theCUBE Research White Paper

Optimizing Neoclouds and Sovereign Clouds:

How Cisco's Nexus One Accelerates GPUaaS and AI Factory Performance

Date: February 2026 **Author:** Bob Laliberte

Executive Summary: A new class of cloud providers, commonly referred to as neoclouds and sovereign clouds, is emerging to meet accelerating demand for GPU-as-a-Service (GPUaaS), AI factories, and region-specific or sovereign AI platforms. Built to support the compute-intensive, highly synchronized nature of modern AI workloads, these environments differ fundamentally from both traditional enterprise data centers and hyperscale cloud platforms.

Based on multiple analyst estimates, a reasonable central view suggests that customers may direct several trillion dollars in cumulative spend toward sovereign and GPU-specialized neocloud and sovereign cloud providers over the next decade, as AI factories and data sovereignty requirements reshape the cloud landscape. Published reports further project that neoclouds alone will invest more than one trillion dollars in data center infrastructure during this period, with networking accounting for a meaningful share of that investment. Sovereign clouds are expected to contribute an additional quarter of a trillion dollars to data center infrastructure, with a significant portion directed to networking technologies and optics.

At the same time, AI workloads are evolving beyond large-language-model training to agentic and now, physical AI use cases. These newer workloads introduce tighter coupling among compute, storage, and networking, significantly increasing sensitivity to latency, jitter, congestion, and performance variability. As a result, networking is no longer a secondary consideration. It has become a determining factor in AI efficiency, job completion time, and overall GPU utilization.

Neocloud and sovereign cloud providers face distinct challenges. They must deliver deterministic, lossless performance at scale while operating with far fewer specialized engineering resources than hyperscalers. They must support multi-tenant environments with strong isolation and security guardrails, often across distributed and sovereign regions. They must also deploy new GPU clusters rapidly, as delays in bringing capacity online directly impact revenue and competitiveness.

To address these challenges, Cisco has developed an integrated Ethernet-based AI networking approach designed for large-scale, multi-tenant AI environments. This approach combines custom silicon using Cisco Silicon One and Cloud Scale ASICs, high-performance Nexus 9000 (N9000) switching platforms, advanced congestion-aware traffic management, integrated security, and centralized operations through Nexus Dashboard and HyperFabric. Cisco further complements this architecture with validated reference designs developed in collaboration with key ecosystem partners.

Together, these capabilities are intended to help neocloud and sovereign cloud providers scale AI infrastructure efficiently, reduce operational complexity, maintain predictable job completion times, and accelerate time-to-market. This positions these providers to capture the next phase of growth in GPUaaS and AI factory offerings.

Evolving AI Requirements Drive Demand

AI workloads are expanding beyond Large Language Models (LLM) to Agentic AI and now Physical AI, increasing the importance of reliable, high-performance networking. These emerging workloads introduce tighter coupling between compute, storage, and network, amplifying sensitivity to latency, jitter, and congestion.

Most began their AI journey with LLMs that rely heavily on GPUs/XPUs and focus on understanding and generating human language. These workloads accelerated adoption due to their broad accessibility and consumer visibility. These LLMs are hosted by hyperscalers and leveraged by enterprises or consumers. However, enterprises are now exploring the benefits of Agentic AI and Physical AI. While LLMs focus on understanding and generating human language, Agentic AI adds planning, memory, and tool use, enabling multi-step workflow orchestration across distributed applications and data sources, and Physical AI enables use cases such as autonomous vehicles or autonomous mobile robots.

As agentic systems scale, the need for deterministic throughput and stable latency grows substantially. Furthermore, Physical AI takes that one step further by extending agentic intelligence into the physical world via robots, drones, autonomous vehicles, and industrial systems. Physical AI requires real-time perception and control, dramatically increasing sensitivity to jitter, latency, and congestion. It also generates significantly more machine data, making reliable network access critical.

New Performance Expectations to Accommodate

To accommodate this shift, networks must scale from dozens to thousands of GPUs to handle complex, high-intensity AI workloads while remaining lossless, non-blocking, and observable. Given the supply chain challenges for GPUs and data center capacity, there is substantial demand for GPU as a Service (GPUaaS), and AI factories. Neoclouds and sovereign clouds have emerged to fill this gap. Their goal is to provide an agile, efficient, and cost-effective solution for organizations to access this technology.

