Palantir NVIDIA

March 2026

Palantir Sovereign AI Operating System Reference Architecture

The Palantir-NVIDIA AI Operating System Reference Architecture (AIOS-RA) is a subset of NVIDIA's Enterprise Reference Architecture, tested and qualified to run Palantir's complete software suite—including AIP, Foundry, Apollo, Rubix, and AIP Hub.

This version of the Reference Architecture delivers a complete, integrated AI infrastructure combining NVIDIA HGX B300 8-GPU nodes with Palantir's hardened compute infrastructure, unified by Apollo deployment automation and Rubix Kubernetes substrate. The architecture utilizes NVIDIA Spectrum-X Ethernet networking and is optimized for on-premises data center deployments.

Notice & Disclaimers

NVIDIA Enterprise Reference Architecture documentation is NVIDIA confidential information and embodies substantial intellectual property and engineering know-how. NVIDIA grants you limited permissions to use NVIDIA Enterprise Reference Architecture to create NVIDIA platform-enabled data center clusters with NVIDIA technologies. You may not use the NVIDIA Enterprise Reference Architecture to develop competing products or technologies or assisting a third party in such activities. Except as expressly stated in this Notice & Disclaimer statement, no other license or right is granted to you by implication, estoppel or otherwise. NVIDIA objects to and rejects any additional or different terms you may propose.

NVIDIA Enterprise Reference Architecture is subject to change without notice. NVIDIA technologies may require enabled hardware, software, or service activation. Your costs and results may vary. This document is not a commitment to develop, release, or deliver any technology, code, or functionality. NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice. It is your sole responsibility to evaluate and determine the applicability of any information contained in this document and ensure that the product is suitable and fit for the application planned by you. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party or a license from NVIDIA.

NVIDIA products or services are subject to the applicable NVIDIA product or sales terms and conditions that are provided in conjunction with the relevant NVIDIA product or service. NVIDIA's provision of these resources does not expand or otherwise alter NVIDIA's applicable warranties or warranty disclaimers for the NVIDIA products or services.

You will not use the NVIDIA Enterprise Reference Architecture or NVIDIA's confidential information to: (i) identify or support an assertion or potential assertion of any intellectual property rights (including patent, copyright, or trade secret); or (ii) prepare, file, amend, or prosecute any patent applications. You agree to grant NVIDIA a limited, worldwide, nonexclusive, perpetual, irrevocable, non-sublicensable, nontransferable, royalty-free, fully-paid license to any patent claim that you may file after accessing NVIDIA Enterprise Reference Architecture that covers the subject matter disclosed herein.

Warranty Disclaimer. NVIDIA Enterprise Reference Architecture materials disclosed are provided "AS IS". To the maximum extent permitted by applicable law, NVIDIA disclaims all warranties and representations of any kind, whether express, implied, or statutory, relating to or arising under this agreement, including, without limitation, the warranties of title, noninfringement, merchantability, fitness for a particular purpose, usage of trade, and course of dealing.

Limitation of Liability. To the maximum extent permitted by applicable law, in no event will NVIDIA be liable for any (i) indirect, punitive, special, incidental, or consequential damages, or (ii) damages for the (a) cost of procuring substitute goods or (b) loss of profits, revenues, use, data, or goodwill arising out of or related to this document, whether based on breach of contract, tort (including negligence), strict liability, or otherwise, and even if NVIDIA has been advised of the possibility of such damages and even if a your remedies fail their essential purpose. Additionally, to the maximum extent permitted by applicable law, NVIDIA's total cumulative aggregate liability for any and all liabilities, obligations, or claims arising out of or related to this document will not exceed one-hundred U.S. dollars (US$100).

Abstract

Together, Palantir and NVIDIA deliver a robust, fully integrated AI operating system, optimized for NVIDIA accelerated compute and Palantir's suite of software capabilities. This architecture accelerates deployment and maximizes performance across on premises, edge, and sovereign cloud environments. NVIDIA and Palantir have qualified the server, network and software specifications in this document via lab testing.

The Palantir AIOS-RA architecture is a subset of the NVIDIA Enterprise Reference Architecture (Enterprise RA) utilizing the 2-8-9-400 node architecture. Palantir AIOS-RA deployment requirements and software integration further constrain the server and network configuration to optimize performance and security for both the control plane of Rubix and Apollo as well as the capabilities suites across Foundry and AIP.

The Enterprise RA is designed to solve the most challenging computational problems of today. Certified configurations based on this architecture can be deployed in enterprise data centers around the world. This Enterprise RA uses the following technologies:

NVIDIA HGX B300 Servers each with eight NVLink connected GPUs

NVIDIA Spectrum-X Platform

NVIDIA ConnectX®-8 SuperNICs

Palantir AI Operating System

This document discusses the components that define the scalable and modular architecture. The configuration is built on a base configuration or partner servers and then can scale based on a predefined configuration of 4 partner servers called scalable units. This provides for rapid deployment of systems of multiple sizes. This RA includes details regarding the SU design and specifics of Ethernet fabric topologies and NVLink networking, and storage system specifications.

The AIOS-RA is managed Palantir Apollo and accelerated by NVIDIA platform software including NVIDIA AI Enterprise, NVIDIA® CUDA®, and NVIDIA Magnum IO™. These technologies help keep the system running at the highest levels of availability. AIOS-RA support is delivered through partnerships with Palantir for Software Support, NVIDIA Enterprise Support (NVEX), and Enterprise RA solution partners, ensuring all components and applications operate smoothly.

Contents

Overview of the Palantir AI Operating System Reference Architecture (AIOS RA)

Overview

The Palantir | NVIDIA AIOS RA delivers a complete, production-ready AI infrastructure that combines:

NVIDIA GPU Infrastructure

↳ HGX B300 8-GPU servers with Spectrum-X Ethernet networking for AI training and inference

Palantir Compute Infrastructure

↳ Hardened Kubernetes substrate running Foundry services (Catalog, Alta, Multipass, etc.)

Unified Management Plane

↳ Rubix (zero-trust Kubernetes) + Apollo (autonomous deployment & lifecycle management)

AIP Platform

↳ Enterprise AI platform connecting LLMs to organizational data and operational systems

NVIDIA acceleration through NVIDIA AI Enterprise Accelerated Computing Libraries, NVIDIA® CUDA®, and NVIDIA Magnum IO

This integrated architecture is sold and deployed as a complete solution, providing customers with a turnkey AI datacenter spanning hardware procurement through application deployment.

Architecture Layers

This architecture is organized into four layers—Application, Platform, Runtime, and Hardware—each providing distinct capabilities and benefits that build on one another for secure, scalable AI deployment and operations.

- **Application Layer (Palantir AIP + NVAIE)**

Palantir's NVIDIA-accelerated app suite delivers application performance optimization, autonomy, NLP, logistics, COA generation, and data analytics as turnkey operational apps that compress time to value and enhance model-driven workflows across edge and enterprise environments. These apps harness GPU acceleration to speed planning and optimization while enabling autonomous and analyst workflows that directly support mission outcomes and decision advantage.

Key capabilities include:

- LLM integration with enterprise data and logic
- Production-grade AI agents and workflows
- Operational decision support applications
- Optimized for NVIDIA GPU acceleration

Architecture Layers Cont.

- **Platform Layer** (**Palantir Foundry**)

The Palantir AIP platform unifies dynamic ontology, data management, model orchestration, open APIs, and builder tools into a secure workspace that integrates and synchronizes data while enforcing governance. By combining flexible data integration with in-platform tooling, it lets teams build, orchestrate, and scale AI workflows across systems quickly, keeping sensitive information protected without sacrificing development speed. The Palantir platform supports integration with a broad array of storage providers, including existing data sources hosted on PowerScale and ObjectScale.

Key capabilities include:

- Dynamic ontology and data management
- Model orchestration and governance
- Open APIs and builder tools
- Secure workspace with enforced data controls

- **Runtime Layer** (**Palantir Rubix & Apollo**)

Palantir Rubix & Apollo provide a zero-trust, containerized runtime for security, compliance, monitoring, delivery, and storage, enabling "build once, run anywhere" deployment with out-of-the-box ATO and day-2 continuous delivery. This operational layer hardens AI services, simplifies multi-environment rollout, and sustains updates and observability for reliable multi-tenant operations at scale.

Key capabilities include:

- Zero-trust, containerized Kubernetes runtime
- Multi-tenant security with namespace isolation
- "Build once, run anywhere" deployment model
- Out-of-the-box ATO compliance support
- Continuous delivery with automated rollback
- Hub-and-spoke fleet management

- **Hardware Layer** (**NVIDIA Certified Systems**)

The AIOS Hardware layer features three server types: two CPU-based node specifications for the control plane and Platform layer, and one GPU server specification in scalable units of four nodes. All nodes are interconnected using NVIDIA Spectrum-X Ethernet networking. The AIOS should be built in pre-designated and tested scalable units. For convivence NVIDIA and Palantir have qualified 3 distinct sizes for Small, Medium or Large Enterprises. The platform nodes have a fixed count and configuration at each of the 3 sizes. The GPU count can scale for each configuration based on demand. A demand calculator has been provided:

Architecture Layers Cont.

