REFERENCE ARCHITECTURE

HYPERSCALE AIRI™

RACK-SCALE, AI-READY INFRASTRUCTURE ARCHITECTED BY PURE STORAGE, NVIDIA, AND MELLANOX

PURESTORAGE®

Powered by NVIDIA.

TABLE OF CONTENTS

INTRODUCTION

Does the rising complexity of infrastructure cause a significant delay in bringing your AI-at-scale initiative to production? Is your data team held back by a DIY infrastructure that doesn't scale with your AI needs? Learn how to deploy a fully integrated, rack-scale solution incorporating multiple racks of NVIDIA® DGX™ servers and Pure Storage FlashBlade™ arrays to improve time-to-insight and drive success in these crucial, investment-heavy projects. Hyperscale AIRI™ is architected by Pure Storage® in partnership with the leaders of AI and supercomputing infrastructure, NVIDIA®, and Mellanox®. Hyperscale AIRI is designed to offer a modular, streamlined approach to building AI infrastructure that can scale to 64 racks of DGX servers and FlashBlades, interconnected by Mellanox Spectrum™ Ethernet switches.

In March 2018 we announced AIRI, the industry's first comprehensive AI-Ready Infrastructure, architected by NVIDIA and Pure Storage. AIRI is purpose-built to enable organizations and their IT teams to extend the power of the DGX-1 servers and operationalize AI-at-scale for every data scientist, researcher, and developer in the enterprise.

Since the announcement, AIRI has been widely deployed across the globe in many industries, including healthcare, financial services, automotive, technology companies, higher education, and research. As AIRI has enabled these organizations to turn data into innovation at an unprecedented pace, they have needed a solution that could scale to multiple racks of DGX servers and FlashBlade storage. However, such rack-scale infrastructure needs to address several additional complexities, including:

- **Multi-Rack Infrastructure Optimization** – Design consideration of compute, storage, and network across racks to deliver maximum performance with no data bottlenecks.
- **Software Stack Support for AI Workflow** – Keeping data science teams productive with a robust software stack for the entire AI pipeline: data collection, data cleansing, data labeling, model exploration, training, and inference.
- **Power & Cooling Considerations** – Utilizing high-density racks to provide the most efficient use of costly data center floor space.

Hyperscale AIRI is already validated and deployed in production by some of the pioneering enterprises of real-world AI.

Designing, configuring, and maintaining infrastructure to satisfy the challenges of large-scale deep learning (DL) requires a significant investment of time and resources to avoid unforeseen delays, bottlenecks, or downtime. Engineers at NVIDIA, Pure Storage, and Mellanox have worked to deliver a fully integrated Hyperscale AIRI that offers scale-out DL out-of-the-box, with time-to-insight in hours rather than multiple weeks.

ACCELERATING MODEL EXPLORATION & TRAINING: NVIDIA DGX-1 AND DGX-2 SERVERS

Originally designed for computer graphics, NVIDIA engineers tapped into the massively parallel architecture of the modern graphics processing unit (GPU) and built a full-stack optimized solution to deliver the performance demanded by the data-parallel, computationally-intensive algorithms that enable DL today. Powered by NVIDIA-engineered, optimized software frameworks to maximize underlying computational capability, NVIDIA GPUs have become the de facto computational platform of AI and DL.

While the advances in GPU performance have been impressive, an AI-infused enterprise requires a state-of-the-art architecture that scales in a predictable, cost-effective way, while ensuring compute capacity for critical workloads. NVIDIA developed the NVIDIA Tesla™ series of GPU accelerators and state-of-the-art GPU interconnection technologies – NVIDIA NVlink and NVIDIA NVSwitch™ – specifically for dense compute, data center-scale systems. However, building a platform maximized for GPU-accelerated performance demands the optimal integration of hardware and software, one that's purpose-built for DL. To simplify the process of selecting and integrating compute hardware with complex DL software, the NVIDIA DGX-1 and NVIDIA DGX-2™ servers were created.

Each DGX-1 server packs the AI computational power of 800 CPUs, delivering one petaFLOPS performance in a 3U appliance, pre-configured with eight Tesla V100 Tensor Core GPUs. DGX-2 servers, on the other hand, integrate 16 Tesla V100 GPUs, using state-of-the-art NVSwitch technology to deliver the low-latency, high-bandwidth inter-GPU communication required for greater levels of model and data parallelism. For multi-node training, both DGX-1 and DGX-2 servers communicate over RDMA networks with an aggregate bandwidth of 400 Gbps and 800 Gbps respectively. In addition, DGX-optimized software frameworks and applications ensure maximized DL training performance.

ACCELERATING DATA ACCESS: PURE STORAGE FLASHBLADE

DL requires more than fast compute and high-bandwidth interconnects. When designing a full-stack platform for large-scale DL, the system architect's goal is to provision as many GPUs as possible, while ensuring linearity of performance as the environment is scaled, all the while keeping the GPUs fed with data. Keeping the GPUs fed requires a high-performance data pipeline all the way from storage to GPUs. When defining storage for deep-learning systems, architects must consider three requirements:

- **Diverse Performance** – DL often requires multi-gigabytes-per-second I/O rates but isn't restricted to a single data type or I/O size. Training deep neural network models for applications as diverse as machine vision, natural-language processing, and anomaly detection requires different data types and dataset sizes. Storage systems that fail to deliver the performance required during neural network training will starve the GPU tier for data, and prolong the length of the run, inhibiting developer productivity and efficiency. Providing consistency of performance at various IO sizes and profiles at a capacity scale will ensure success.
- **Scalable Capacity** – Successful machine learning projects often have ongoing data acquisition and continuous training requirements, resulting in continued growth of data over time. Furthermore, enterprises that succeed with one AI project find ways to apply these powerful techniques to new application areas,

resulting in further data expansion to support multiple use cases. Storage platforms with inflexible capacity limits result in challenging administration overheads to federate disparate pools.

