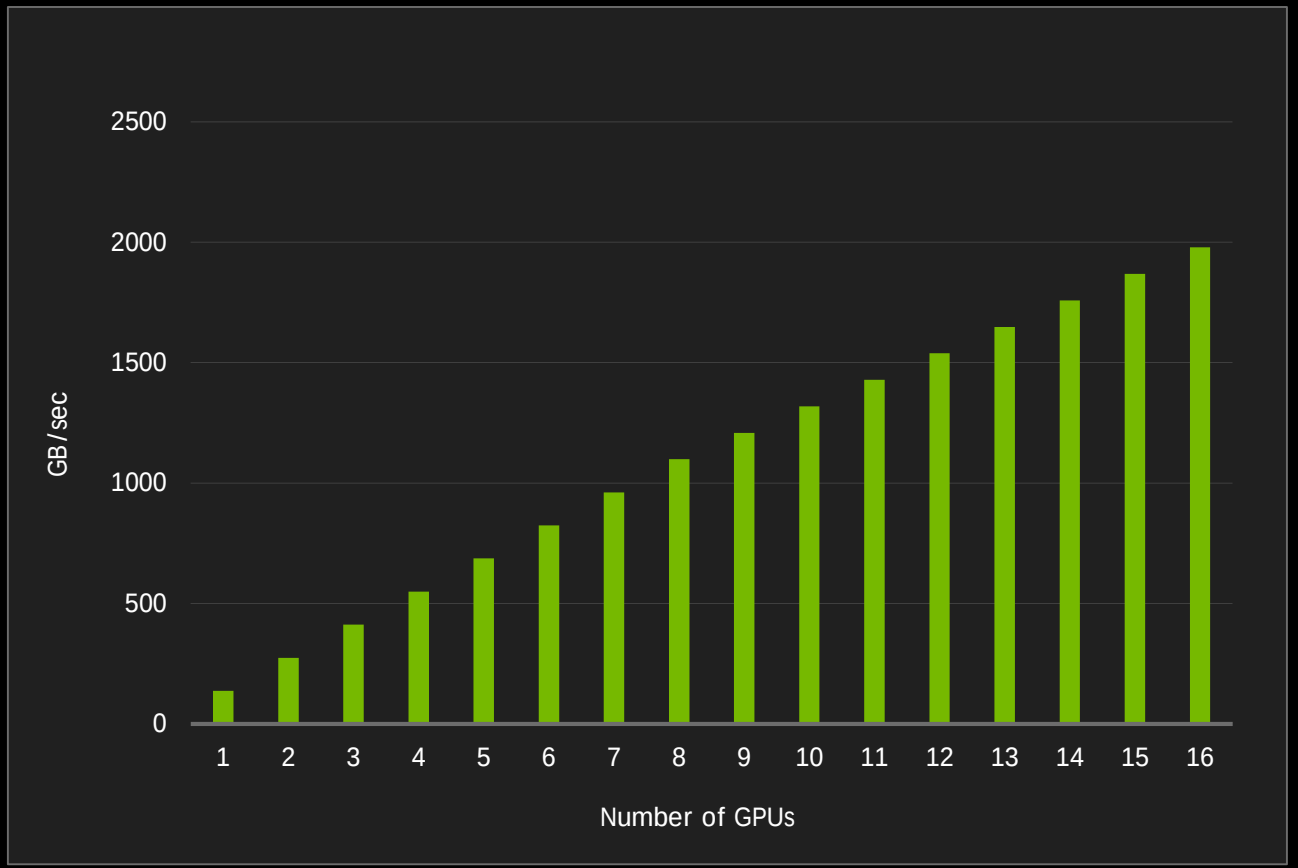
DGX-2: ACHIEVED BISECTION BW

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