The needs of existing and emerging neoclouds and sovereign clouds can best be addressed by leveraging AI-ready data center solutions, with particular focus on the network. The ability to maintain a lossless operation even under heavy and hybrid loads, ensures predictable and repeatable job completion times (JCT) and high-bandwidth east-west performance, which are a must for these providers. Neoclouds must have fine-grained visibility into congestion and insights into jitter and latency levels. Technologies such as RDMA over Converged Ethernet (RoCE v2), advanced load balancing, packet-level intelligence, alignment with the Ultra Ethernet Consortium initiatives, reference architecture blueprints, and readiness for industry benchmarks and frameworks (including MLCommons and NVIDIA Collective Communications Library (NCCL)) should be considered as core requirements.

Neoclouds and Sovereign Clouds: Represent A New Class of AI Infrastructure Providers

Neoclouds and sovereign clouds are non-hyperscalers offering GPUaaS and/or AI factory capabilities, typically targeting regional, sovereign, or vertical-market needs. They provide enterprises, Independent Software Vendors (ISVs), and startups with accelerated access to GPU resources without the overhead of building hyperscale-level infrastructure. Examples of emerging neocloud providers include CoreWeave, Lambda, Nebius, and sovereign cloud providers include OVHcloud, Humain, Core42, and Deutsche Telekom T-Systems.

This new class of provider prioritizes architectures and commercial models that maximize speed and efficiency: pre-validated designs, simplified procurement, and repeatable deployment patterns that shorten time-to-market and reduce implementation risk, while preserving architectural openness and flexibility.

Network Requirements for Neoclouds and Sovereign Clouds

Neoclouds and sovereign clouds have stricter network requirements than traditional enterprise data centers or hyperscalers, as they must operate multi-tenant environments where customers share GPU clusters and network infrastructure. The requirements include:

- Providing robust multi-tenant micro-segmentation using technologies such as EVPN-VXLAN and VRFs, with the ability to enforce policy and isolate traffic down to the GPU/DPU level for guaranteed security and performance.
- Maintaining uniform performance across noisy multi-tenant environments.
- Unified management across front-end and back-end fabrics, with the option to run a common Ethernet-based fabric for server and storage access (10/25/50/100G+) and GPU interconnects (100/400/800G+ over RoCEv2 Ethernet or InfiniBand) to simplify operations and increase flexibility. Deliver uniform traffic distribution across all links to eliminate traffic congestion.
- Proactively detect congestion and mitigate it to ensure optimized performance.
- Rapidly detect link degradation or failure to minimize delays or outages.
- Ensuring consistent job completion times to meet customer SLAs and obtain repeat customers.
- Operational simplicity and centralized observability reduce operational overhead, enable small teams to operate large-scale platforms, and improve insight into performance and faults.

Together, these capabilities form the baseline for reliable, multi-tenant network environments that support GPUaaS and AI-factory services at scale, ensuring predictable job completion times and customer satisfaction.

Challenges Facing Neocloud and Sovereign Cloud Providers

While there is a tremendous opportunity for neoclouds and sovereign clouds, several challenges may arise. The decisions these providers make today could have a profound impact on their ability to operate efficiently, remain agile, scale cost-effectively over time, and harness sustainable profitability. Those challenges include:

1. Choosing Between InfiniBand and Ethernet Networks for back-end fabric: Many neocloud and sovereign cloud providers initially deployed InfiniBand-based GPU bundles for the back-end network, but often they must reconsider that choice as they scale. Challenges when choosing the network include:

- Limited InfiniBand expertise in the market. These technologies have largely been confined to supercomputing environments, and equivalent operational expertise is not widely available. As a result, neoclouds and sovereign cloud providers may face delays and increased costs as organizations compete for scarce specialized talent.
- Operational inconsistency between Ethernet-based front-end networks and InfiniBand back-ends. Neoclouds and sovereign cloud providers have universally adopted Ethernet for their front-end networks; however, organizations that bundled InfiniBand must manage two separate technologies, which increases complexity.
- Tooling and ecosystem fragmentation with two different technologies, providers would have to stock extra components, potentially deal with multiple different network vendors and ecosystem partners.
- Higher costs at scale. It is imperative to compare the economies of scale for Ethernet vs InfiniBand, not just on day one, but as the environment scales up, out, and across.
- Ethernet Performance remains a key consideration: Can RoCE v2-enabled Ethernet deliver lossless, low-jitter fabrics on par with InfiniBand at scale?