- **Strong Resiliency –** As the value of AI grows within an organization, so does the value of the infrastructure supporting its delivery. Storage systems that result in excessive downtime or require extensive administrative outages can cause costly project delays or service disruptions.

Existing storage systems sacrifice one or more of these dimensions or force architects and administrators to suffer through excessive deployment and management complexity.

Initial deployments for DL often start with direct-attached storage (DAS), resulting in hard capacity limits and challenges in sharing data sets across multiple compute units. Collecting multiple DAS servers into a shared file system with the Hadoop Distributed File System (HDFS) can alleviate the capacity concerns, but comes at a stark performance cost for small, random I/O patterns that are common in many DL use cases. Furthermore, burdening the CPUs in a GPU server with storage management can lead to bottlenecks in the overall pipeline and poor resource utilization.

Parallel file systems such as GPFS and Lustre, designed specifically for the needs of high-performance computing (HPC), can be tuned by expert-level administrators to meet the requirements of a particular workload. However, a new data set or training paradigm inevitably requires a new configuration and tuning process, resulting in project delays and potential stranded capacity. Traditional NAS offerings can provide strong resilience and scalable capacity but often fail to deliver the performance required across a range of I/O patterns and at large-scale compute clusters.

Pure Storage FlashBlade™, with its scale-out, all-flash architecture and a distributed file system purpose-built for massive concurrency across all data types, is the only storage system to deliver on all of these dimensions while keeping required configuration and management complexity to a bare minimum. In addition, with multi-chassis support, FlashBlade seamlessly scales from TBs to PBs all in one namespace.

ACCELERATING DATA MOVEMENT: MELLANOX SPECTRUM SWITCHES

Mellanox accelerates data movement with a unique family of composable switches & NICs using silicon purpose-built to bring a level of performance and visibility not possible with off-the-shelf silicon. High performance compute platforms including DGX servers use Mellanox ConnectX adapter cards for external connectivity. Following are the key attributes that make Mellanox Spectrum Ethernet switches ideal for AI workloads:

- **Consistent Performance** – Mellanox Spectrum switches coupled with ConnectX adapters leverage hardware-accelerated end-to-end congestion management to provide a robust data path for RoCE based GPUDirect traffic. Mellanox Spectrum Ethernet switches support high bandwidth storage traffic with fair and predictable performance.
- **Intelligent Load Balancing** – Mellanox Spectrum switches support Adaptive Routing (AR) to maximize cross-sectional bandwidth in data center fabrics. AR leverages end-to-end congestion notification mechanisms

to maximize the cross-sectional bandwidth of fabric. Maximizing cross-sectional bandwidth reduces the remote access-related performance penalty that often limits scale-out system performance.

- **Hardware Accelerated Visibility** – Mellanox Spectrum provides detailed and contextual telemetry to answer the “When, What, Who, Where and Why” questions as soon as an issue happens. Hardware-accelerated histograms track and summarize queue depths at a sub-microsecond granularity. This avoids false-alerts common to simple watermarks/thresholds.

ACCELERATING TIME-TO-VALUE IN AI: HYPERSCALE AIRI™

Launched in March 2018, AIRI is a converged infrastructure stack, purpose-built solution for DL at-scale. Now engineers at NVIDIA, Mellanox, and Pure Storage have come together to architect Hyperscale AIRI to offer supercomputing capabilities for the pioneering enterprises of AI. With Hyperscale AIRI, AI visionaries can now build the most advanced neural networks with a multi-rack, scale-out DL solution. In addition, Hyperscale AIRI keeps the data team productive at any scale and empowers them to quickly iterate over various models and deploy in production in days rather than months.

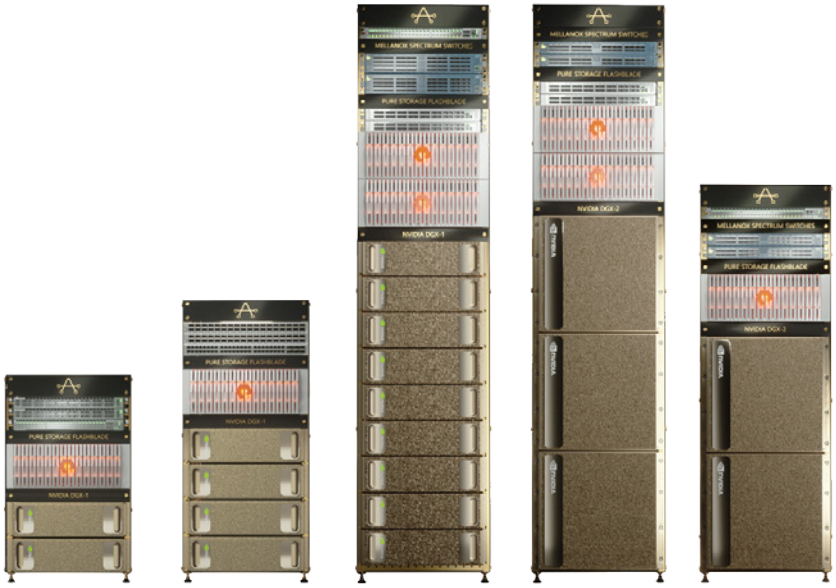


FIGURE 1. AIRI portfolio

AIRI’s rack-scale architecture, which is already deployed in production by leading companies and research labs, is configured and tested as a complete solution, avoiding the intricate configuration and tuning required otherwise. In addition, with AIRI’s flexibility, enterprises can scale based on their growing AI initiative: from a compact AIRI “Mini” to Hyperscale AIRI with no downtime or data migration.