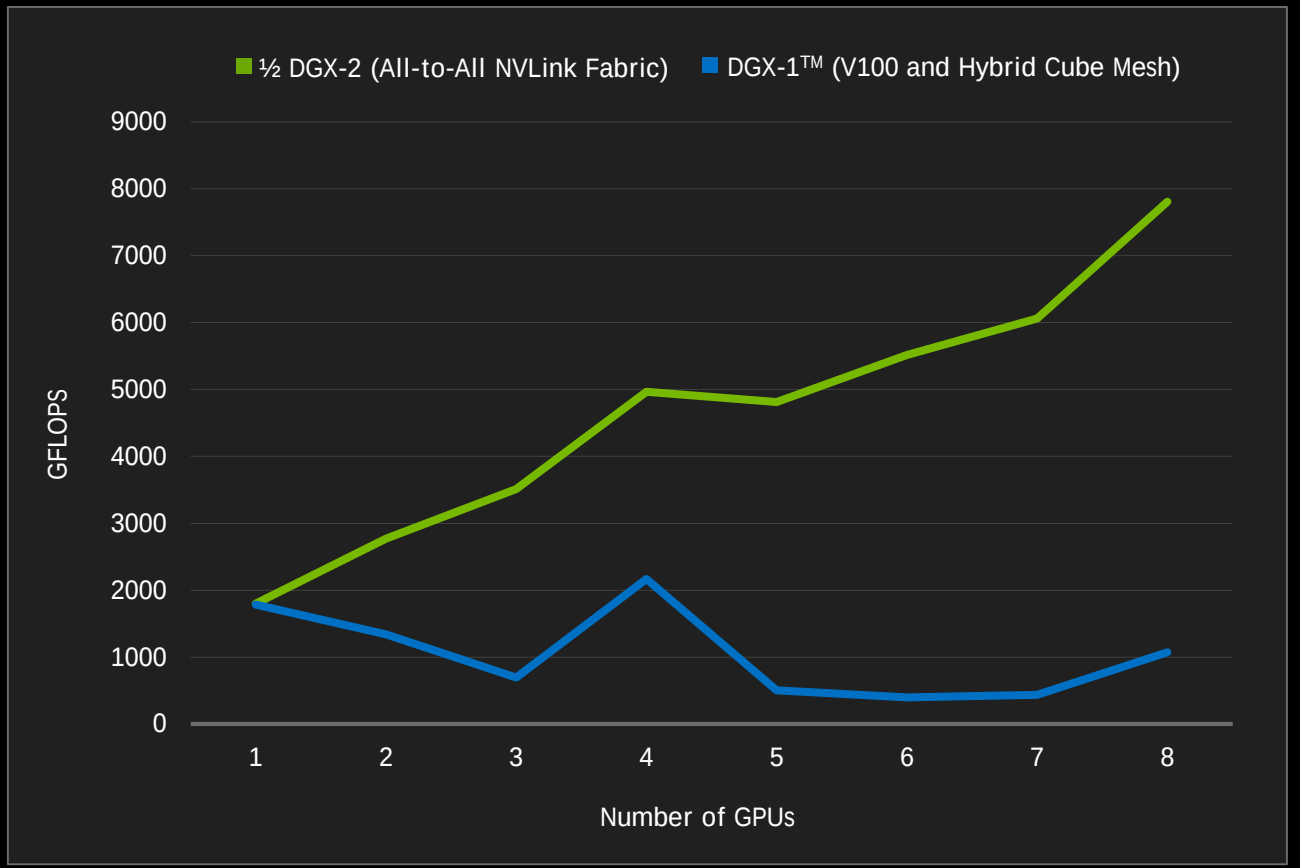


DGX-2: CUFFT (16K X 16K)