2. Delivering True Multi-Tenant Segmentation: Neoclouds must securely isolate tenants at multiple layers, network, compute, and GPU, while preserving performance. This includes enforcing segmentation through EVPN-VXLAN, VRFs, and policy-based controls, and extending isolation down to the GPU/DPU level for both security and Quality of Service (QoS). The challenge is to deliver strong multi-tenant separation, including micro-segmentation and traffic shaping, without introducing congestion or increasing job completion times (JCT) for shared GPU clusters.

3. Operational Complexity and Skill Constraints: InfiniBand skill sets are scarce; managing large AI clusters requires a vast team of IT-skilled professionals; overseeing 32,000+ GPU clusters require a massive, skilled workforce. Additionally, even Ethernet-based RoCEv2 fabrics require careful Explicit Congestion Notification (ECN)/Priority Flow Control (PFC)/Date Center Quantized Congestion Notification (DCQCN) tuning, QoS, and dynamic load balancing (DLB). Many neoclouds and sovereign clouds lack hyperscaler-style engineering resources, and CLI-only operations don't scale. GPU clusters are still relatively new, and neoclouds and sovereign clouds will face challenges in finding specialized, skilled resources to build robust environments.

4. Time-to-Market Pressures: GPUaaS revenue depends on how quickly new capacity can go live or achieve the "time to first job". Neoclouds and sovereign Clouds can't afford to engage in science experiments or take the time to research and train resources over time. Trying to piece together a complete solution from multiple vendors can be time-consuming, not to mention the conflicts and communication issues that arise during deployment. Assembling a solution from multiple vendors raises compatibility and support challenges and may not meet unpredictable scaling demands. Additionally, these piecemeal solutions may lack zero-touch provisioning and full automation, further delaying the "time to first job".

5. Scaling Up, Out, and Across Regions: Inability to Scale Up, Out, and Across in required timeframes. Neoclouds and sovereign clouds must consider their ability to grow beyond a single site. Power and space limits drive multi-site and regional deployments. Network capabilities must accommodate all three dimensions (Up, Out, Across). As AI factories expand, they will need:
    - Scale-up capabilities to drive denser fabrics per site
    - Scale-out requires replicating clusters across campuses
    - Scale-across requires multi-region fabrics with high-performance Data Center Interconnects (DCI)

To be successful networks must support all three dimensions reliably.

6. Ensure High Availability Environments: Downtime isn't an option for neoclouds and sovereign clouds. In a rapidly emerging space, competition is always seeking an advantage, and it will be imperative that the infrastructure be highly available at every level. This is especially true for the network. Are you working with a network vendor that offers high availability? Does the Network Operating System (NOS) have supervisor redundancy, graceful restarts, and in-service software upgrades, which are critical for fixed switches without redundant supervisors that require hitless software upgrades and critical patching, without impacting the network? Does it leverage dual-plane designs where the data center network is built using two completely independent network planes operating in parallel? At the topology level, does the AI fabric use inherently fault-tolerant designs such as leaf–spine with emerging/dynamic load balancing at larger scales, dual- or multi-plane architectures built as independent network planes in parallel?

7. Securing the Network Environment: Neocloud and sovereign cloud providers face a blend of hyperscaler-level technical challenges and enterprise-grade compliance requirements. The most acute risks center on multi-tenant GPU isolation, data sovereignty enforcement, AI pipeline integrity, silicon-level security, and zero-trust enforcement on high-speed AI fabrics. These challenges only compound as the neoclouds and sovereign clouds expand globally. Are you able to deliver zero-trust segmentation, and at what levels can that be accomplished? Does the solution incorporate security at the DPU level, and are you able to quickly find, mitigate, and bounce back from any vulnerabilities?

8. Addressing High-Speed Optics and Physical Layer Constraints: As AI fabrics move to 800G, and beyond, the network becomes constrained not just by topology design but by optics choice, fiber plant, and power budgets. Neoclouds and sovereign clouds must balance reach, density, and cost across Active Optical Cables (AOCs), Direct Attach Copper cables (DACs), and pluggable optics, while managing the significant power and thermal footprint of high-speed ports.