Table 1

Hardware Arch Node Scaling

Size	Platform Nodes	GPU Count (SU Count)	Application
Small	3 Master + 9 Worker	32 – 128 (1-4 SU)	Departmental AI pilots, specialized workloads, tactical deployments
Medium	3 Master + 21 Worker	128 – 320 (4-10 SU)	Enterprise-wide AI, multi-tenant production, mission applications
Large	3 Master + 45 Worker	320 - 640 (10-20 SU)	Mission-critical AI at scale, organization-wide deployment, multi-workload

- **Platform Nodes**

These systems are CPU nodes for running Palantir Management, Provisioning, Monitoring and Control Software from Palantir.

Platform Nodes based on the Master Node Specification and Worker Nodes specifications below must be used to build and connect this architecture.

↳ AIOS requires dedicated CPU compute infrastructure to run Foundry services (Catalog, Alta, Multipass, etc.), Kubernetes control plane, and orchestration layers.

↳ This infrastructure is co-located with GPU nodes and managed via the same Apollo deployment system.

Hardware support is available through the fulfilment OEM. Software support from NVIDIA and Palantir through respective licensing agreements.

Architecture Layers Cont.

○ **GPU Nodes**

A best-of-breed GPU stack from NVIDIA—spanning air-cooled and liquid-cooled B200 and B300—delivered via an appliance model with OEM partners provides future-proofed architecture and managed supply chain. The options align thermals and procurement to workload density, preserving performance headroom and lifecycle longevity while easing deployment logistics.

The AIOS RA is a modular architecture based on NVIDIA-Certified HGX B300 systems, each with eight B300 SXM GPUs. Using a four-node scalable unit (SU), this RA scales up to 128 NVIDIA-Certified systems for a total of 1024 B300 SXM GPUs.

↳ This reference design scales up to 20 SUs (Scalable Units). Larger clusters can be built based on customer requirements.

↳ Flexible rail-optimized end-of-row network architecture that can accommodate modifications in the rack layout and number of servers per rack

● **Network Architecture**

The architecture uses NVIDIA Spectrum-X Ethernet networking throughout:

GPU-to-GPU Communication (East-West):

- NVIDIA ConnectX-8 SuperNICs (8x per GPU node)
- 800GbE per adapter
- RDMA over Converged Ethernet (RoCE)
- Rail-optimized spine-leaf topology

Storage, Customer and Management (North-South):

- NVIDIA ConnectX-8 SuperNIC (2x per GPU node 2-port Ethernet configuration)

Platform Node Networking (East-West):

- NVIDIA ConnectX-8 SuperNIC (2-port Ethernet configuration)
- 2 ports LACP: OpenShift internal communications

Platform Node Networking (North-South):

- NVIDIA ConnectX-8 SuperNIC (2-port Ethernet configuration)
- 2 ports LACP: Upstream/customer connectivity

Management Network (Out-of-Band):

- 1GbE connections for BMC/iDRAC access
- Physical isolation from production traffic

∗ **Critical**

This architecture is standardized on Ethernet networking only. InfiniBand is not supported due to Palantir Rubix/Apollo requirements for IP-based networking and multi-tenant isolation. DPU is currently not supported in this generation of AIOS-RA, but is planned for future support pending qualifications timelines.

Architecture Layers Cont.

- **Deployment Model**

Customers purchase and deploy the complete integrated stack:

1. **Hardware Procurement**
   ↳ NVIDIA GPU nodes + Palantir compute nodes + networking
2. **Apollo Bootstrap**
   ↳ Cluster connected as Apollo "Environment" (Kubernetes cluster)
3. **Rubix Deployment**
   ↳ Hardened substrate installed via Apollo spoke control plane
4. **Foundry Services**
   ↳ Catalog, Alta, Multipass, etc. deployed to worker nodes
5. **AIP Activation**
   ↳ Platform layer activated with customer data integration
6. **Model Deployment**
   ↳ LLMs and AI workflows deployed to GPU infrastructure

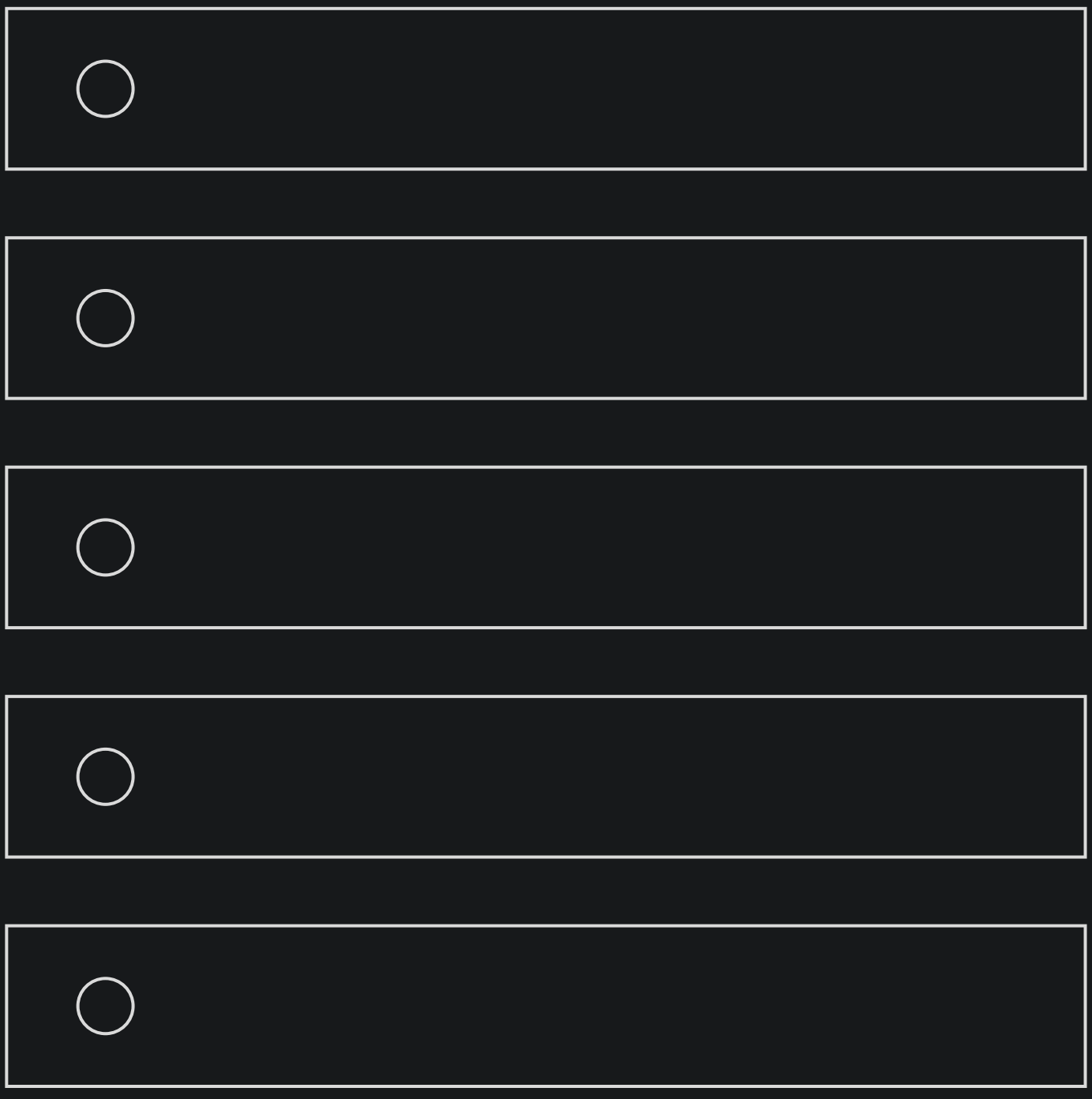
The reference architecture is designed for air-gapped and on-premises deployments with full autonomy after initial installation.

Table 2

Components and Layers of AIOS

Component	Details
East/West Network	Required, Single Plane
North/South Network	Single Tenancy
Storage Network	Required
Provisioning Stack	Apollo Spoke Control Plane
Runtime Stack	Rubix
Platform Layer	Foundry
Application Stack	AIP + NVIDIA NVAIE

Node Hardware



Hardware Components

This section describes the hardware components of the AIOS, including the scale out GPU Nodes as well as the Master and Worker Platform Node specification. Tables 4-6 List Node configurations from Dell Technologies that of the same generation, that are interoperable.

It is recommended to source Master, Worker and GPU nodes from the same manufacturer, who should also be able to pre-rack and cable the servers and networking equipment.

Dell PowerEdge R670

The Dell PowerEdge R670 is a compact, high-performance rack server well suited for running Palantir Ontology workloads that rely on CEPH for Ontology objects. Its balanced combination of compute density, memory bandwidth, and I/O flexibility aligns with the performance needs of Ontology object hydration, transform pipelines, and large-scale analytical operations.

With DDR5 memory across up to 32 DIMM slots, the R670 provides the high capacity and bandwidth required to efficiently load, process, and compute on complex Ontology objects, especially in environments where rapid in-memory access and parallel workflows are critical.

For storage-intensive Ontology workloads, the R670 supports NVMe, SAS, and SATA configurations, including PCIe Gen5 NVMe drives that enable the low latency and high throughput needed for CEPH object reads/writes. This allows CEPH OSDs or caching tiers to deliver consistent performance for Ontology-backed pipelines and fast-growing datasets.

The platform includes multiple PCIe Gen5 expansion slots, enabling integration of high-speed networking, storage controllers, or accelerators as Ontology and CEPH requirements scale. Support for OCP 3.0 NICs provides additional flexibility for deploying high-bandwidth Ethernet or InfiniBand, which is essential for CEPH cluster replication, durability, and recovery operations.