Hyperscale AIRI brings together all the benefits of the DGX-1 and DGX-2 servers, and Pure Storage FlashBlade, wrapping them in a high-bandwidth, low-latency Ethernet or InfiniBand fabric by Mellanox that unifies storage and compute interconnects with RDMA-capable 100Gb/s network. AIRI enables seamless scaling for both GPU servers and storage systems. As compute demands grow, additional DGX-1 and DGX-2 servers can be provisioned in the high-performance fabric and instantly access all available datasets. Similarly, as storage capacity or performance demands grow, additional blades can be added to the FlashBlade systems addressing petabytes of data in a single namespace with zero downtime or re-configuration.

AIRI SOFTWARE STACK

The software stack in AIRI is designed to keep the data science team productive at any scale and deliver capabilities to accelerate their AI projects. The AIRI software stack contains three key components:

NVIDIA GPU Cloud (NGC)

An optimized container registry that provides a comprehensive catalog of GPU-accelerated AI frameworks such as TensorFlow, PyTorch, MXNet, NVIDIA TensorRTTM, open-source RAPIDS, and more. NGC helps data scientists, software developers, and researchers to rapidly build, train, and deploy an end-to-end AI pipeline. More information on NGC and documentation available at ngc.nvidia.com.

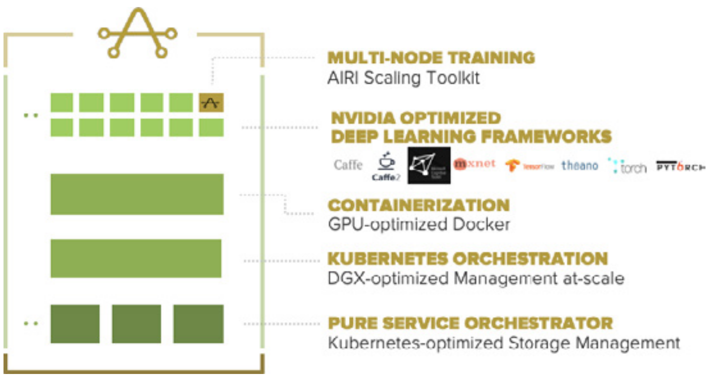


FIGURE 2. AIRI Software Stack

AIRI Scaling Toolkit

A toolkit with pre-optimized configuration and parameters to support multi-DGX server training using Horovod and OpenMPI. The AIRI Scaling Toolkit abstracts several complex parameters required to run large training jobs with popular AI frameworks and delivers linear scalability as data teams scale their large model training jobs with multiple DGX-1 and DGX-2 servers.

Kubernetes

Kubernetes provides a high level of flexibility in load balancing and node failover, easily integrating with other modern analytics applications. With Kubernetes, a data science team can quickly automate the deployment and management of containerized AI workflows, including feature extraction and data collection verification and analysis. For persistent storage to Kubernetes, FlashBlade supports both S3 and NFS, which can be easily configured, or Pure Service Orchestrator for more complex AI pipelines.

In addition to Kubernetes for faster model exploration and tuning, Slurm-managed domains can also co-exist in AIRI, and idle systems can be moved back and forth between these two environments. Slurm supports a more static cluster environment but provides advanced HPC-style batch scheduling features that simplify the multi-node training that some teams require for large-scale training jobs.

More information on the Kubernetes management stack and documentation are available as an open source project on GitHub.

SYSTEM ARCHITECTURE

DL training workloads are one of the most challenging workloads that run in data centers. Even the fastest, biggest, and most expensive scaled-up systems available today will not have the capacity to handle modern DL workloads. In hyper-growth environments that have DL workloads, it is critical that the IT team grow their infrastructure by scaling out horizontally. Scaling out entails interconnecting regular and repetitive compute and storage building blocks to form a larger coherent distributed system. Scale-out systems deliver exceptional performance and can be expanded to meet future demands. The architecture for Hyperscale AIRI is designed for two types of workloads:

- **Model Intensive** – Each AIRI rack has the highest density of compute, storage, and network scaled in a modular fashion
- **Data Intensive** – Separate racks of either high-density compute or high-density storage that can be scaled independently

While several network topologies were considered to deliver maximum scalable performance across racks for both these workloads, we believe that a non-blocking leaf-spine topology delivers the maximum training performance and faster data ingestion (See the **Network Architecture** section for more details). The architecture of a single rack of AIRI with DGX-1 servers for both model-intensive and data-intensive workloads are shown below. AIRI in both configurations can be seamlessly scaled to 64 racks of DGX servers and FlashBlades:

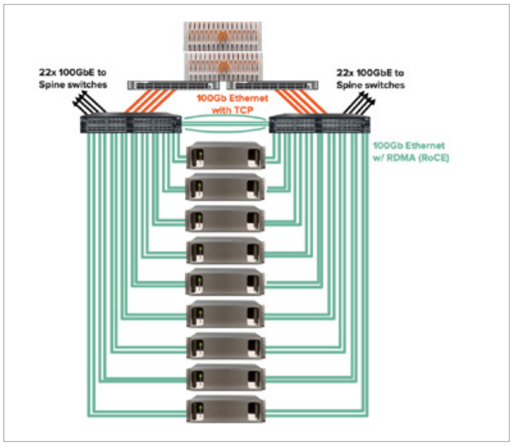


FIGURE 3. AIRI Rack with DGX-1 servers for Model-Intensive Workload

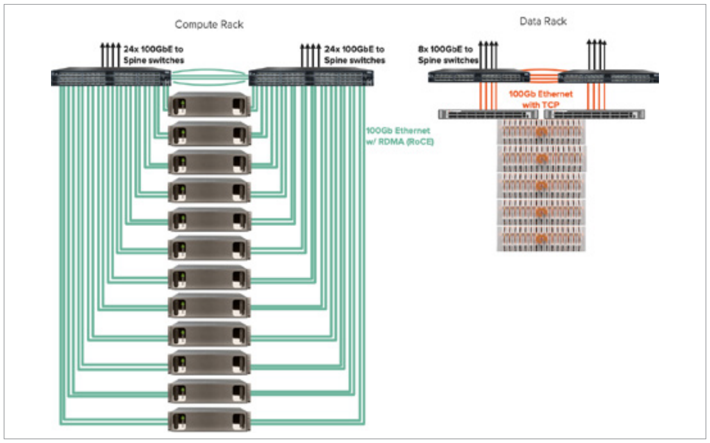


FIGURE 4. AIRI Racks with DGX-1 servers for Data-Intensive Workload

A model intensive AIRI rack with DGX-1 servers is optimized to support up to nine DGX-1 servers and two FlashBlades (unified into a single, multi-chassis namespace via Pure Storage External Fabric Modules) interconnected by 64-port Top-of-Rack (ToR) Mellanox SN3800 switches. A data-intensive AIRI is optimized to support up to 12 DGX-1 servers in a compute rack and up to 10 petabytes of flash storage (up to a five FlashBlade chassis) in a data rack. The ToR switches in the compute and storage racks are 64-port Mellanox SN3800 switches and 32-port Mellanox SN2700 switches, respectively. In addition, multiple SN3800 switches are used as spine switches for both these workloads.

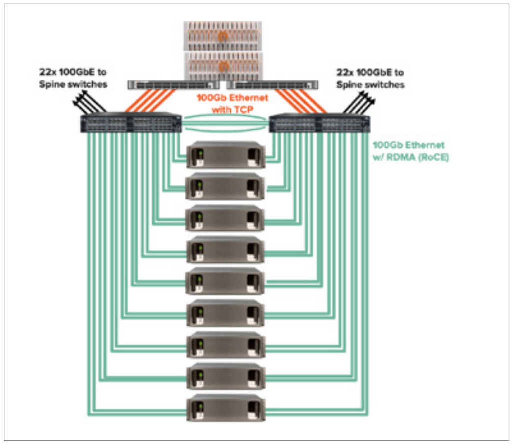


FIGURE 5. AIRI Rack with DGX-2 servers for Model Intensive Workload

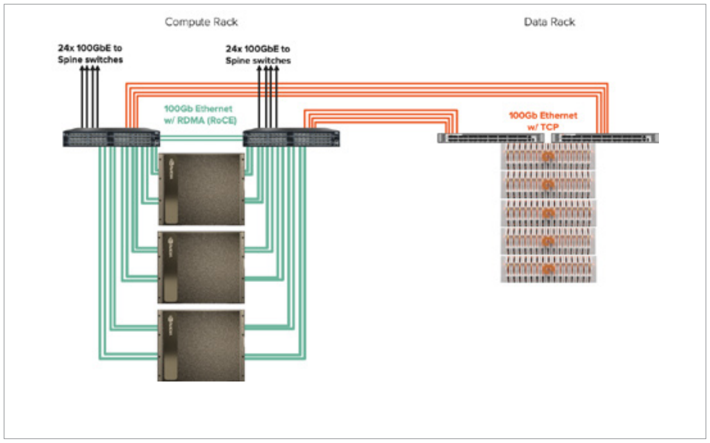


FIGURE 6. AIRI Rack with DGX-2 servers for Data Intensive Workload

The architecture of a single rack of AIRI with DGX-2 servers for both Model Intensive and Model & Data Intensive workloads are shown in Figure 5 and Figure 6 respectively. A model intensive AIRI rack with DGX-2 servers is optimized to support up to three DGX-2 servers and two FlashBlades (with multi-chassis) interconnected by 64-port Top-of-Rack (ToR) Mellanox SN3800 switches. A data intensive AIRI is optimized to support up to three DGX-2 servers in compute rack and up to 10 petabytes of flash storage in a data rack. The shared ToR switches in compute and storage racks are 64-port Mellanox SN3800 switches.

The connectivity between the storage and Mellanox Ethernet switches is through FlashBlade's eXternal Fabric Module (XFM). XFM simplifies the FlashBlade storage and exposes as a single namespace to the DGXs. The XFM module is within the product boundary of FlashBlade and is managed by the Purity//FB operating system. Hence, even when more storage is added, it requires no further additional cabling between the leaf-spine switches. In addition, to scale across multiple AIRI racks, SN3800 switches may be used as spine switches for both these workloads.

The Hyperscale AIRI architecture is designed to scale across racks for large DL workloads and is not restricted to these sizes. As datasets and workload requirements scale, additional DGX-1 and DGX-2 servers can be provisioned and instantly access all available data. Similarly, as storage capacity or performance demands grow, additional blades – or additional FlashBlade chassis – can be added to the FlashBlade system with zero downtime or re-configuration.

NETWORK ARCHITECTURE

DL workloads are characterized by high bandwidth GPU-to-GPU and GPU-to-Storage communications. Libraries such as NVIDIA Collective Communications Library (NCCL) implement smart topology discovery to optimize complex collective operations across GPUs. Communication patterns include broadcast, scatter/gather, reduce, and all-to-all. Small transfers between the nodes are latency sensitive while larger transfers are bandwidth sensitive. Picking the right interconnect platforms and appropriate topology that supports the communication patterns is crucial to unleash scale-out performance. The right interconnect will provide:

- **Consistent Performance** – DL workloads are distributed in nature with frequent high bandwidth communications between local as well as remote compute elements. The interconnect should deliver consistent performance both for remote and local traffic even in the presence of congestion. Each element in the network should provide consistent performance and the overall topology should support consistent connectivity across endpoints.
- **High Cross-Sectional Bandwidth** – DL workloads frequently have simultaneous, many-to-many, high bandwidth communications between compute elements. The interconnect should be able to support high bandwidth flows concurrently without being a bandwidth bottleneck. Smart congestion management and link-level flow distribution by network elements can enhance cross-sectional bandwidth. In addition, the choice of topology also has a big impact on cross-sectional bandwidth.
- **Scale** – The interconnect should not only meet today's demands but also scale to meet future demands as needed without impacting performance. With a scalable interconnect, customers can purchase and deploy additional resources as demand grows.
- **Failure Resiliency** – End-point, link, or network element failures should only have a minimal impact on the overall infrastructure. The interconnect infrastructure should provide sufficient visibility into the traffic to troubleshoot and reduce mean time to recovery.