The Case for Ethernet in AI Fabrics

The discussion surrounding Ethernet vs InfiniBand has been active since Generative AI technology gained significant traction several years ago. While InfiniBand took an early lead due to performance concerns and its bundling with GPU packages, Ethernet continues to adapt through the work of the Ultra Ethernet Consortium and the development of RoCEv2. As a result, the market is shifting toward Ethernet adoption in back-end GPU environments. According to theCUBE research[1], respondents prefer Ethernet-based network solutions to support their AI workloads. 59% of organizations surveyed prefer Ethernet for AI workloads, compared with 38% for InfiniBand. See Figure 1. When asked why an organization would prefer one over the other, Ethernet supporters stated that Ethernet is widely deployed across the network, they have the appropriate in-house skills, and even major cloud providers use Ethernet. Outside of the top three reasons for Ethernet, organizations cited that they believed RoCE v2 could provide comparable performance.

[1] Source: theCUBE Research Report: The Impact of AI on the Network

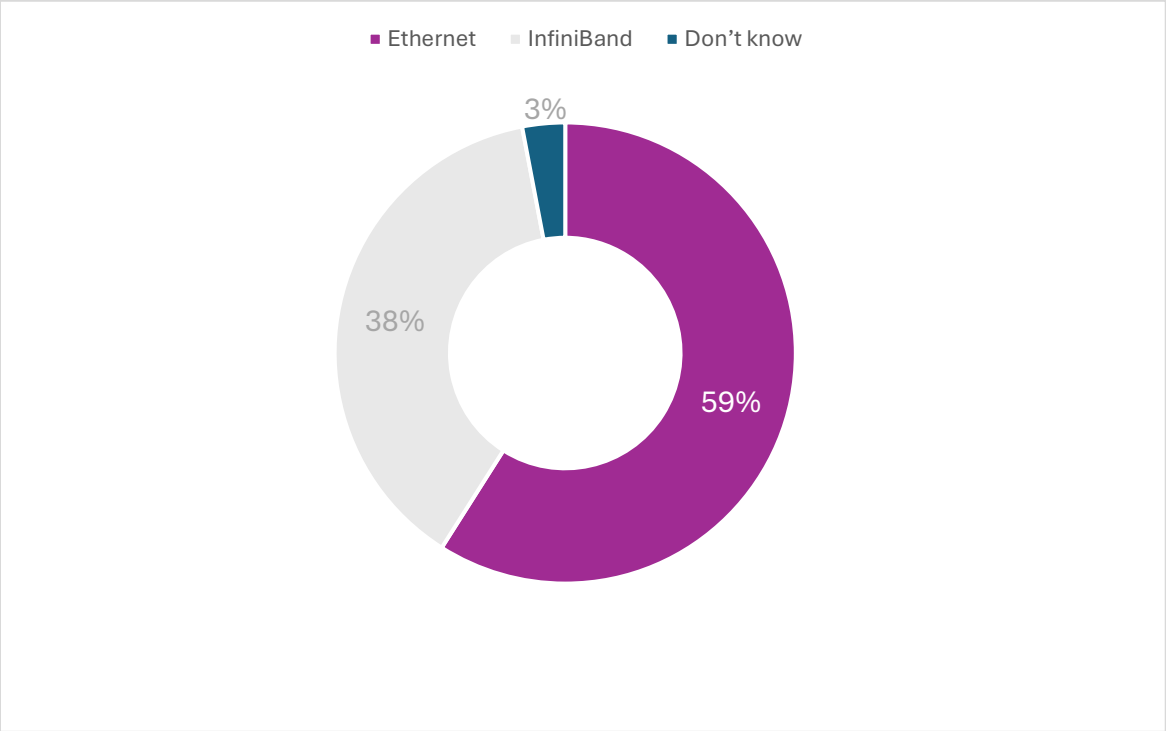

Figure 1. Qn. What networking technology do you prefer to support AI workloads/clustered AI workloads? Select one

It should also be noted that validated reference architectures are the top requirement cited by respondents considering Ethernet for back-end AI networks. This highlights that emerging neoclouds and sovereign cloud providers, are looking to their vendors to minimize the risk of deploying these new environments and accelerate time to market and value through proven configurations optimized for performance and security.