Designed for efficient large-scale operation, the R670's optimized air cooling and power delivery in a 1U chassis make it ideal for dense rack environments typical of Palantir Foundry deployments. Management through iDRAC with OpenManage enables secure remote lifecycle management, automation, and fleet-level operational simplicity. Security features—including Silicon Root of Trust, Secure Boot, System Lockdown, and Data-at-Rest Encryption with SEDs—help protect Ontology objects stored in CEPH throughout the server's lifecycle, ensuring consistent integrity and trust for mission-critical data. Overall, the PowerEdge R670 offers the compute, memory, and storage profile needed to support scalable, secure, high-performance Ontology workloads backed by CEPH object storage.

Foundry Platform Systems

Table 3

Platfrom Master Node Specification (Kubernetes Control Plane)

Parameter	Requirement
Quantity	3 nodes (high availability)
Purpose	Kubernetes control plane operations
Manufacturer	Dell Technologies
Server Model	Dell PowerEdge R670 (17th Gen)
CPU	Minimum: 2x Intel® Xeon® 6 Performance 6530P 2.3G, 32C/64T, 24GT/s, 144M Cache, Turbo, (225W) DDR5-6400
Memory	16x 64GB RDIMM, 6400MT/s 2Rx4 (1TB total)
Boot Controller	BOSS-N1 monolithic controller card with 2x M.2 960GB
Direct Attach Storage	2x 3.2TB Data Center NVMe Mixed Use AG Drive E3s Gen5 with carrier 10x 7.68TB Enterprise NVMe Read Intensive AG SED Drive E3s Gen5 with carrier, FIPS
Network Cards	2x NVIDIA ConnectX-8 Dual Port Ethernet Host Bus Adapter
Network Configuration	4-port Ethernet (see Network Requirements below)
Power Supplies	Redundant (1+1) 1500W PSU
Rails	ReadyRails Sliding, no cable management
iDRAC Version	iDRAC10, Enterprise 17G

Foundry Platform Systems Cont.

Table 4

Platform Worker Node Specification (Foundry Services)

Parameter	Requirement
Quantity	N nodes (scaled based on storage requirements - see Sizing Guidance)
Purpose	Foundry services (Alta, Multipass, Catalog, streaming, etc.)
Manufacturer	Dell Technologies
Server Model	Dell PowerEdge R670 (17th Gen)
CPU	Minimum: 2x Intel® Xeon® 6 Performance 6530P 2.3G, 32C/64T, 24GT/s, 144M Cache, Turbo, (225W) DDR5-6400
Memory	16x 64GB RDIMM, 6400MT/s 2Rx4 (1TB total)
Boot Controller	BOSS-N1 monolithic controller card with 2x M.2 960GB
Direct Attach Storage	2x 3.2TB Data Center NVMe Mixed Use AG Drive E3s Gen5 with carrier 10x 7.68TB Enterprise NVMe Read Intensive AG SED Drive E3s Gen5 with carrier, FIPS
Network Cards	2x NVIDIA ConnectX-8 Dual Port Ethernet Host Bus Adapter
Network Configuration	4-port Ethernet (see Network Requirements below)
Power Supplies	Redundant (1+1) 1500W PSU
Rails	ReadyRails Sliding, no cable management
iDRAC Version	iDRAC10, Enterprise 17G

NVIDIA HGX B300 Baseboard

The HGX B300 baseboard (Figure 1) is an AI powerhouse that enables enterprises to expand the frontiers of business innovation and optimization. It is capable of training AI models scaling up to 10 trillion parameters. HGX B300 is a unified platform consisting of 8 Blackwell GPUs internally connected by NVLink and externally connected by a network interface of 800 Gb/s (2 x 400Gb/s Ethernet) per each GPU. This connectivity is provided by 8x ConnectX-8 SuperNICs on the baseboard for the E/W (East-West) networking.

Figure 1

HGX B300 Baseboard

Table 5

HGX B200 or B300 GPU and per Node Specifications

Specification	NVIDIA B200 SXM	NVIDIA B300 SXM
Memory per GPU	180GB HBM3e	288GB HBM3e
Memory per 8-GPU Node	1.44TB HBM3e	2.30TB HBM3e
GPU Bandwidth	Up to 8TB/s	Up to 8TB/s
GPU Aggregate Bandwidth per 8-GPU Node	Up to 64TB/s	Up to 64 TB/s

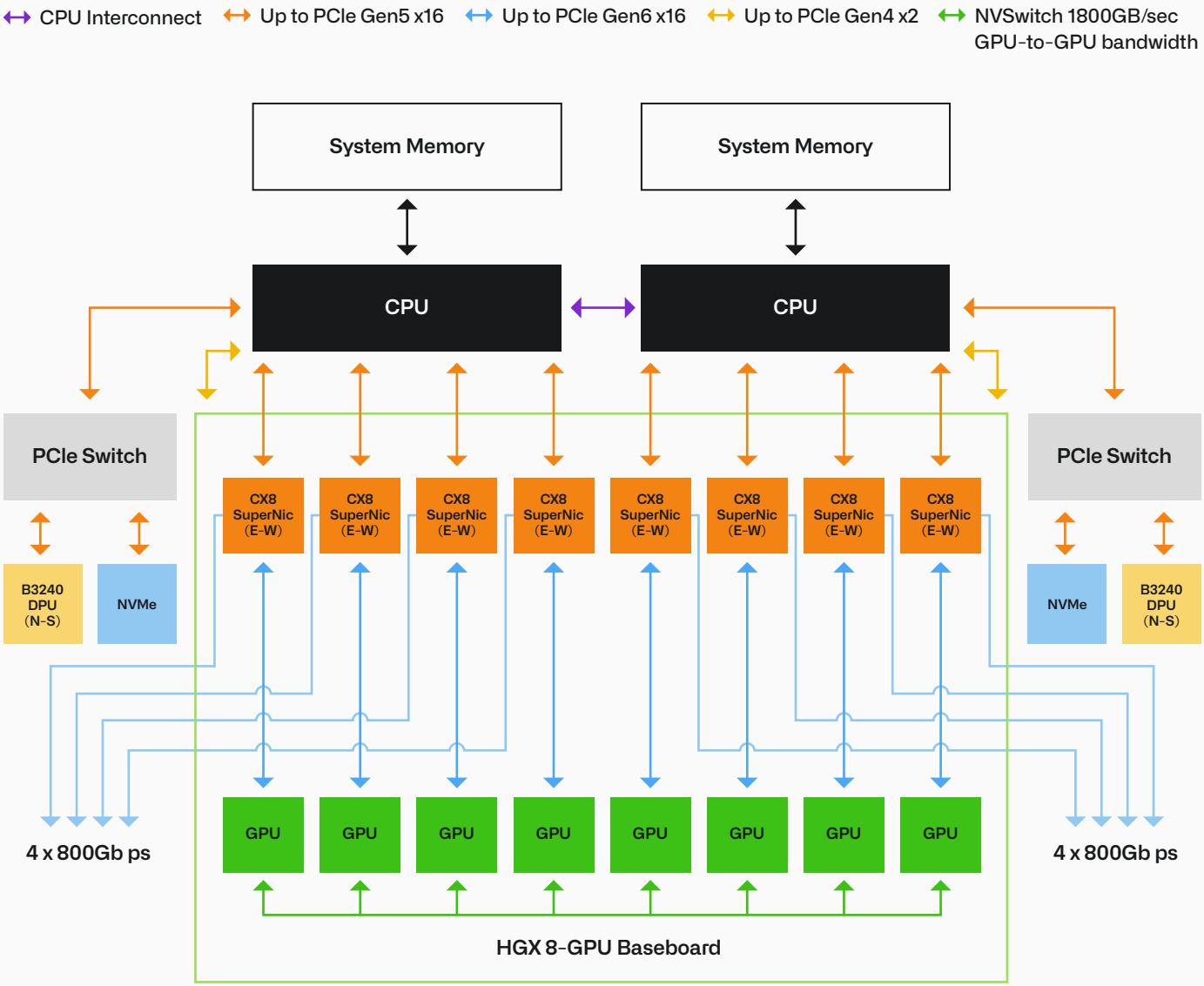
The HGX B300 baseboard provides 144 petaflops of processing power, making it a leading accelerated platform for AI and HPC. It offers advanced networking up to 800 Gbps with ConnectX-8 SuperNICs, enabling AI cloud networking, composable storage, zero-trust security, and GPU compute elasticity in AI designs.

HGX B300 Systems

NVIDIA-Certified HGX B300 systems are based on a common system design with flexibility for optimizing the configuration to match the cluster requirements. An example of this design is shown in Figure 2.

Figure 2

Example of a HGX B300 8 GPU System Configuration



GPU node specification is shown in the table below. This version of the AIOPS RA assumes a Dell Technologies configuration.

HGX B300 Systems Cont.

Figure 3

Dell PowerEdge XE9780

The Dell PowerEdge XE9780 is a high-performance, GPU-optimized platform ideal for powering Palantir AIP workloads and large-scale ontology-driven AI pipelines. With two 6th Gen Intel® Xeon® processors (up to 86 cores each) and configurations supporting up to eight NVIDIA HGX B300 NVL8 or B200 GPUs connected via NVIDIA NVLink, the XE9780 delivers the compute density needed for Palantir AIP model customization, RAG pipelines, and real-time agentic workflows grounded in the Palantir Ontology.