We analyzed several network topologies for Hyperscale AIRI, their trade-offs, as well as some of the existing deployments. Let's go over three of the topologies that looked promising to support the needs of Hyperscale AIRI, their associated attributes, and matching traffic patterns.

2-D Torus

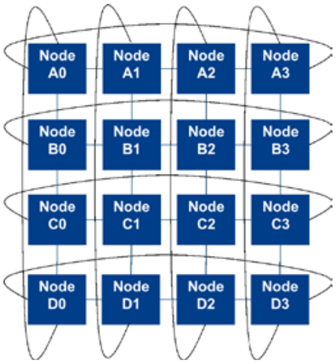


FIGURE 7. 2-D Torus topology

Some of the largest supercomputers in the world use torus topology to minimize node-to-node distance for message passing. This works well for certain applications where the data processing pipeline matches the physical topology. However, the same torus topology can cripple applications with many-to-many communication patterns due to the lack of cross-sectional bandwidth. Even a few node failures in a Torus network can impact performance.

3-Tier Hierarchical

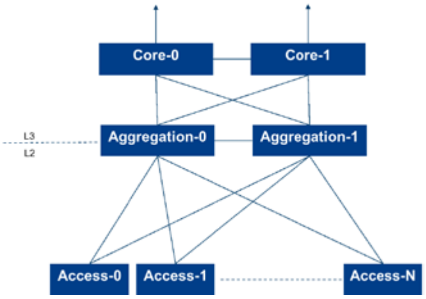


FIGURE 8. 3-Tier Hierarchical topology

During the early 2000s, when client-server applications were popular, data center networks were designed in a hierarchical fashion to deliver reliable north-south communication between the servers residing in the data centers and the clients outside. The access switches connect to the servers and a redundant pair of aggregation switches are used to interconnect the access switches. The same network design cannot scale-out to support modern web 2.0 workloads that are dominated by east-west traffic patterns, as the topology does not support more than two aggregation switches. The east-west bandwidth of the entire fabric is restricted to the bandwidth that can be supported by just the two aggregation switches.

Leaf/Spine Topology

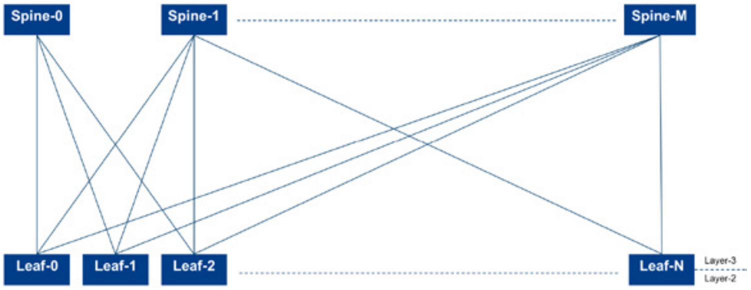


FIGURE 9. 2-Tier Leaf/Spine Topology

Layer-3 ECMP based Leaf/Spine topology has become the de facto network topology in modern data centers. The following are key design principles for a 2-Tier Leaf/Spine design:

1. All Leaf switches are connected to all Spines via uplinks. The maximum number of Spines is determined by the number of uplink ports in the Leaf switches.

2. Spine switches are not connected to each other.
3. Servers and endpoints can be single or dual homed but connect ONLY to Leaf switches.

Leaf/Spine topology satisfies all the requirements to build a scalable high-performance infrastructure: **consistent performance, high cross-sectional bandwidth, scale,** and **failure resiliency**.

CONSISTENT PERFORMANCE

Independent of the size of the network, every server that is connected to a Leaf/Spine network is just three network switches away from any other server. This inherent symmetry in the topology makes a consistent performance at scale possible.

In addition to the topology, the switches used to build the network also play a crucial role in delivering consistent performance. With a fully shared and monolithic packet buffer architecture, Mellanox Spectrum switches support a robust 100GbE data path with smart Explicit Congestion Notification (ECN)-based congestion management. Using Spectrum based 100GbE optimized leaf and spine switches will enable end-to-end cut-through performance.

HIGH CROSS-SECTIONAL BANDWIDTH

In a Leaf/Spine topology, every leaf switch is connected to every spine switch and there are multiple choices of interconnection path between the leaf switches. For example, Leaf-0 can communicate to Leaf-1 over M different paths where M is the number of spine switches. With Layer-3 Equal Cost Multipathing (ECMP), all these paths can be simultaneously activated. With intelligent hashing, traffic flows can be distributed across the paths maximizing link utilization and bandwidth.

Mellanox Spectrum switches implement Adaptive Routing (AR) to further enhance the cross-sectional bandwidth. AR facilitates the flows on your leaf/spine network, providing active route adjustments based on a global end-to-end view but with a distributed load-balancing technology (proven in IB) that works at ASIC speeds. The benefit is higher fabric utilization and cross-sectional bandwidth than static ECMP solutions.

SCALE

The number of ports on the spine switch is equal to the maximum number of racks the leaf/spine topology can support. For example, a design with Mellanox SN3800 spine switches with 64x100GbE ports can support up to 64 racks. The topology can initially contain just a couple of spine switches and can be expanded by adding more spine switches as needed.

FAILURE RESILIENCY

The interconnect should continue to function without having a substantial performance hit during link or even during switch failures. Leaf/Spine networks running time-tested Layer-3 routing protocols provide just that. When a link fails, the traffic that originally would have traversed that link is redistributed over other links. When a leaf switch fails, only the single homed servers connected to the leaf switch will lose connectivity. When a spine switch fails, only a fraction of the bandwidth related to the failed spine switch is affected.

While using a larger spine device enables a bigger network, it also increases the failure blast radius. It is ideal to have at least four or more spine switches in the network. In the event of a spine switch failure, only 25% of the bandwidth is lost.