Cisco's Integrated AI Networking Stack

To enable neoclouds and sovereign clouds to accelerate their time-to-market, Cisco has created a comprehensive AI networking stack. Purpose-built for AI pipeline workloads, Cisco Nexus One integrates silicon, hardware, software, management tools, and optics. It supports lossless, low-jitter, low-latency, high-throughput fabrics and enables seamless scaling from dozens to thousands of GPUs while preserving performance and job completion time. This integrated approach simplifies deployment, reduces operational complexity, and minimizes risk for neoclouds and sovereign clouds as they build their GPUaaS infrastructure. The integrated architecture includes:

- Silicon: Cisco provides two families of custom silicon, Cloud Scale ASICs and Silicon One, to accommodate a wide range of networking applications. Its Cloud Scale ASICs are well-suited for environments that require scale, performance, and flexibility. Cloud Scale ASICs are available in Cisco N9000 series switches and can be used for high-density fabrics. Silicon One offers unified, programmable, scalable architecture that can span all major networking domains, reducing complexity, unifying operations, enabling feature consistency, and future-proofing networks against evolving data-intensive applications and AI/ML demands. The Silicon One G200 in the Nexus 9364E-SG2 (51.2Tbps full-duplex switching, high 512-radix) delivers 800G for scale-out training and inference. It is designed to enable AI cluster fabrics that require low latency, deterministic traffic, and massive throughput. Neoclouds and

sovereign clouds can also take advantage of the Silicon One P200 switching platform to deliver (51.2 Tbps full-duplex deep buffer routing) scale across distributed GPU cluster environments.

  - Advantage: These advanced custom ASICs provide neoclouds and sovereign clouds with the scalable performance and efficiency required to support growing AI workloads and ensure future-resilient infrastructure supporting next-generation GPU AI factories. By owning the complete silicon and software stack, Cisco can optimize holistically, delivering higher scale, richer features, and better cost/performance ratios for neoclouds and sovereign clouds. This delivers "future-proofing" and better investment protection: more endpoints, bigger fabrics, and easier growth without wholesale hardware changes. Given Cisco's accelerated innovation cadence, future generations of switching platforms are expected to continue scaling bandwidth, performance, and efficiency, providing additional headroom for evolving AI pipelines.

- Systems: The Nexus 9000 (N9000) Series is Cisco's flagship data center switching platform, engineered to power high-performance, scalable, cloud and AI-ready networks for front-end and back-end networks. The series supports 1/10/25 GbE up to 800 GbE, with a solid path to 1.6 Tbps, and is offered in both fixed-port and modular configurations. Examples include the Nexus 9364E-SG2 series (Silicon One G200), which delivers 64 x 800G ports (2RU), up to 512 x 100G via breakouts, and has a 256MB fully shared packet buffer for burst absorption and lossless RoCEv2 fabrics. The Nexus 9300-GX2 Series (Cloud Scale) offers 64/48/32 x 400G variants, with MACsec and IPsec options and shared-memory buffers. The recently announced Cisco N9100 Series switches, powered by NVIDIA Spectrum-X Ethernet switch silicon, deliver 51.2 Tbps of switching bandwidth, ultra-low-latency performance, and support for lossless Ethernet fabrics, enabling high-throughput AI workloads in neoclouds and sovereign cloud environments. The N9100 series offers operational flexibility through support for NX-OS or SONiC (Software for Open Networking in the Cloud), ensuring seamless integration into existing data center architectures and alignment with open networking preferences. Cisco offers a wide range of validated reference designs for both back-end and front-end AI networks based on the N9000 systems.

  - Advantage: With validated reference blueprints, these switches provide neoclouds and sovereign clouds with proven, high-performance building blocks that accelerate time-to-market and ensure predictable job completion times for their GPUaaS offerings. To further optimize power and cooling efficiencies, these providers should expect options for 100% liquid cooling. Liquid cooling matters for AI fabrics because thermal constraints now shape network architecture, performance, and economics. Liquid cooling is a critical enabler of scalable, high-performance AI networking; without it, AI fabrics are forced to compromise on density, determinism, and cost efficiency.

- Optics: Cisco offers a wide range of transceiver/optical modules. Its full portfolio spans from 1G to 800G and commitments for 1.6T, including 400G BiDi, QSFP-DD (Quad Small Form-factor Pluggable- Double Density) & OSFP (Octal Small Format Pluggable), DAC (Direct Attach Copper cables), and Linear Pluggable Optics (LPO)-ready designs. Cisco has rigorously tested interoperability and publishes a comprehensive optics-to-device compatibility matrix that removes any guesswork. These optics help to reduce the footprint and energy consumption in GPUaaS environments.

  - Advantage: Cisco's optics portfolio including the latest OSFP and LPO technologies, backed by a strong supply chain, reduces total cost of ownership, ensures scalability and future proofing, and simplifies procurement for neoclouds and sovereign clouds. The energy-efficient solutions ensure providers can

optimize sustainability goals. The optics are optimized for AI clusters and multi-tenant fabrics to support high throughput, low latency, and scalable interconnects.