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

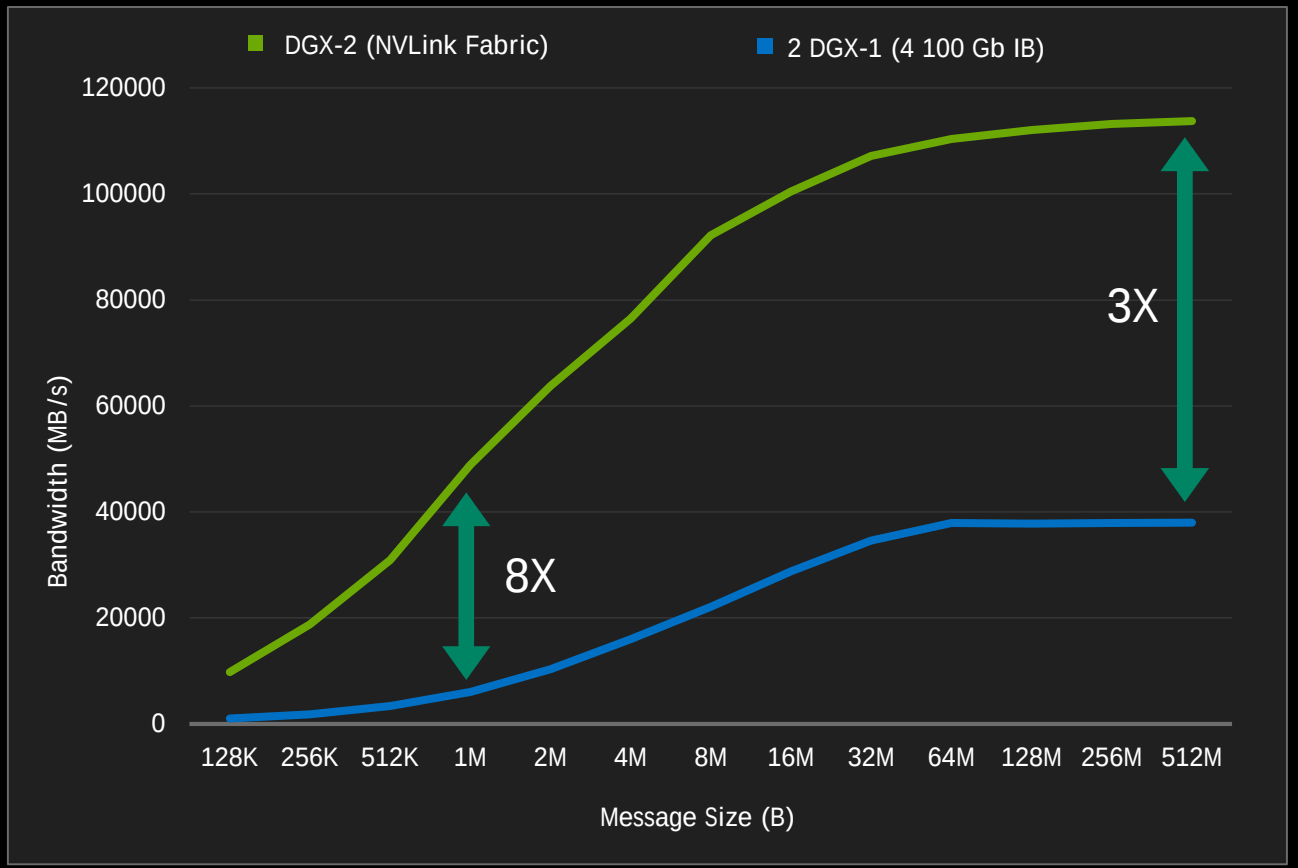


DGX-2: ALL-REDUCE BENCHMARK

Important communication primitive in Machine-Learning apps

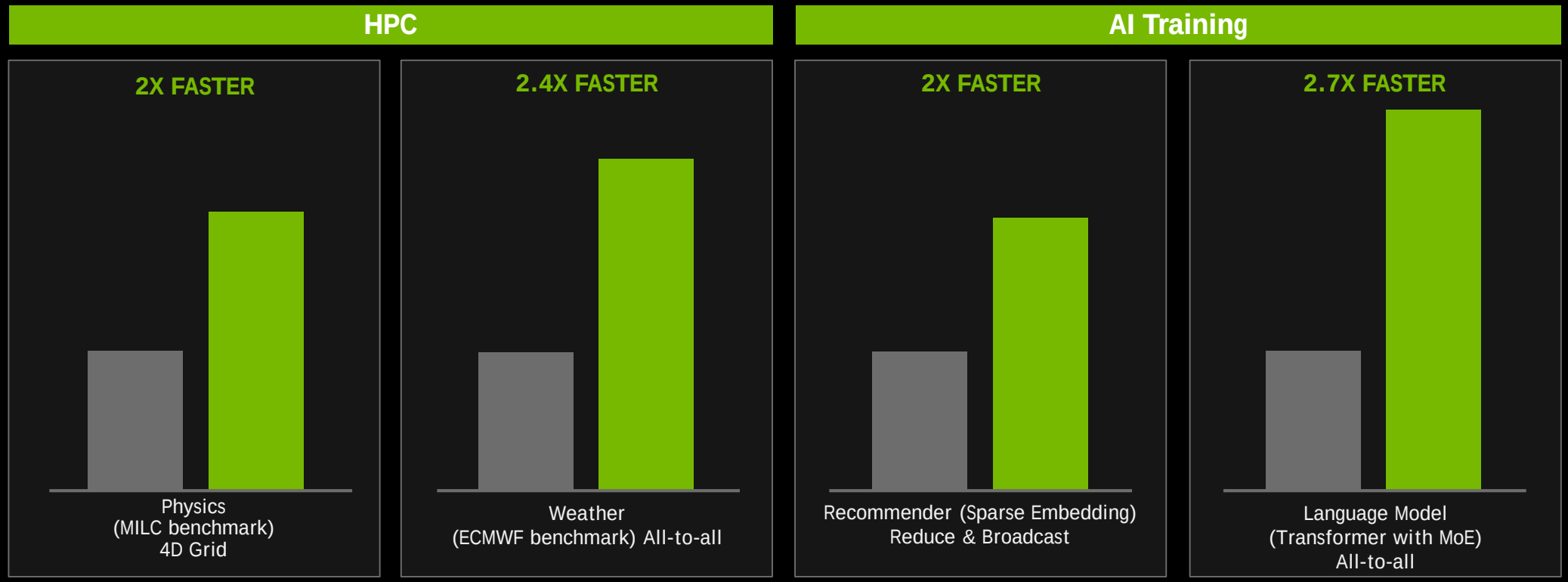
DGX-2 provides increased bandwidth and lower latency compared to two 8-GPU servers (connected with InfiniBand)