TOPOLOGY	CONSISTENT PEFORMANCE	BANDWIDTH	SCALE	FAILURE RESILIENCY	APPLICATIONS
TORUS			✓		Niche High-Performance Applications
HIERARCHICAL	✓	✓			Client-Server Applications
LEAF/SPINE	✓	✓	✓	✓	DL and Modern distribution and web-scale applications

TABLE 1. Topology attributes

Rack Level Topology

The mix of storage resources, GPU, and CPU resources will vary according to the nature of the AI workload. To address this, we are proposing the following rack configurations as part of the reference architecture:

1. DGX-1 server based Model Intensive Rack
2. DGX-1 server based Data Intensive Rack Scale solution comprised of a mix of
    a. DGX-1 server based compute racks
    b. Pure Storage Flash Blade based storage racks
3. DGX-2 server based Model Intensive Rack
4. DGX-2 server based Model and Data Intensive Rack

The rest of this subsection will cover the connectivity for these different rack configurations

DGX-1 SERVER BASED RACK CONNECTIVITY

Model Intensive Rack Connectivity

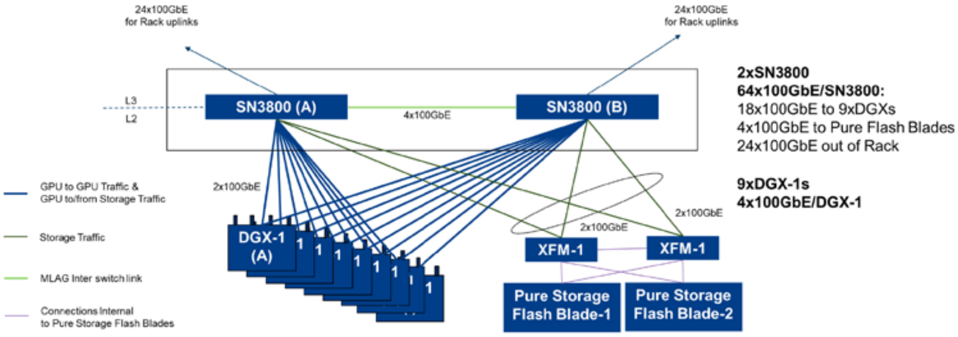


FIGURE 10. DGX-1 server based Model Intensive Rack

The **DGX-1 server based Model Intensive Rack** comprises nine DGX-1 server and two Pure Storage FlashBlade systems interconnected by two Mellanox SN3800 64x100GbE switches. The Mellanox SN3800 Ethernet switches support robust high bandwidth 100GbE GPU to GPU and GPU to Storage communications. RoCE transport with ECN and PFC is enabled for GPU to GPU communications. The storage traffic runs over TCP.

The Pure Storage FlashBlade arrays connect to the two SN3800 switches in MLAG over a redundant pair of 4x100GbE interfaces. The 4x100GbE ISL connection between the switches will carry storage traffic in the event of storage of host link failure.

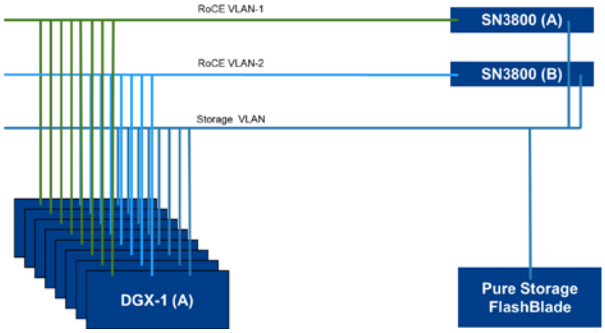


FIGURE 11. DGX-1 server based Model Intensive Rack VLAN connectivity

Two VLANs with unique IP ranges, RoCE VLAN-1 and RoCE VLAN-2, are provisioned to support RoCE traffic. A third VLAN with unique IP ranges, Storage VLAN, is provisioned for storage traffic. DGX server ports connected to SN3800 (A) belong to RoCE VLAN-1 and DGX server ports connected to SN3800 (B) belong to RoCE VLAN-2. SN3800 (A) acts as the default gateway for RoCE VLAN-1 and SN3800 (B) as the default gateway for RoCE VLAN-2. Traffic between the two RoCE VLANs is routed via the spine switch. Port-channels and switch MLAGs are configured to provide redundant storage connectivity.

Data Intensive Rack Connectivity

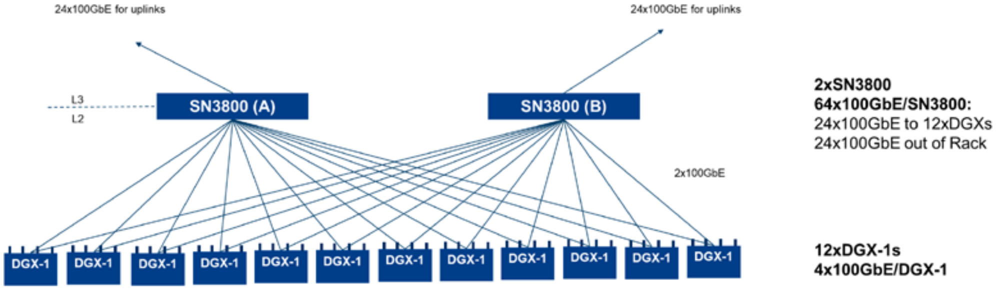


FIGURE 12. Compute Rack

The **DGX-1 server based Compute Rack** comprises 12 DGX-1 servers interconnected by a couple of Mellanox SN3800 64x100GbE switches. The Mellanox SN3800 Ethernet switches support robust high bandwidth 100GbE GPU to GPU communications. RoCE transport with ECN and PFC is enabled for GPU to GPU communications.