Up to 4 TB of DDR5 memory enables high-throughput ontology queries and simultaneous AIP model operations, while flexible storage—up to 16 E3.S Gen5 NVMe or 10 U.2 Gen5 NVMe SSDs—supports low-latency access to vector stores, feature sets, and operational data streams. The platform's PCIe Gen5 expansion and OCP 3.0 Gen5 networking integrates easily into high-bandwidth AI fabrics, making it suitable for distributed AIP clusters where model-serving, orchestration, and ontology services run at scale.

With iDRAC10 for fleet-level management and hardware security features such as Silicon Root of Trust and Secure Boot, the XE9780 meets enterprise and government-grade requirements—making it a strong foundation for running Palantir AIP's end-to-end AI workflows reliably and securely.

HGX B300 Systems Cont.

Table 6

HGX B300 GPU Node Specification (8-GPU System Component)

Parameter	Requirement
Quantity	4 node scalable units. Order in multiples of 4.
Target Workloads	Palantir AIP, Large Language Model; Traditional DL Inference Models
Manufacturer	Dell Technologies
Server Model	PowerEdge XE9780 (Gen 17)
GPU Configuration	Eight B300 SXM GPUs on a B300 baseboard with up to 2304 GB of GPU memory (See the topology diagrams for details)
NVIDIA® NVLink™ and NVSwitch™	HGX B300 Baseboards use a combination of fifth generation NVSwitch and fifth generation NVLink ↳ Total Aggregate Bandwidth 14.4TB/s ↳ GPU-to-GPU Bandwidth 1800GB/s
gCPU	Intel Emerald Rapids, or Sapphire Rapids
CPU Sockets	Two CPU sockets minimum
CPU Speed	2.0 GHz minimum base CPU clock
CPU Cores	Minimum of 48 physical CPU cores per socket ↳ Recommendation of 56 physical CPU cores per socket
System Memory (Total Across All CPU sockets)	Minimum of 2TB system memory and 500GB/s memory bandwidth ↳ For optimal performance, system memory should be evenly distributed across all CPU sockets and memory channels and should be fully populated and symmetrically placed on all CPU memory controller (MC) channels.
DPU (North-South)	2x NVIDIA ConnectX-8 Dual Port Ethernet Host Bus Adapter
PCI Express	Eight Gen5 x16 links and one Gen4 x2 link per HGX B300 baseboard One Gen5 x16 link per DPU, SuperNIC or adapter
PCIe Topology	Balanced PCIe topology with connectivity spread evenly across CPU sockets and PCIe root ports.
Network Adapters/NICs Speed (East-West)	Eight NVIDIA® ConnectX-8 SuperNICs per HGX B300 baseboard ↳ Up to 800 Gbps per adapter
Local Storage	Local storage recommendations are as follows: ↳ Inference Servers: Minimum 1 TB NVMe drive per CPU socket ↳ Training / DL Servers: Minimum 2 TB NVMe drive per CPU socket ↳ HPC Servers: Minimum 1 TB NVMe drive per CPU socket ↳ 1 TB NVMe boot drive
Remote Systems Management	SMBPBI over SMBus (OOB) protocol to BMC PLDM T5-enabled SPDM-enabled
Security	TPM 2.0 module (secure boot)

Network Design

Configuration

The AIOS RA configuration use three physical network fabrics:

Converged (North-South) Network

Compute (East-West) Network

Out-of-Band Management Network

Each network is discussed in this section.

East-West Network

To deliver the highest AI performance, NVIDIA recommends the ConnectX-8 SuperNIC network adapters.

The on baseboard, ConnectX-8 SuperNIC offers up to 800 Gb/s, low-latency network connectivity between GPUs in the AI cluster, featuring RDMA and RoCE acceleration, with NVIDIA® GPUDirect® and GPUDirect Storage technologies. For data centers that deploy Ethernet, ConnectX-8 offers a range of advancements including RoCE optimizations and multi-tenancy, which are key to in AI environments. Additionally, ConnectX-8 offers advanced hardware offloads for overlay networks such as VXLAN and NVGRE, which help reduce CPU load and improve network efficiency.

Beyond raw speed, ConnectX-8 enhances network performance with features like advanced congestion control, telemetry-based routing, and quality of service (QoS) capabilities to meet diverse AI application and training requirements. Security is built in with hardware acceleration for cryptographic protocols like IPSec and MACSec, taking these tasks over from the CPU to maintain performance.

Multi-node deployments with an NVIDIA B300 platform should adhere to the following total compute network bandwidth per GPU recommendations.

The ConnectX-8 SuperNIC for the East-West Network is integrated onto the baseboard, maintaining a **1:1 GPU-to-NIC ratio** to ensure optimal bandwidth and direct connectivity for each GPU.

Total Compute Network Bandwidth

400 GB/s (8x 400 Gb/s Ethernet NICs)

Palantir Platform nodes also connect over this same fabric is a non-blocking topology. This network is used for both distributed storage and OpenShift/Kubernetes internal communication. (pod networking, service mesh). This design allows the most efficient communication for clustered applications within and across the nodes.

The leaf and spine architecture allows for a scalable and reliable network that can fit varied sizes of clusters using the same architecture.

East-West Network Cont.

The compute fabric is designed to maximize bandwidth and minimize network latency required to connect GPUs within a server and within a rail.

Many common pre-trained models do not exceed the size of a single B300 SXM GPU. Models beyond 120B parameters may require model parallelism and require more than one B300 SXM GPU. This parallelism will still reside within the same node as each server node can hold up to 8x B300 SXM GPUs. For foundation model inference hosting, as well as training and fine-tuning workloads multi-node scaling >8 GPUs will be required. This architecture supports these workloads, and sizing of GPU scalable units should be aligned to these uses cases.

North-South Network

A converged network for both storage and in-band management is used in the Enterprise RA. This provides Enterprises with flexible storage allocation and easy network management. The converged network has the following attributes:

- Provide high bandwidth to shared storage and connects the customer through the converged network.
- It is independent of the compute fabric to maximize both storage and application performance.
- Provides Palantir Foundry platform layer with ceph storage and connectivity to external service.
- Each compute and management node are connected with two 400 GbE ports to two separate switches to provide redundancy and high storage throughput that can reach up to 40 GB/s per node.

The fabric is built on Ethernet technology with RDMA over Converged Ethernet (RoCE) support and utilizes NVIDIA ConnectX-8 in each compute node to deliver existing and emerging cloud and storage services.

- It is flexible and can scale to meet specific capacity and bandwidth requirements.
- Tenant-controlled management nodes provide tenant the flexibility to deploy OS and their job scheduler of choice.
- Hybrid storage fabric design with support for tenant isolation provides access to both shared and dedicated storage per tenant.
- Used for node provisioning, data movement, Internet access, and other services that must be accessible by the users.

Out-of-Band Management (Node) Network

The OOB management network connects all the base management controller (BMC) ports, as well as other devices that should be physically isolated from system users to allow the infrastructure management. This includes the 1 GbE switch management ports.

NVIDIA NetQ™

NVIDIA® NetQ™ is a scalable, modern network operations tool set that provides visibility into your overlay and underlay networks, enabling troubleshooting in real-time. NetQ delivers data and statistics about the health of your data center—from the container, virtual machine, or host, all the way to the switch and port. NetQ correlates configuration and operational status, and tracks state changes while simplifying management for the entire Linux-based data center. With NetQ, network operations change from a manual, reactive, node-by-node approach to an automated, informed, and agile one. Visit Network Operations with NetQ to learn more.

Rack Elevations and Network Topology

Overview

The OOB management network connects all the base management controller (BMC) ports, as well as other devices that should be physically isolated from system users to allow the infrastructure management. This includes the 1 GbE switch management ports.

Table 7

Hardware Arch Node Scaling

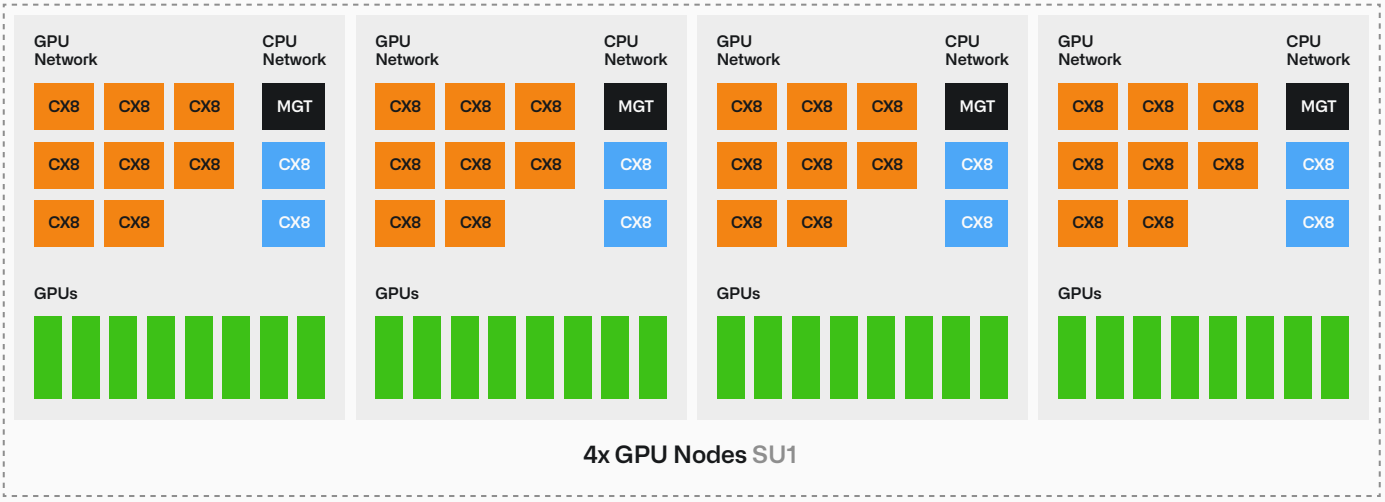
Size	Platform Nodes	GPU Count (SU Count)	Application
Small	3 Master + 9 Worker	32 – 128 (1-4 SU)	Departmental AI pilots, specialized workloads, tactical deployments
Medium	3 Master + 21 Worker	128 – 320 (4-10 SU)	Enterprise-wide AI, multi-tenant production, mission applications
Large	3 Master + 45 Worker	320 - 640 (10-20 SU)	Mission-critical AI at scale, organization-wide deployment, multi-workload

The Enterprise RA is built on scalable units (SU) based on 4 compute nodes.