NVLink network has efficiency superior to InfiniBand on smaller message sizes



DGX-2: >2X SPEED-UP ON TARGET APPS

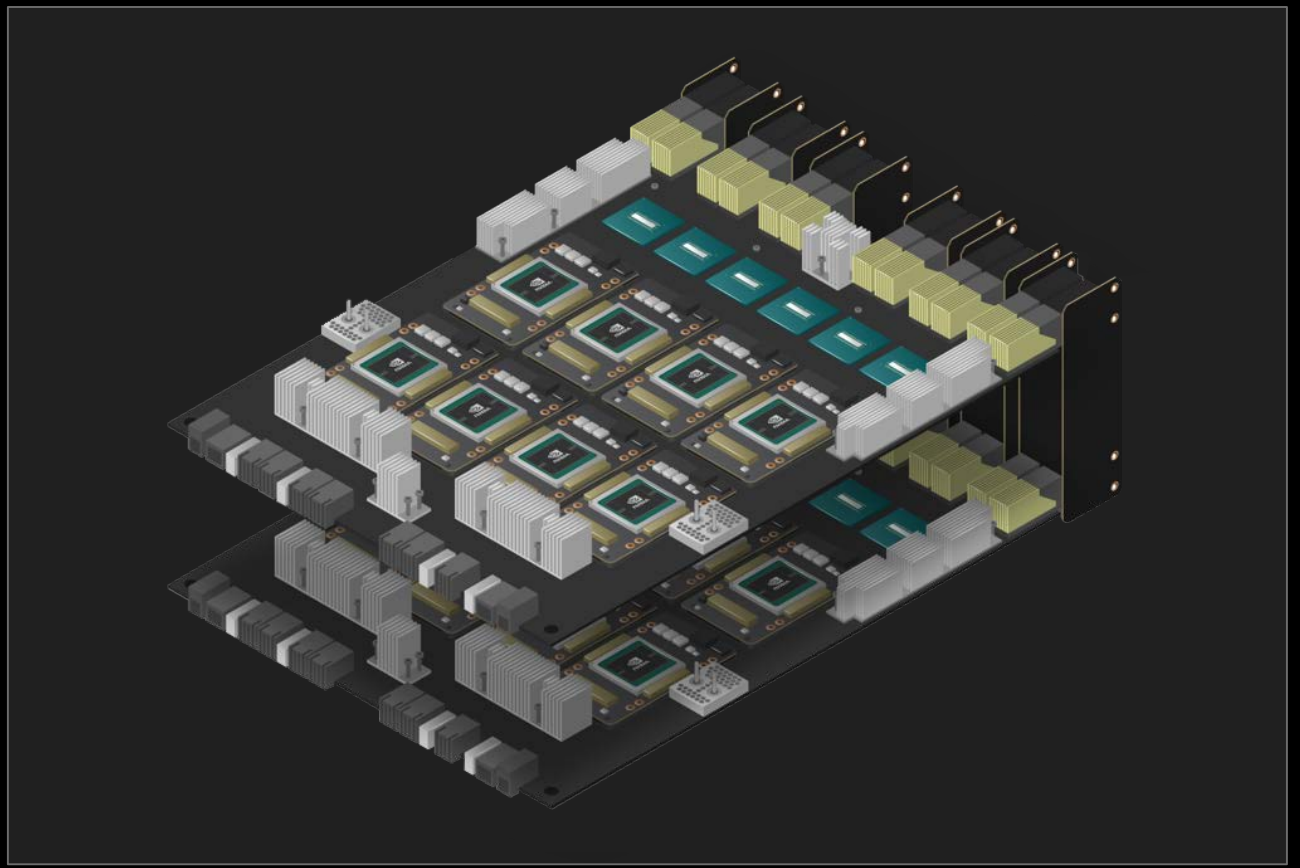
Two DGX-1™ Servers (V100) Compared to DGX-2 – Same Total GPU Count



