TECHNICAL OVERVIEW

NVIDIA NVSWITCH

The World's Highest-Bandwidth On-Node Switch

As deep learning neural networks become more sophisticated, their size and complexity continue to expand. And so do the available datasets they can ingest to deliver next-level insights. The result is exponential growth in required computing capacity to train these networks in a practical amount of time. To meet this challenge, developers have turned to multi-GPU implementations, which have demonstrated near-linear performance scaling. In these multi-GPU systems, one of the keys to continued performance scaling is flexible, high-bandwidth inter-GPU communications.

NVIDIA introduced NVIDIA® NVLink™ to connect multiple GPUs at 10X the PCIe bandwidth and boost computing capacity. This solution enabled 8 GPUs in a single server to be connected together in a point-to-point network called a hybrid cube mesh. This implementation was an important step forward and elevated the performance of 8 GPU servers. But in "all to all" communications, where all GPUs need to communicate with one another, this implementation requires certain GPU pairs to communicate over a much slower PCIe data path. To take GPU server performance to the next level and scale beyond 8 GPUs in a single server, a more advanced solution was needed.

Enter NVSwitch.

NVSwitch enables a fully NVLink-connected 16-GPU system with an uncompromised 300 GB/s of connectivity. This interconnect fabric eliminates bottlenecks and intermediary GPU hops to enable 16 GPUs to behave as one, unleashing an incredible 2 petaFLOPS of deep learning computing capacity to train the next generation of AI networks.

NVSwitch: The World's Highest-Bandwidth On-Node Switch

NVSwitch is an NVLink switch chip with 18 ports of NVLink per switch. Internally, the processor is an 18 x 18-port, fully connected crossbar. Any port can communicate with any other port at full NVLink speed, 50 GB/s, for a total of 900 GB/s of aggregate switch bandwidth.

VITAL STATISTICS:	
Port Configuration	18 NVLINK ports
Speed per Port	50 GB/s per NVLINK port (total for both directions)
Connectivity	Fully connected crossbar internally
Transistor Count	2 billion

Each port supports 25 GB/s in each direction. The crossbar is non-blocking, allowing all ports to communicate with all other ports at the full NVLink bandwidth. Consider Figure-1 shown below. All 8 GPUs on one baseboard are connected with a single NVLink to all 6 NVSwitches. Eight ports on each of the NVSwitches are used to communicate with the other baseboard. Each of the 8 GPUs on a baseboard can communicate with any of the others on the same baseboard at the full bandwidth of 300 GB/s with a single NVSwitch traversal. Each of the GPUs can also communicate at full bandwidth with any GPU on the second baseboard. In this case, there are two NVSwitch traversals. The bi-section bandwidth between the boards is 2.4 TB/s (48 links at 25 GB/s in each direction). Note that the NVIDIA DGX-2™ platform uses only 16 of the available ports per switch. The remaining ports are reserved.

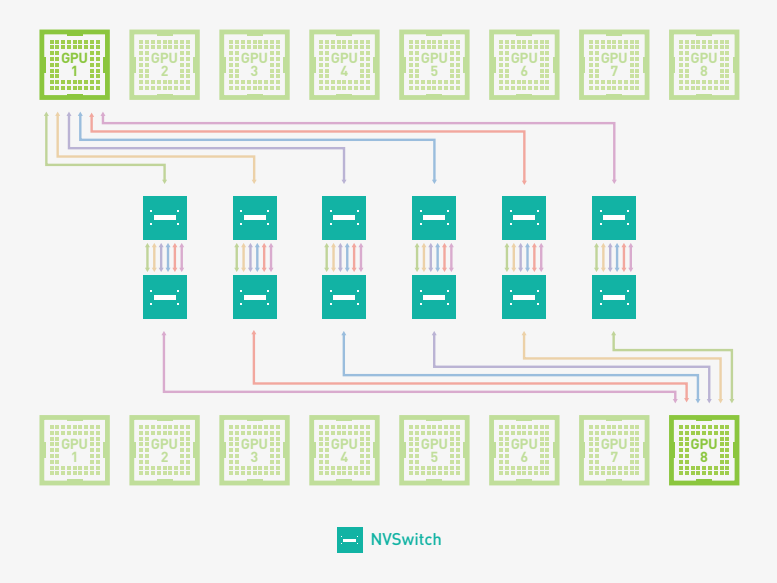

Figure-1: NVSwitch Topology Diagram - Two GPUs' connections shown for simplicity. All 16 GPUs connect to NVSwitch chips in the same way.

With such high-bandwidth data movement, data integrity is paramount. Data traversing NVLink itself is protected using cyclical redundancy coding (CRC) to detect errors and replay the transfer. NVSwitch's datapaths, routing, and state structures are protected using error-correcting codes (ECC). NVSwitch also supports final hop-address fidelity checks and buffer over- and underflow checks. For security, NVSwitch's routing tables are indexed and controlled by the NVIDIA fabric manager, providing protection by limiting an application's access to its specific ranges.