- Software: Cisco's software stack is designed to provide scalable, automated, secure, low-latency data-center fabrics capable of supporting GPU clusters, distributed AI pipelines, and multi-tenant cloud services. It combines network operating systems, advanced telemetry, AI-assisted operations, built-in security, and congestion-aware traffic management into a cohesive AI fabric.

    - NX-OS for Lossless AI Fabrics: Cisco NX-OS is the network operating system that powers Cisco data center switching and storage portfolio in data-center environments. When paired with the N9000 series, it provides a modular, high-performance, highly available OS designed for cloud-scale fabrics, automation, multi-tenant networks, and AI-ready architectures. More specifically, NX-OS delivers multi-tenancy via VRFs (Virtual Routing and Forwarding), EVPN (Ethernet VPN)-VXLAN (Virtual Extensible LAN), and granular policy constructs. It has comprehensive RoCEv2 support and AI-specific lossless fabric features. Cisco NX-OS achieves high availability for AI data centers through features such as supervisor redundancy, vPC (Virtual Port Channel), In-Service Software Upgrades (ISSU), and graceful restarts, enabling resilient, fault-tolerant networks that minimize downtime and support mission-critical applications. Optional SONiC support enables open networking preferences.
        - Advantage: Cisco NX-OS delivers the scalable, automated, highly available, and high-performance network foundation required for neoclouds and sovereign cloud environments, combining flexible operations with cloud-native automation and deep visibility.
    - Proactive congestion management: AI workloads create unique, burst-heavy traffic patterns. Cisco's proactive congestion management includes:
        - Explicit Congestion Notification (ECN) and Weighted Random Early Detection (WRED) for early congestion signaling and intelligent queuing.
        - Priority Flow Control (PFC) for tuning RoCEv2 environments + PFC watchdog for lossless transport and pause storm protection.
        - Per flow congestion awareness in all ASICs offered by Cisco (Cloud Scale, Silicon One, and extended to Spectrum-X technology).
        - Data Center Quantized Congestion Notification (DCQCN) combining Explicit Congestion Notification (ECN) and PFC for end-to-end RDMA control.
        - Dynamic Packet spraying/ adaptive load balancing.
        - Approximate Fair Drop (AFD) to throttle large long-lived flows (elephant flows) without penalizing short flows (mice flows) to ensure fair treatment of different traffic types.

        - Advantage: These capabilities enable neocloud and sovereign cloud operators to maintain highly available, lossless, low-jitter fabrics essential for multi-rack GPU clusters, distributed AI training jobs, high-performance storage backplanes and real-time interference pipelines.

- Intelligent Packet Flow: Cisco's Intelligent Packet Flow capabilities (spanning hardware analytics, adaptive routing, flow-aware telemetry, and algorithmic traffic management) optimize AI networking performance through a combination of advanced load-balancing techniques and real-time traffic awareness, including:
    - Optimal path utilization. The system maximizes path utilization using strategies such as flowlet-based and per-packet load balancing, WCMP (Weighted Cost Multi Path) with DLB, policy-based algorithms, ECMP (Equal Cost Multi Path) pinning, and packet trimming.
    - Congestion-aware traffic management. This provides real-time visibility into microbursts, congestion events, and latency impacts via mechanisms such as congestion signaling, INT (In-band Network Telemetry), and tail timestamping.
    - Autonomous recovery and resiliency. Rapidly reroute traffic around faults to prevent hotspots and maintain fabric stability. Flowlet-based dynamic load balancing creates a "smart highway" for traffic, demonstrating balanced transmission rates and improved job completion times, with additional mixed-mode DLB enhancements available for converged storage and front-end fabrics. Per-packet load balancing further increases utilization by distributing packets across all available links, supported by NIC (Network Interface Card)-level reordering technologies such as NVIDIA's BlueField-3 SuperNICs.
    - Advantage: Intelligent Packet Flow turns the network into an intelligent fabric that continuously adapts to the demands of AI pipelines and distributed compute clusters. It provides higher GPU utilization and more predictable JCT, even in multi-tenant, bursty environments. It directly addresses the requirements of neoclouds and sovereign clouds for uniform traffic distribution and robust congestion avoidance and management, ensuring optimal GPU utilization and consistent performance. New innovations include mixed-mode Dynamic Load Balancing (DLB) for converged storage & front-end fabrics