Each SU is a discrete entity of computation that is tied to the port availability size of the network devices. SUs can be replicated to adjust the scale of the deployment with more ease.

Figure 4

Example Diagram of 4-Node SU



Rack Elevations and Network Topology Cont.

The 4-Node scalable unit provides the following connectivity building blocks:

- For the Compute (E-W) fabric: 4 servers, each with 8x ConnectX-8 SuperNICs, providing 32x 400Gb/s connections and a total aggregate bandwidth of 12.8Tb/s
- For the Converged (N-S) fabric: 4 servers, each with 1x ConnectX-8 providing 8x 200Gb/s connections and a total aggregate bandwidth of 1.6Tb/s
- For the Out-of-band Management fabric, 4 servers, each with 6x 1Gb/s connections providing 24x 1Gb/s for management

Spine-Leaf Networking

The network fabrics are built using switches with NVIDIA Spectrum-X Ethernet technology in a full nonblocking fat tree topology to provide the highest level of performance for the application running over the Enterprise RA configuration. The networks are RDMA compliant based fabrics in a leaf-spine architecture where the GPUs are connected using a rail-optimized network topology through their respective SuperNICs. This design allows the most efficient communication for multi-GPU applications within and across the nodes.

The leaf and spine architecture allows for a scalable and reliable network that can fit varied sizes of clusters using the same architecture. The compute network is designed to maximize bandwidth and minimize network latency required to connect GPUs within a server and within a rail.

In addition to the supported compute fabrics the converged spine-leaf network also supports the following attributes:

- Provides high bandwidth to high performance storage and connects to the data center network.
- Each compute and management node is connected with two 200 GbE ports to two separate switches to provide redundancy and high storage throughput.

Details of the converged network are found in NVIDIA Cloud Partner Reference Architecture: Common Networking (NVOnline 1116743).

Out-of-Band (OOB) Management Network

The OOB management network connects the following components:

- Base management controller (BMC) ports of the server nodes
- Server node 1G Base-T in-band management connections
- OOB management ports of switches

This OOB network can also connect other devices that should have management connectivity physically isolated for security purposes. The SN2201 is used to connect to the BMC/OOB 1 Gbps ports of these components. Scale out this network using a 25 Gbps or 100 Gbps spine layer to connect the SN2201 uplink switches.

Details of the OOB management network are found in NVIDIA Cloud Partner Reference Architecture: Common Networking (NVOnline 1116743).

AIOS RA "Small"

The rack elevation below shows the small AIOS RA architecture, featuring 3 master servers, 9 workers, and up to 4 scalable GPU scalable units for a total of 128 GPUs. Each GPU unit occupies one rack. Networking is centralized in a single rack, with fiber connections directly linking servers. There are two fabrics, east-west Spectrum-X for GPU compute, and converged north-south for Storage, in-band and OOB 1 Gigabit management traffic.

For the respective design point please see the switch, transceiver and cable counts in **'Error! Reference source not found'** and **'Error! Reference source not found'**.

For management connectivity this design uses a SN2201 switch per SU. For example, in Figure 5 there are 4x SN2201 switches uplinked to the core via 2x 100G links per switch.

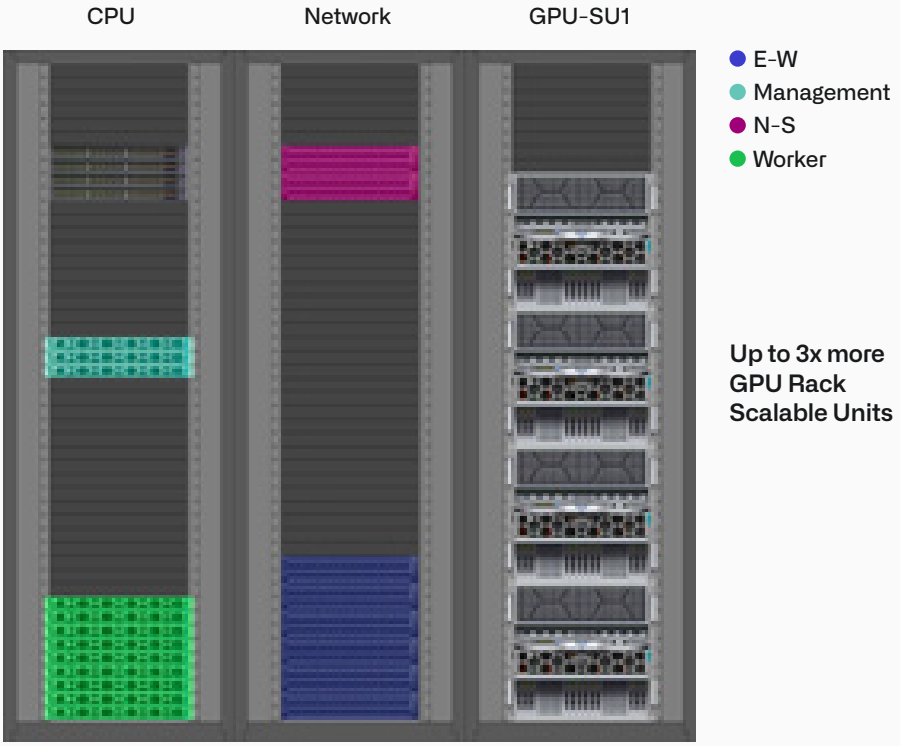
AIOS RA "Small" Cont.

For management connectivity this design uses a SN2201 switch per SU. For example, in Figure 5 there are 4x SN2201 switches uplinked to the core via 2x 100G links per switch.

Allowable variations on the rack elevation include separating the GPU servers into multiple racks if power density isn't available in the datacenter. The network switch rack and the platform server rack can also be combined, but this will limit future extendibility, see the medium rack elevation below.

Figure 5

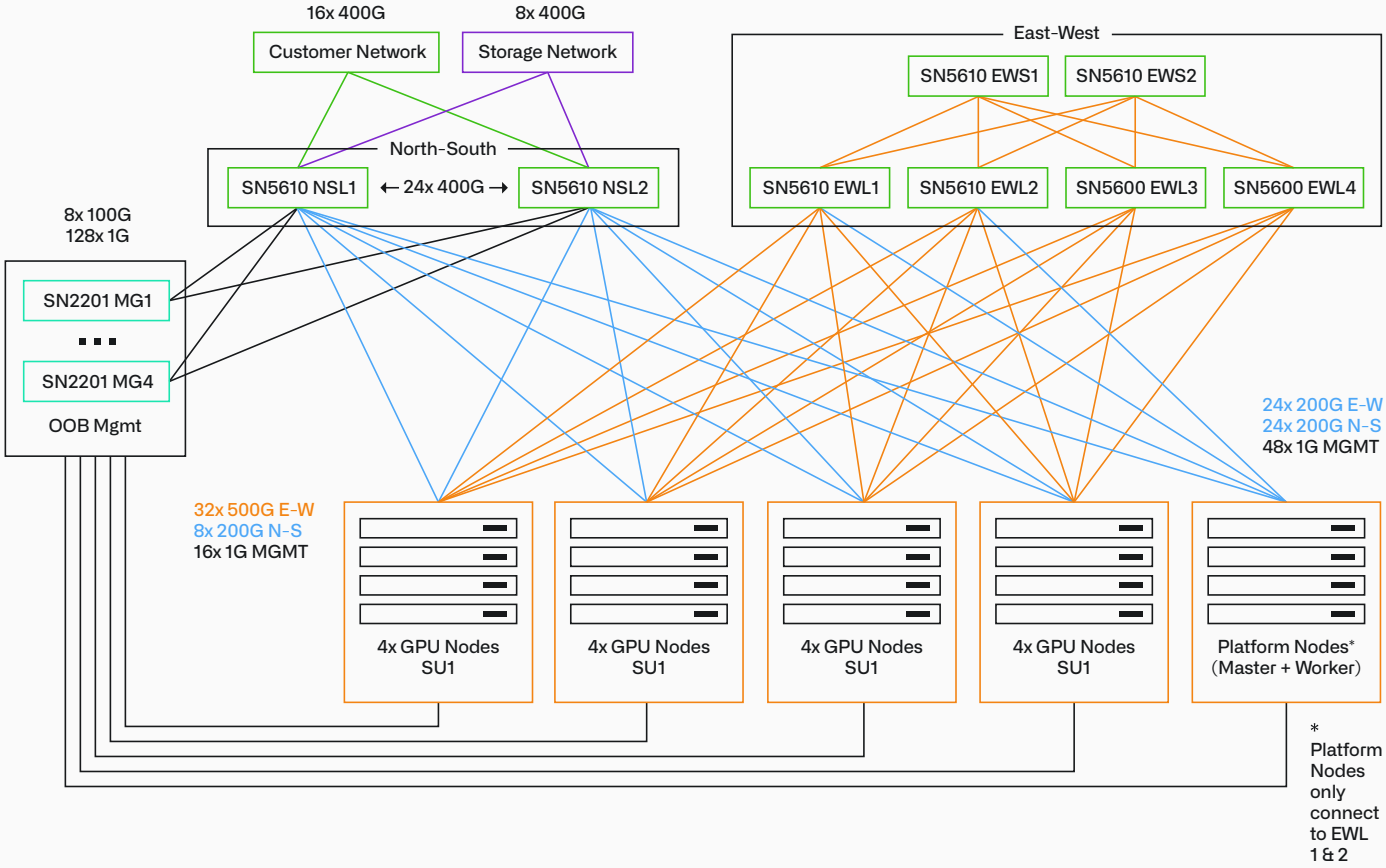
Overview – Base (Expandable to 4 SU)



AIOS RA "Small" Cont.