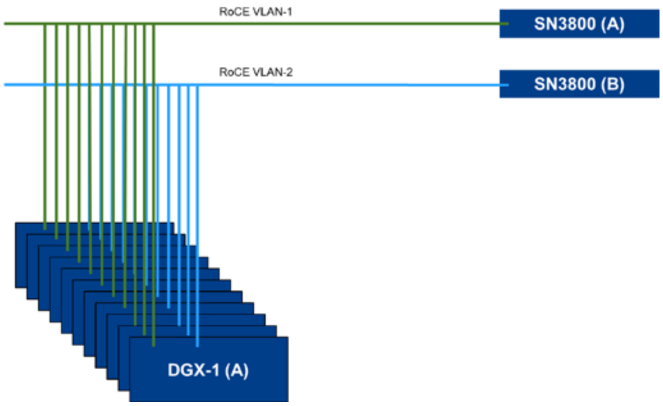


FIGURE 13. Compute Rack VLAN connectivity

Two VLANs with unique IP ranges, RoCE VLAN-1 and RoCE VLAN-2, are provisioned to support RoCE traffic. DGX server ports connected to SN3800 (A) belong to RoCE VLAN-1 and DGX server ports connected to SN3800 (B) belong to RoCE VLAN-2. SN3800 (A) acts as the default gateway for RoCE VLAN-1 and SN3800 (B) as the default gateway for RoCE VLAN-2. Traffic between the two RoCE VLANs is routed via the spine switch.

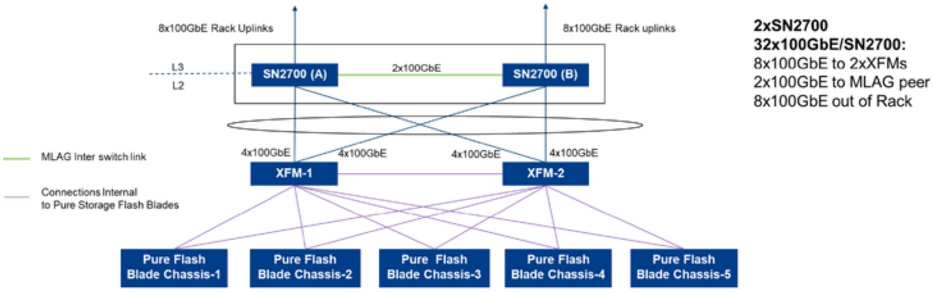


FIGURE 14. Storage Rack

The **Storage Rack** comprises of five FlashBlades with multi-chassis and two Mellanox SN2700 32x100GbE switches in MLAG to provide redundant external connectivity.

DGX-2 Server Based Rack Connectivity

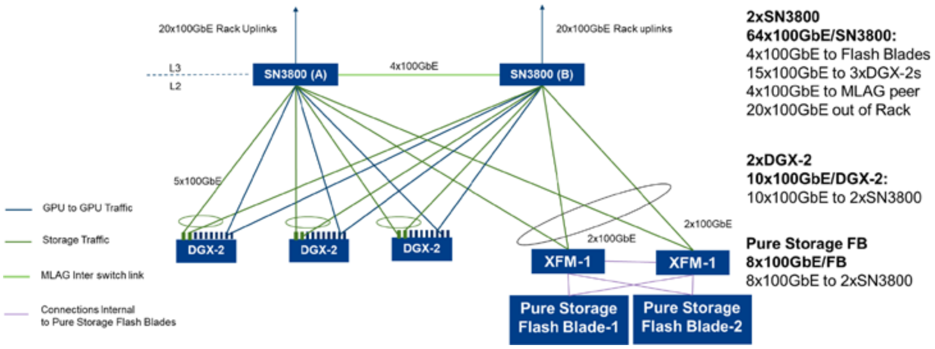
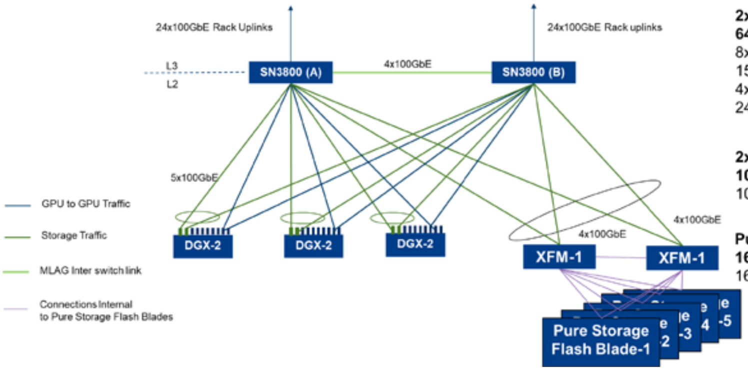


FIGURE 15. DGX-2 server based Model Intensive Rack topology



2xSN3800
64x100GbE/SN3800:
8x100GbE to Flash Blades
15x100GbE to 3xDGX-2s
4x100GbE to MLAG peer
24x100GbE out of Rack

2xDGX-2
10x100GbE/DGX-2:
10x100GbE to 2xSN3800

Pure Storage FB
16x100GbE/FB
16x100GbE to 2xSN3800

FIGURE 16. DGX-2 server based Model and Data Intensive Rack topology

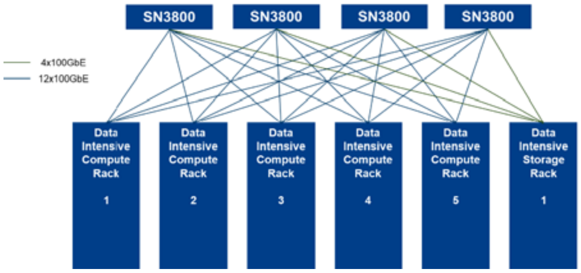
The **DGX-2 Server Based Racks** comprise three DGX-2 servers and a Pure Storage FlashBlade system interconnected by two Mellanox SN3800 64x100GbE switches. The pair SN3800 in MLAG provide redundant connectivity to the Pure Storage FlashBlades.