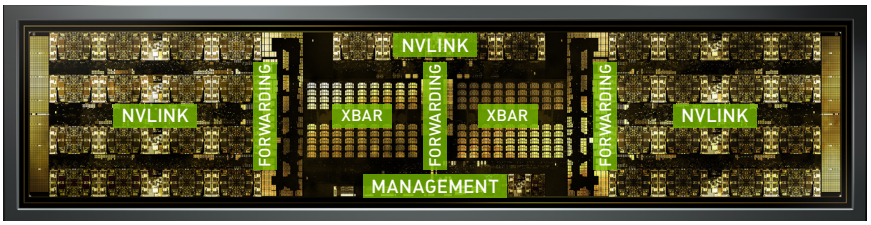

Figure-2: NVSwitch Die Shot

NVSwitch enables larger GPU server systems with 16 GPUs and 24X more inter-GPU bandwidth than 4X InfiniBand ports, so much more work can happen on a single server node. A 16-GPU server offers multiple advantages: It reduces network traffic hot spots that can occur when two GPU-equipped servers are exchanging data during a neural network training run. With an NVSwitch-equipped server like NVIDIA's DGX-2, those exchanges occur on-node, which also offers significant performance advantages. In addition, it offers a simpler, single-node programming model that effectively abstracts the underlying topology.

Chart 1: NVIDIA's new DGX-2 server with 16 GPUs is able to deliver up to 2.4X more high-performance computing (HPC) performance than two 8-GPU servers communicating over a 4X InfiniBand connection. The European Centre for Medium-Range Weather Forecasts' (ECMWF) mini-app workload executes a large number of Fast Fourier Transform (FFT) operations, which involve significant inter-GPU communication. The MILC-based workload is a numerical simulation application that uses a batched conjugate gradient (CG) solver to study quantum chromodynamics (QCD), the theory surrounding strong interactions of subatomic physics. This benchmark corresponds to a batched mixed-precision CG solver.

See notes about workload details.

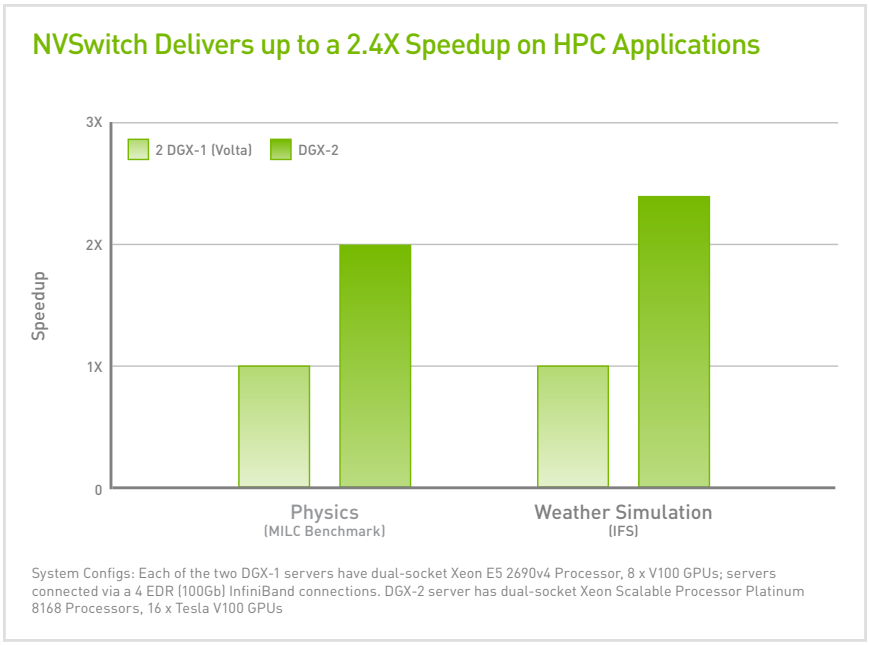

Chart-1

Chart 2: NVIDIA's new DGX-2 server with 16 GPUs is able to deliver up to 2.7X more deep learning training performance than two 8-GPU servers communicating over a 4X InfiniBand connection. The mixture of experts (MoE) workload is a combination of neural networks that collaborate to produce more sophisticated language translations. Sparse embedding networks are used for recommender systems to match users with relevant product and service information.

See notes about workload details.