- Joint Reference Architectures and Ecosystem Integration: Research highlights that organizations believe validated blueprints from network vendors are the best way to become comfortable deploying Ethernet solutions in front and back-end networks. Cisco has worked with major technology vendors to create vendor-agnostic, interoperable, and open reference architectures across its ecosystem of partners. They include:
    - NVIDIA Reference Architectures (RAs) with Spectrum-X, extended via Cisco RAs (AI Infrastructure with N9000, HyperFabric).
    - Cisco switches based on Cisco Silicon One with NVIDIA SuperNICs, and roadmap for switches using NVIDIA Spectrum-X ASICs.
    - Seamless integration of Cisco fabrics with NVIDIA SuperNICs, enabling high performance and uniform traffic distribution.
    - Vendor-agnostic support for NVIDIA, AMD, Intel, VAST, WEKA, and other ecosystem partners.

- Advantage: TheCUBE Research[2] highlights that validated and reference architecture blueprints from network vendors are the best way to accelerate the deployment of Ethernet solutions in AI environments. For neocloud and sovereign cloud providers, Cisco's joint reference architectures, including NVIDIA Spectrum-X and Cisco Ras, provide pre-validated, high-confidence blueprints. They compress design and qualification cycles, reduce integration risk, and directly address skill and time-to-market constraints. This allows providers to stand up Ethernet-based GPUaaS and AI-factory infrastructure quickly, with predictable performance and supportable, multi-vendor interoperability across compute, storage, and network environments.

- Flexible Network Management: Cisco offers flexibility and choice. Neoclouds and sovereign clouds can select on-premises or cloud-based management solutions. Nexus Dashboard can be deployed on-premises and with Nexus One fabric experience provides a centralized platform for different fabric types, and multi-site environments for automated provisioning and real-time visibility (configuration, automation, monitoring, AI jobs, telemetry, and analytics) across data center fabrics. The dashboard provides several AI-specific capabilities, including congestion scores, detailed insights, anomaly detection, end-to-end visibility, and sustainability insights. Cisco Nexus One unifies NX-OS VXLAN EVPN and Cisco ACI fabrics through open standards, delivering a consistent operational and policy experience via Nexus Dashboard while enabling seamless integration across data center, campus, and cloud environments. Cisco Nexus Hyperfabric is a cloud-managed, full-stack AI infrastructure that delivers pre-configured, low-latency AI fabrics with unified lifecycle management and monitoring across compute, GPUs, storage, and networking, enabling fast, scalable, and repeatable AI cluster deployment. Providers can also leverage Hyperfabric in a "bring your own" AI infrastructure that combines their own servers, GPUs, and storage with Cisco networking and Nexus Hyperfabric.

  - Advantage: For Neo and Sovereign clouds that span multiple regions, edge sites, and colocation footprints, Nexus Dashboard provides the single control plane needed to operate distributed fabrics with consistency and speed. Nexus Dashboard with Nexus One directly addresses neoclouds' and sovereign clouds' need for operational simplicity and observability, providing end-to-end telemetry and empowering them to shift from reactive operations to data-driven, semi-autonomous AI networking across different fabric types. Hyperfabric takes that to the next level with turnkey solutions leveraging certified reference architectures or can be used in a BYO AI infrastructure model. These options are crucial for these providers that lack hyperscaler-style engineering resources, as they simplify management and reduce reliance on scarce specialized skills. This shift enables operators to move from reactive troubleshooting to proactive, data-driven operations.

- Integrated Security: Cisco integrates security natively into its cloud-networking software stack. Cisco secures multitenant neocloud environments with end-to-end protections that begin in the fabric itself. VRF and EVPN-VXLAN provide strong tenant isolation, while inline IPsec and MACsec deliver line-rate encryption for both east-west and data-center interconnect traffic. Continuous threat detection and anomaly analytics are enabled through rich streaming telemetry, giving operators real-time visibility into emerging issues. Policies

[2] Source: theCUBE Research Report: The Impact of AI on the Network

are consistently enforced through segmentation-driven controls in Nexus Dashboard, ensuring intent is applied across the entire environment. These capabilities are anchored by secure boot, a hardware root of trust, and signed software images across the N9000 portfolio, providing a trusted foundation for AI-scale operations. Providers can also leverage the Isovalent Enterprise Platform, utilizing eBPF (extended Berkeley Packet Filter), to strengthen security with a path to Zero Trust and multi-tenancy control. Cisco Live Protect enforcement mode provides real-time, kernel-level security for N9000 switches, mitigating vulnerabilities without reboots or downtime and delivering automated protection, centralized management, and continuous operational resilience across NX-OS and ACI environments. Cisco Live Protect enforcement mode, enables operators to mitigate vulnerability exposure with a single click and deploy live shield across all devices.