Figure 6

Small AIOS RA with 28 servers (128 GPUs)



In the small RA there are 4 HGX B300 server standard units (SU), with 4 server nodes per unit.

The network has been designed to be cost efficient with the CPU (N-S) Network on a non-blocking, converged two switch fabric. The GPU Compute (E-W) Network has a separate Leaf-Spine network to provide high-bandwidth, low-latency node to node communications.

Each SU has 32x 400G uplinks to the fabric for GPU Compute traffic (East-West) and 8x 400G uplinks to the fabric for North-South traffic.

The cluster is designed for up to 12 Platform Nodes all dual-connected at 200G to the North-South Network and dual connected to the east-west fabric. Platform nodes are only connected to east-west leaf switches one and two.

The North-South Network is overprovisioned to support 64 100G/200G network connections towards the customers' network. Additionally, there are there are 32 100G/200G network connections for storage attachment. This is designed to provide at least 12.5Gb of bandwidth per GPU.

AIOS RA “Small” Cont.

The design assumes that rack layout provides power supply redundancy within a rack. Otherwise, another rack layout should be considered. The number of HGX B300 servers per rack is also dependent on the rack power available.

The single plane Spectrum fabric for the Compute (Node E-W) Network is rail-optimized and non-blocking using 64-port switches facilitating high availability, resiliency and load-balancing. Principally in this design as there are 4 leaf switches in a plane, each leaf switch supports 2 rails e.g. (1+5 [EWL1], 2+6 [EWL2], 3+7 [EWL3] and 4+8 [EWL4]) respectively. An example of a recommended single-ported optic that can be used for this configuration at the ConnectX-8 OSFP port is shown here.

VLAN isolation is used for the various fabrics on the Converged (Node North-South) Network that have been collapsed into the physical infrastructure.

The 200 GbE management fabric supports high availability (HA) using eight100GbE connections to the core network. All nodes are connected to 1Gb management switches which in turn are uplinked into the core network.

AIOS RA “Medium”

The rack elevation below shows the medium AIOS RA architecture, featuring 3 master servers and 21 workers. The GPU portion of the architecture starts at 4 scalable units of GPU compute and can scale up to 10, for a total of 320 B300 GPUs.

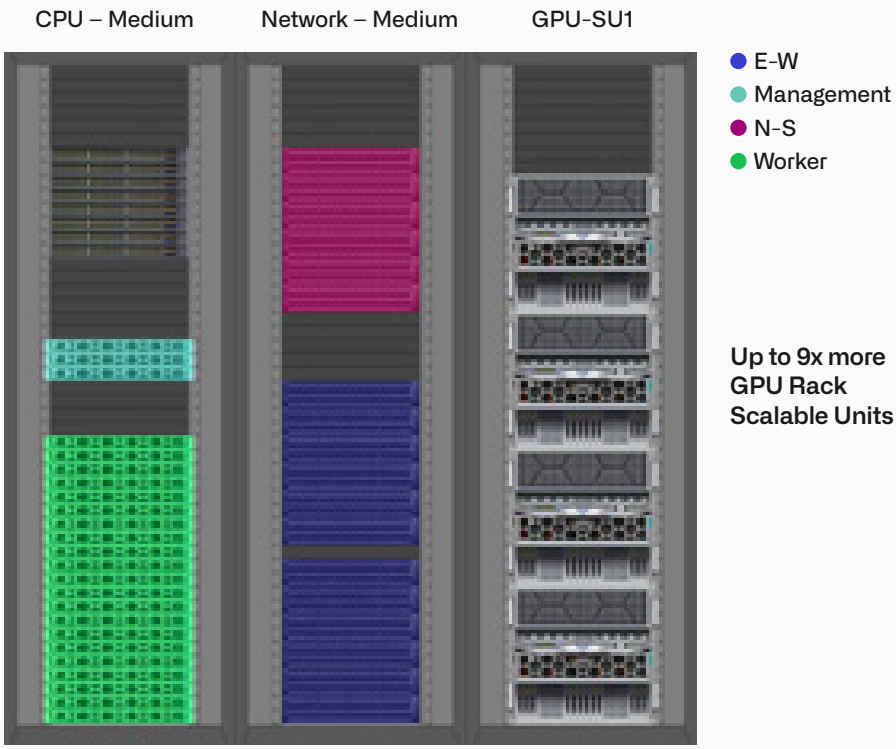
Like in the small RA the networking is centralized in a single rack, with fiber connections directly linking servers. Two separate Spectrum-X fabrics support east-west and north-south traffic, and an OOB 1 Gigabit management network is also included.

For the respective design point please see the switch, transceiver, and cable counts in **‘Error! Reference source not found’** and **‘Error! Reference source not found’**.

For management connectivity, this design uses a SN2201 switch per SU. For example, in Figure 7 there are 8x SN2201 switches uplinked to the core via 2x 100G links per switch.

Figure 7

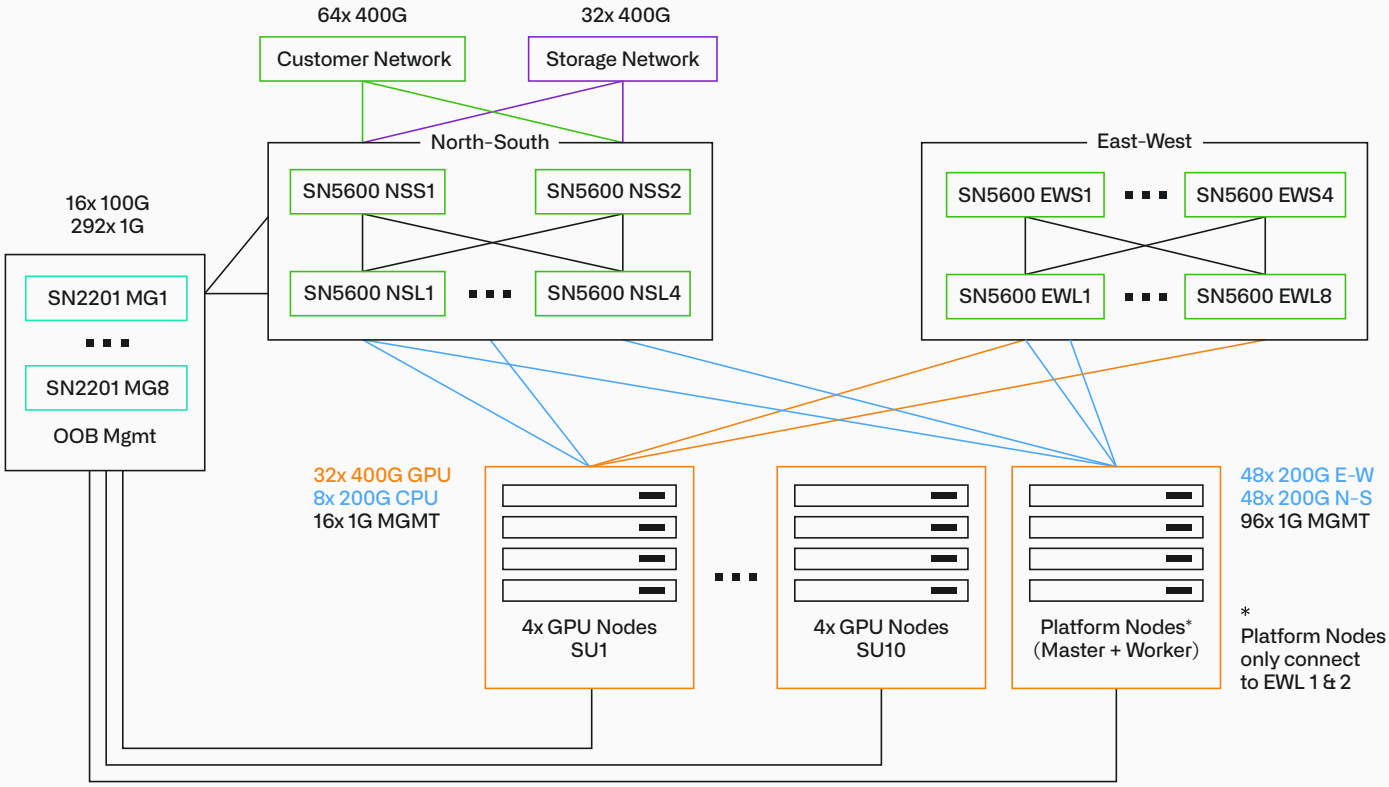
Overview – Base (Expandable to 10 SU)



AIOS RA “Medium” Cont.