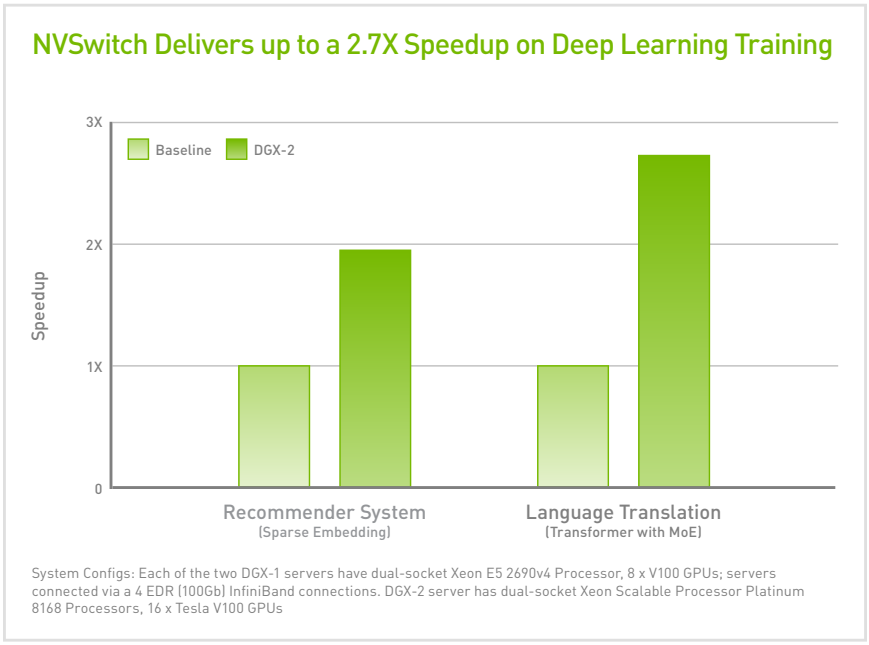

Chart-2

Ready to Tackle Tomorrow's Workloads Today

With the continuing explosive growth of neural networks' size, complexity, and designs, it's difficult to predict the exact form those networks will take, but one thing remains certain: the appetite for deep learning compute will continue to grow along with them. In the HPC domain, workloads like weather modeling using large-scale, FFT-based computations will also continue to drive demand for multi-GPU compute horsepower. And with a 16-GPU configuration packing a half-terabyte of GPU memory in a unified address space, applications can scale up without requiring knowledge of the underlying physical topology.

Similarly, HPC and graph analytics workloads continue to grow and take advantage of GPU acceleration. NVSwitch, which lies at the heart of NVIDIA's new DGX-2 server, is the critical connective tissue that enables 16 GPUs in a single server to accelerate even the most aggressive workloads, and bring the next wave of deep learning innovation.

To learn more about NVIDIA Tesla® V100 and the Volta architecture, download the **VOLTA ARCHITECTURE WHITEPAPER**.

To learn more about NVIDIA's new DGX-2 server, visit the **DGX-2** page.

WORKLOAD NOTES:

MILC Benchmark: A numerical simulation application that uses a batched CG solver to study quantum chromodynamics (QCD), the theory surrounding strong interactions of subatomic physics. This benchmark corresponds to a batched mixed-precision conjugate gradient solver that includes 64-bit floating-point (FP64) double-precision and FP16 half-precision calculations. This kind of algorithm can be used in the analysis phase of lattice QCD. The problem size here is 48 x 48 x 48 x 64. The high dimension requires each GPU to fetch data from many of its neighbors to execute its computations, creating all-to-all traffic that NVSwitch dramatically accelerates.

ECWMF's IFS: A global numerical weather prediction model developed by the European Centre for Medium-Range Weather Forecasts (ECMWF) based in Reading, England. ECMWF is an independent intergovernmental organization supported by most of the nations of Europe, and it operates one of the largest supercomputer centers in Europe for frequent updates of global weather forecasts. The Integrated Forecasting System (IFS) mini-app benchmark focuses its work on a spherical harmonics transformation that represents a significant computational load of the full model. The benchmark speedups shown in the graph are better than those for the full IFS model, since the benchmark amplifies the transform stages of the algorithm (by design). However, this benchmark demonstrates that ECMWF's extremely effective and proven methods for providing world-leading predictions remain valid on NVSwitch-equipped servers such as NVIDIA DGX-2, since they're such a good match to the problem.

Recommender: A mini-app built at NVIDIA that's modelled after Alibaba's paper "**Deep Interest Network for Click-Through Rate Prediction.**" The mini-app uses a batch size of 8,192, indexing into a 1 billion-entry embedding table. Each entry is 64 dimensions wide with each dimension in FP16 single-precision. The resultant data table requires more than 256 GB of memory, so this use case requires at least two DGX-1V 32 GB servers or a DGX-2 to run. The app includes reduce and broadcast operations between GPUs that NVSwitch accelerates. The performance of this workload is driven by how many embedding table lookups a system can deliver, hence the metric used is billions of lookups per second.

Mixture of Experts (MoE): Based on a network published by Google at the Tensor2Tensor github, this workload uses the Transformer model with MoE layers. The MoE layers each consist of 128 experts, each of which is a smaller feed-forward deep neural network (DNN). Each expert specializes in a different domain of knowledge, and the experts are distributed to different GPUs, creating significant all-to-all traffic due to communications between the Transformer network layers and the MoE layers. The training dataset used is the "1 billion-word benchmark for language modeling" according to Google. Training operations use Volta Tensor Core and run for 45,000 steps to reach perplexity equal to 34. This workload uses a batch size of 8,192 per GPU.