- Advantage: These security tools support sovereign cloud requirements, tenant isolation for GPUaaS providers, and secure interconnects for regulated industries. In line with Cisco's innovation engine, this integrated approach to security aligns directly with sovereign cloud and regulated-industry requirements: strong tenant isolation for GPUaaS, secure interconnects for distributed AI pipelines, and continuous protection of critical network infrastructure. Neocloud and sovereign cloud providers get a fabric that is secure by design, not bolted on later, and can evolve towards zero-trust, multi-tenant AI networking without sacrificing uptime or operational efficiency. Cisco's software stack for neoclouds and sovereign cloud environments delivers a unified platform combining NX-OS, intelligent packet flow, Nexus Dashboard, embedded security into every layer, and real-time congestion management to provide scalable, automated, secure, and AI-optimized cloud fabrics for next-generation GPU and sovereign cloud architectures.

Cisco Secure AI Factory and Full-Stack Options

Cisco's Secure AI Factory provides an end-to-end blueprint for building modern AI infrastructure, spanning compute, storage, networking, and security, optimized for GPU clusters, high-performance AI pipelines in neoclouds, sovereign clouds, and enterprise AI environments. It can be delivered as a complete, integrated solution or adopted in modular building blocks to meet evolving cloud architectures.

Neoclouds and sovereign clouds increasingly seek turnkey solutions that accelerate deployment while reducing risk. The Secure AI Factory integrates four core components to deliver a high-performance, scalable, and secure foundation for modern AI workloads. AI-optimized Cisco UCS servers provide GPU-dense compute with high-bandwidth interconnects and automated lifecycle management. VAST Data's unified AI storage platform delivers ultra-low-latency, highly parallel access to training datasets, model artifacts, and inference pipelines. DPUs add secure, hardware-accelerated services, offloading networking, storage, and encryption tasks while enabling multi-tenant isolation. Finally, N9000 AI fabrics supply the high-speed, congestion-aware backbone that connects the entire AI Factory, ensuring predictable performance and efficient scaling across training, fine-tuning, and inference clusters. Together, these elements form a cohesive end-to-end architecture or can be adopted modularly to meet evolving neocloud requirements.

The Cisco Secure AI Factory reduces operational complexity and shortens time-to-revenue. Cisco also supports modular adoption, allowing customers to integrate only the networking stack or mix components. Cisco's forthcoming Live Protect enforcement mode enhances security by enabling real-time shielding of vulnerabilities without downtime or reboots.

Our ANGLE

The transition from large language model-focused AI to agentic and now physical AI represents more than a change in application design. It fundamentally reshapes the infrastructure and operational foundations required to deliver next-generation digital services. As AI workloads become more distributed, synchronized, and latency sensitive, the network increasingly determines whether AI platforms can scale efficiently and deliver consistent outcomes.

Neoclouds and sovereign cloud providers are uniquely positioned to meet this demand, but they face constraints that hyperscalers do not. Limited access to specialized talent, pressure to deploy GPU capacity quickly, and the need to operate multi-tenant and sovereign environments all raise the bar for infrastructure design. In this context, experimental architectures and fragmented multi-vendor solutions introduce significant operational and business risk.

The analysis in this paper highlights a clear shift toward Ethernet-based AI fabrics supported by validated reference architectures. Organizations evaluating Ethernet for back-end AI networks consistently emphasize the importance of proven designs that reduce deployment risk, simplify operations, and deliver predictable performance at scale. Success in these environments depends not only on bandwidth but on lossless behavior, congestion awareness, observability, resiliency, and integrated security.

Cisco's end-to-end Nexus One combines custom silicon, a broad portfolio of AI-optimized switching platforms, advanced congestion management capabilities, embedded security, centralized operations, and joint reference architectures developed with ecosystem partners. Together, these elements are designed to reduce operational complexity, mitigate skill gaps, and accelerate time-to-first job and revenue. Neocloud and sovereign cloud providers should explore Cisco's validated Ethernet AI architectures and consult with Cisco experts to accelerate their AI journey.