Figure 8

Medium AIOS RA with (3+21+40) servers (320 GPUs)



In the medium RA there are up to 10 HGX B300 server Scalable Units (SU), with 4 server nodes per unit.

Both the North-South and East-West fabrics are implemented as non-blocking spine-leaf topologies.

The East-West fabric receives 2 uplinks per platform node at 200Gbps, for a total of 48 connections. CPU and GPU Nodes are intended to be treated interchangeably, for this reason, they will connect to EWL1 & 2, while maintaining unpopulated ports on EWL3-8 to maintain plane consistency and flexibility for replacement with CPU or GPU node. Additionally, the GPU SUs contribute 8 400Gbps connections per node. The connections are rail optimized and divided among the 8 East-West Leaf switches shown in the configuration. Platform node East-West connections are only made to leaf switches one and two.

AIOS RA “Medium” Cont.

The North-South Network takes 2x 200Gbps connections from each GPU and platform node. The cluster is designed for up to 24 Platform Nodes.

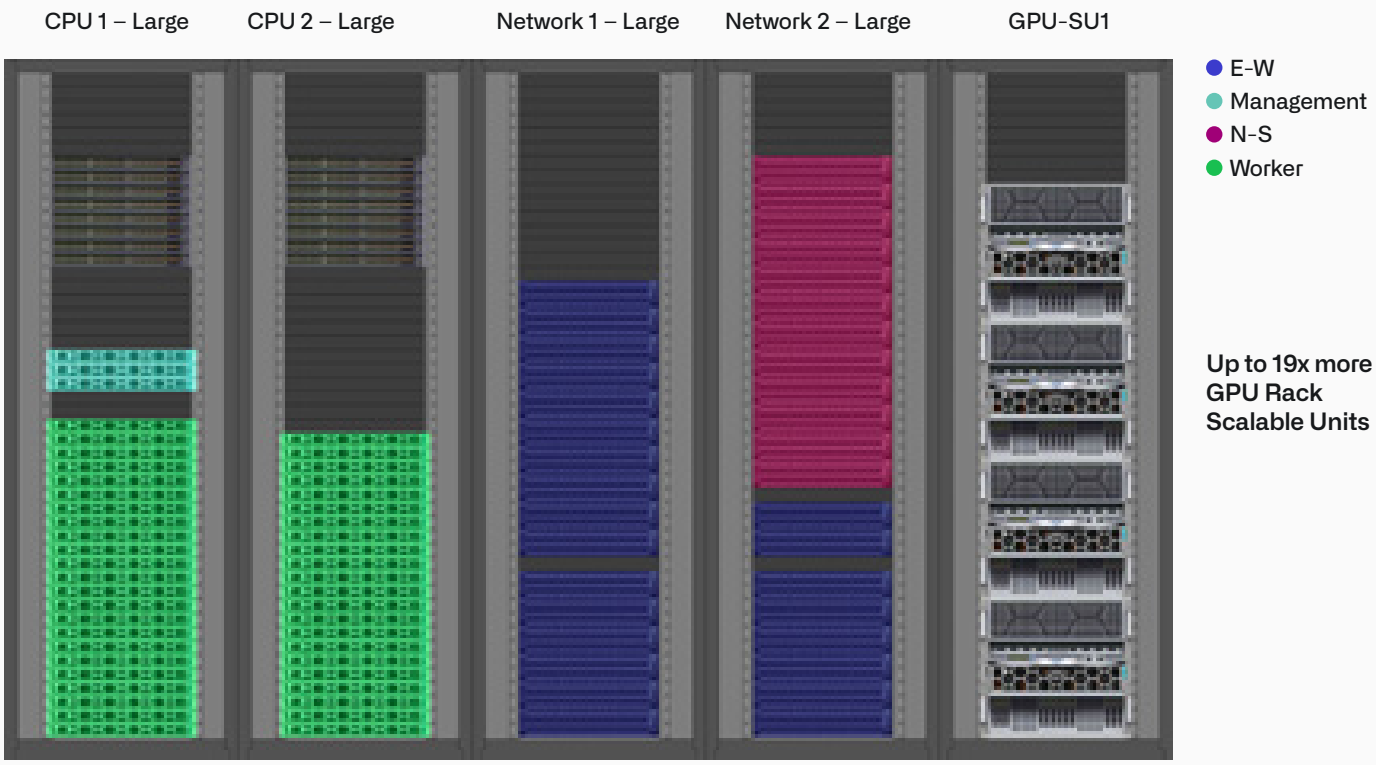
The North-South Network is overprovisioned to support 64 400G network connections towards the customers’ network. Additionally, there are 32 400G network connections for storage attachment. This is designed to provide at least 12.5Gb of bandwidth per GPU.

The 200 GbE management fabric supports high availability (HA) using sixteen100GbE connections to the North-South network. All nodes are connected to 1Gb management switches which in turn are uplinked into the North-South network.

AIOS RA “Large”

Figure 9

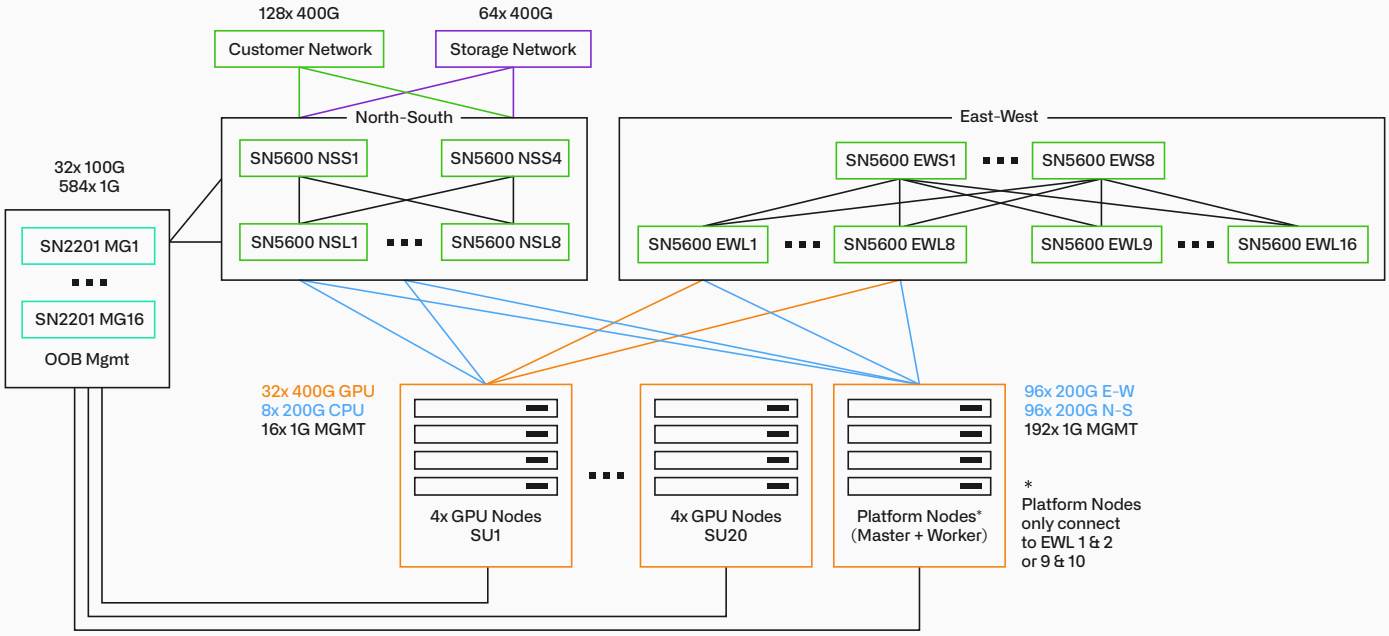
Overview – Base (Expandable to 20 SU)



AIOS RA "Large" Cont.

Figure 10

Large AIOS RA



In the Large RA there are up to 20 HGX B300 server Scalable Units (SU), with 4 server nodes per unit.

Both the North-South and East-West fabrics are implemented as non-blocking spine-leaf topologies.

The East-West fabric receives 2 uplinks per platform node at 200Gbps, for a total of 96 connections. CPU and GPU Nodes are intended to be treated interchangeably, for this reason, they will connect to EWL1-2, or EWL9-10, while maintaining unpopulated ports on EWL3-8 or EWL11-16 to maintain plane consistency and flexibility for replacement with CPU or GPU node. Additionally, the GPU SUs contribute 8 400Gbps connections per node. The connections are rail optimized and divided among the 16 East-West Leaf switches shown in the configuration. In the Large RA Platform node East-West connections are only made to leaf switches one and two, or nine and ten.

The North-South Network takes 2x 200Gbps connections from each GPU and platform node. The cluster is designed for up to 48 Platform Nodes.

The North-South Network is overprovisioned to support 128 400G network connections towards the customers' network. Additionally, there are 64 400G network connections for storage attachment. This is designed to provide at least 12.5Gb of bandwidth per GPU.

The 200 GbE management fabric supports high availability (HA) using 32100GbE connections to the North-South network. All nodes are connected to 1Gb management switches which in turn are uplinked into the North-South network.