Two VLANs with unique IP ranges, RoCE VLAN-1 and RoCE VLAN-2, are provisioned to support RoCE traffic. DGX server ports connected to SN3800 (A) belong to RoCE VLAN-1 and DGX server ports connected to SN3800 (B) belong to RoCE VLAN-2. SN3800 (A) acts as the default gateway for RoCE VLAN-1 and SN3800 (B) as the default gateway for RoCE VLAN-2. Traffic between the two RoCE VLANs is routed via the spine switch.

DGX-2 servers are equipped with a separate PCIe slot to host a 2x100GbE adapter dedicated for storage traffic. The 2x 100GbE ports from this dedicated adapter can redundantly be connected to the pair of SN3800 switches over MLAG.

Multi-Rack Leaf/Spine Topology

A typical deployment will include some combination of several Compute and Storage Racks. **With Mellanox SN3800 as the spine switch, we can build a scale-out system comprising up to 64 racks.**


SN3800
SN3800
SN3800
SN3800
4x100GbE
12x100GbE
Data Intensive Compute Rack 1
Data Intensive Compute Rack 2
Data Intensive Compute Rack 3
Data Intensive Compute Rack 4
Data Intensive Compute Rack 5
Data Intensive Storage Rack 1

4xSN3800
64x100GbE/SN3800:
60x100GbE to 5xCompute Racks
4x100GbE to 1xData Rack

Data Intensive Compute Rack
48x100GbE Uplinks :
12x100GbE to each of the 4 SN3800 Spines

Data Intensive Storage Rack
16x100GbE Uplinks :
4x100GbE to each of the 4 SN3800 Spines

FIGURE 17. Multi-rack topology with five Compute Racks and a Storage Rack

The above shows a leaf/spine topology with five Compute and one Storage Rack. Each Compute Rack has 12x100GbE ports connecting to each of the 4xSN3800 spine switches. The storage rack has 4x100GbE ports connecting to each of the 4xSN3800 spine switches. Each compute rack has two subnets and storage rack has one subnet. Traffic across subnets is routed through the spine.

POWER AND COOLING MANAGEMENT

To some extent, deploying multiple racks of DGX-1 or DGX-2 server with FlashBlade arrays in an AIRI configuration is like deploying traditional servers and networking in a rack. However, with the high power consumption and corresponding cooling needs of DGX servers, server weight, and multiple networking cables per servers, additional care and preparation is needed for a successful deployment.

While additional DGX server and FlashBlade site requirements are detailed in their respective Site Preparation Guides, here are some important items for consideration to support both AIRI with three DGX-2 servers or nine DGX-1 servers in a rack.

DESIGN GUIDELINES	
RACK	SUPPORT 3000 LBS. OF STATIC LOAD DIMENSION OF 1200 MM DEPTH X 700 MM WIDTH STRUCTURED CABLING PATHWAYS PER TIA 942 STANDARD
COOLING	REMOVAL OF 119,420 BTU/HR ASHRAE TC 9.9.2015 THERMAL GUIDELINES “ALLOWABLE RANGE”
POWER	NA: A/B POWER FEEDS, EACH THREE-PHASE 400V/60A/33.2KW; (OR THREE-PHASE 208V/60A/17.3 KW WITH ADDITIONAL CONSIDERATIONS FOR REDUNDANCY AS NEEDED) INTERNATIONAL: A/B POWER FEEDS, EACH 300/400/415V, 32A, THREE PHASE 21-23KW EACH

TABLE 2. Source: DGX Server Ready Datacenter Guidelines

CONCLUSION

Artificial intelligence, fueled by rapid innovation in deep learning ecosystems, is becoming prevalent in a wide range of industries. Experts now believe new industry leaders will arise, led by enterprises who invest in AI and turn their data into intelligence. While many enterprises want to jumpstart their AI initiatives, challenges in building a scalable and AI-optimized infrastructure often hold them back. Engineers at NVIDIA, Mellanox, and Pure Storage partnered to architect a scalable and powerful infrastructure for enterprises pushing the boundaries of AI innovation. With AIRI’s modularity and simplicity, enterprises can start with AIRI “Mini” and easily hyperscale as their teams and projects start to grow. In addition, AIRI aims to solve infrastructure complexities, providing pioneering enterprises of AI a solution with supercomputing capabilities. Now, every enterprise can focus on developing powerful new insights with AI by deploying the simple, scalable, and robust AIRI system.

APPENDIX

For more information on AIRI Reference Architecture, see:

https://github.com/PureStorage-OpenConnect/AIRI

For more information about NVIDIA DGX, see:

https://www.nvidia.com/en-us/data-center/dgx-systems/

For more information about Pure Storage FlashBlade, see:

https://www.purestorage.com/solutions/applications/artificial-intelligence.html

For more information on Mellanox Spectrum switches, see:

https://www.mellanox.com/page/ethernet_switch_overview

For more information on NVIDIA GPU Cloud (NGC), see:

https://ngc.nvidia.com

For more information on NVIDIA Kubernetes Deepops, see:

https://github.com/NVIDIA/deepops

The Pure Storage products described in this documentation are distributed under a license agreement restricting the use, copying, distribution, and decompilation/reverse engineering of the products. The Pure Storage products described in this documentation may only be used in accordance with the terms of the license agreement. No part of this documentation may be reproduced in any form by any means without prior written authorization from Pure Storage, Inc. and its licensors, if any. Pure Storage may make improvements and/or changes in the Pure Storage products and/or the programs described in this documentation at any time without notice.

THIS DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID. PURE STORAGE SHALL NOT BE LIABLE FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH THE FURNISHING, PERFORMANCE, OR USE OF THIS DOCUMENTATION. THE INFORMATION CONTAINED IN THIS DOCUMENTATION IS SUBJECT TO CHANGE WITHOUT NOTICE.

ps_wp18p_hyperscale-airi_01