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

NVSWITCH AVAILABILITY: NVIDIA HGX-2™

- 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



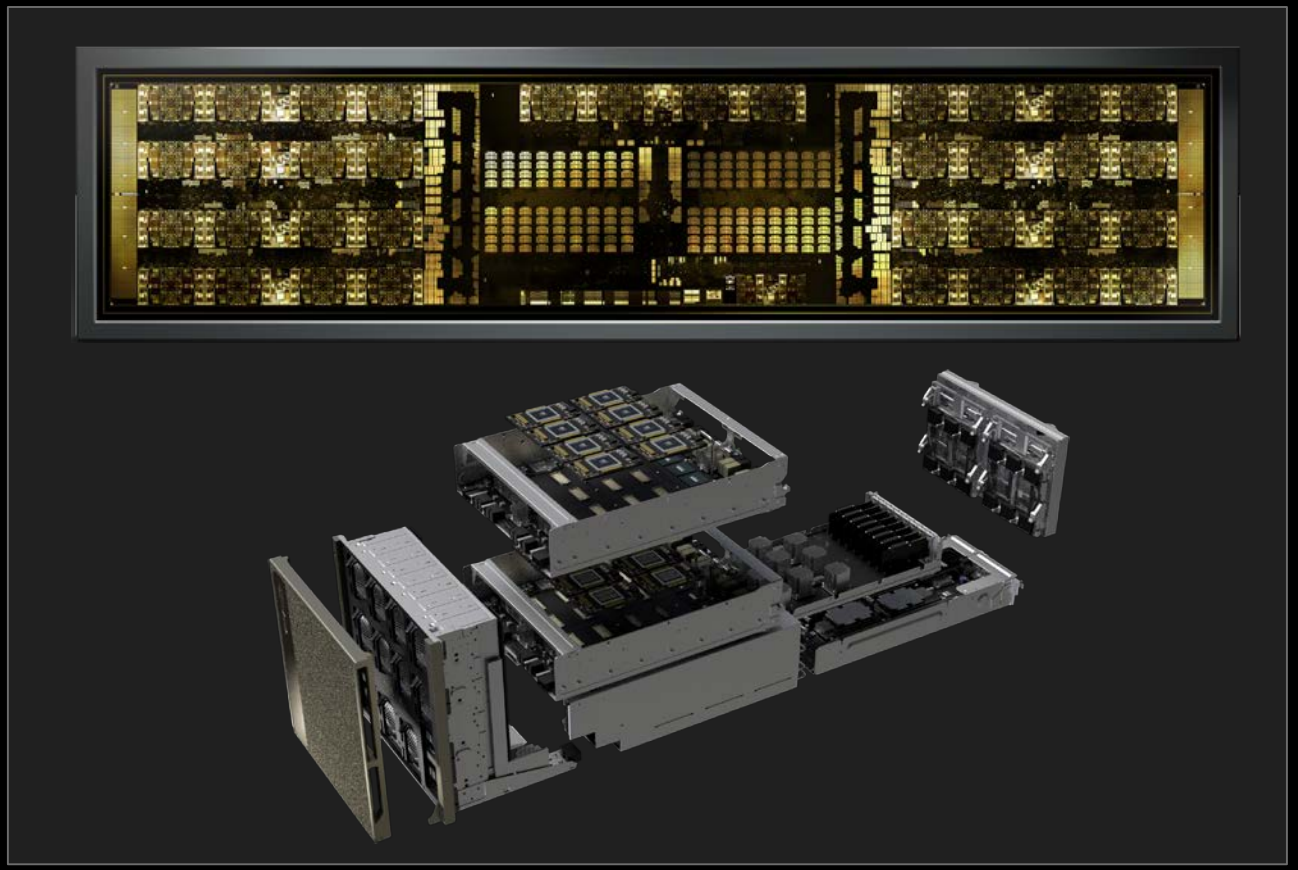
SUMMARY

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




NVIDIA