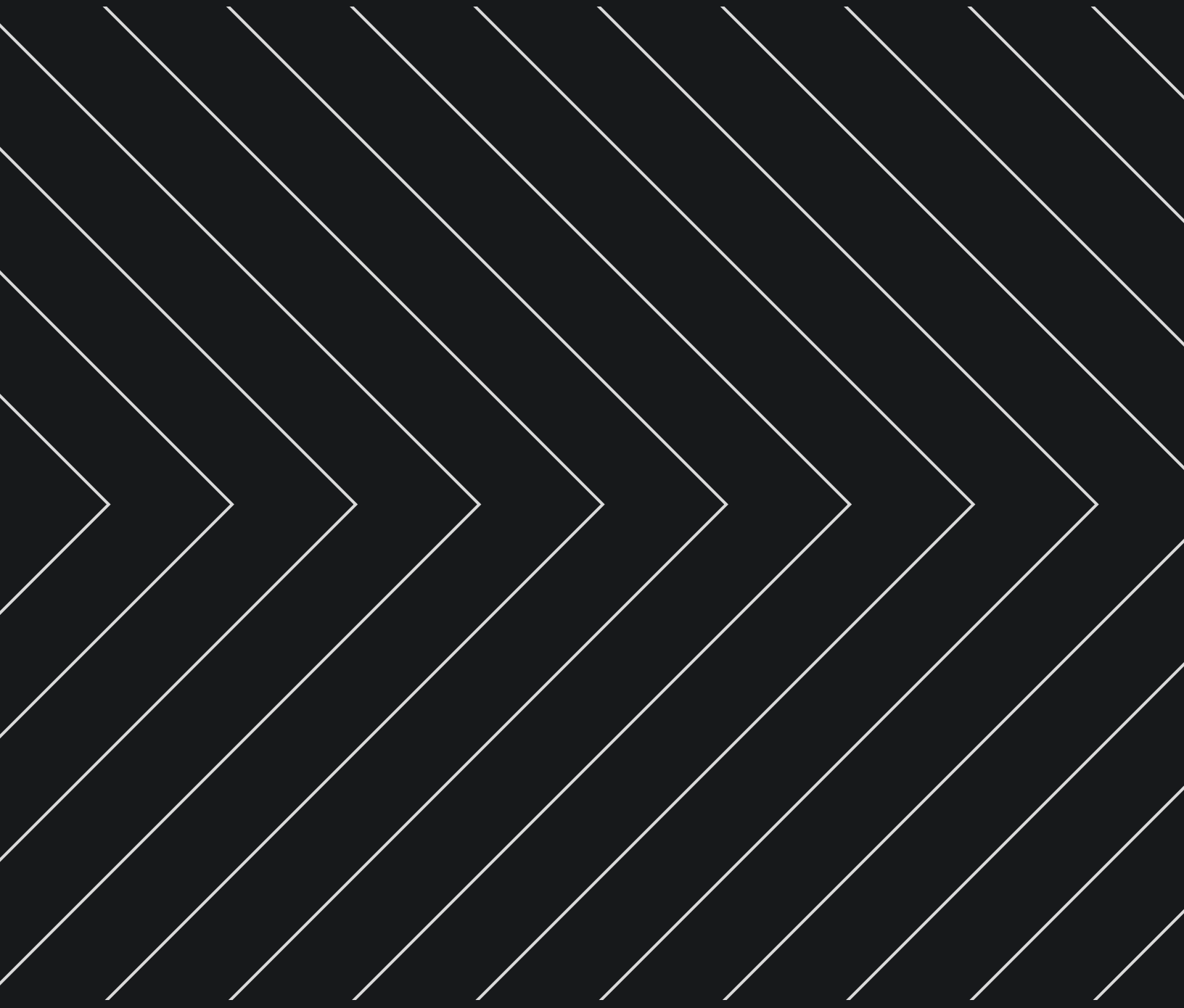
Summary

The Palantir AI Operating System Reference Architecture, co-developed by Palantir and NVIDIA, delivers a secure, high performance foundation to deploy, integrate, and expand AIP across enterprise and government environments while staying current with rapid GPU and networking advances. It combines Palantir's production grade AI platform practices with NVIDIA's accelerated computing roadmap to guarantee optimal performance for real world AIP workloads and seamless scaling from pilot to multi site operations. The design emphasizes out of the box compliance and governance patterns aligned to regulated missions, accelerating authority to operate outcomes with built in access controls, auditing, and privacy guardrails integral to Palantir's platforms.

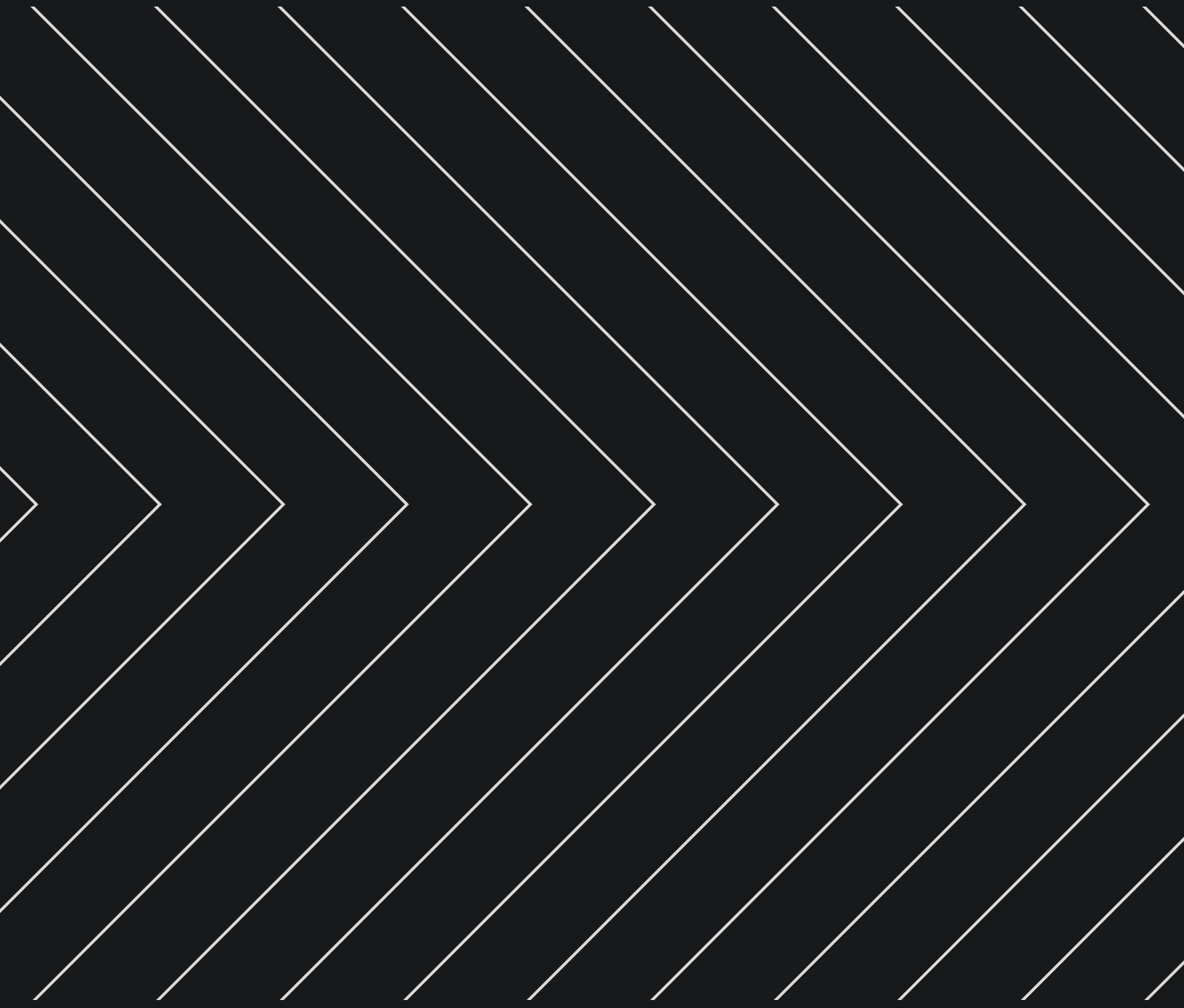
Security is engineered from the ground up, featuring a zero trust, containerized runtime that isolates services, enforces least privilege, and maintains continuous auditability for sensitive AI operations in production. Integration and deployment are pre tested to reduce risk and shorten time to value, pairing validated software patterns with reference infrastructure so enterprises can stand up AIP quickly and confidently across diverse environments. The architecture keeps pace with NVIDIA's newest GPU generations and high performance interconnects, aligning with reference designs for modern AI data centers to ensure sustained throughput and efficiency as models and inference demands grow.

Enterprises retain choice through a best of breed hardware approach, selecting among multiple server vendors while inheriting a consistent AIP software stack and lifecycle model across on prem, cloud, and edge footprints. Scale up and scale out pathways are explicitly supported: densify nodes for larger models or expand clusters and sites for capacity and resilience, all managed with production controls for governance and observability. The result is a future proofed reference architecture that marries operational security with cutting edge acceleration, enabling organizations to deploy AIP today and evolve alongside next wave GPU and networking innovations without disruptive redesigns

The future of cloud



is on-prem.




Palantir
NVIDIA