Gcore AI Cloud Stack and Nokia Data Center Fabric Reference Architecture

Contents

Gcore AI Cloud Stack and Nokia Data Center Fabric Reference Architecture

Combining enterprise cloud orchestration with high-performance networking for AI infrastructure

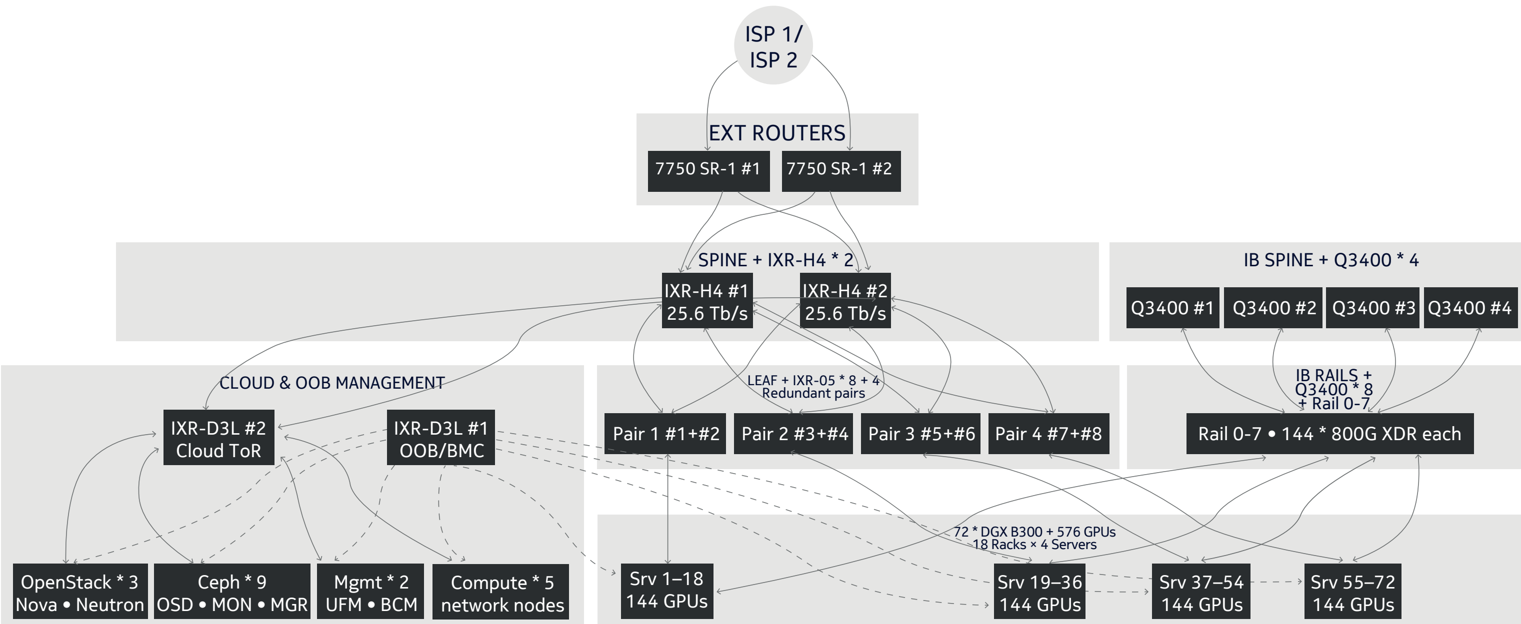
This reference architecture integrates Gcore AI Cloud Stack (OpenStack-based) with Nokia's Data Center Fabric (DCF) to deliver a production-ready foundation for GPU-accelerated AI workloads. The architecture employs a dual-fabric design where the Nokia Datacenter Fabric IXR switches handle north-south Ethernet traffic (frontend network) while NVIDIA Quantum InfiniBand switches manage east-west GPU communication (backend network). This design separates concerns between the East-West InfiniBand fabric for lossless GPU-to-GPU communication and the North-South Ethernet fabric for management, storage access, and external connectivity.

The architecture enables seamless scaling from a single Scalable Unit (72 servers, 576 GPUs) to multi-rack deployments exceeding 500 servers.

Gcore's cloud orchestration layer provides multi-tenant isolation, API-driven automation, and enterprise billing, while Nokia's SR Linux delivers model-driven network automation with 100% YANG coverage for infrastructure-as-code workflows.

This document covers the complete infrastructure stack including compute, networking, storage, security, monitoring, high availability, and facility requirements for production deployment.

Gcore AI Cloud Stack delivers orchestration



Gcore is a European-headquartered company operating infrastructure over **210+ points of presence** across six continents with network capacity exceeding **200 Tbps.** AI Cloud Stack provides IaaS capabilities with a proprietary abstraction layer (Gcore Cloud API) delivering enhanced features and a consistent user experience.

Core platform services

The compute layer supports both virtual machines and bare metal servers with Intel Xeon Scalable and ARM Ampere Altra Max processors. GPU infrastructure includes NVIDIA B300, B200, H200, H100, A100, and L40S with planned GB200 (Blackwell) support. Multi-GPU configurations scale **to 8 GPUs per instance** with InfiniBand interconnects for low-latency communication. Billing starts at **€1.25/hour** for GPU instances with per-second granularity.

Storage services span three tiers:

- Ceph-backed block storage with hot-plug attach capabilities
- S3-compatible object storage integrated with CDN delivery
- NFS-based file shares with a **VAST Data backend** for GPU regions requiring high-performance parallel access.

VAST integration delivers sustained throughput of **40+ GBps per node** that AI training workloads demand.

Networking leverages OpenStack Neutron for virtual private clouds, load balancing (Octavia-backed L7 with mutual TLS), and security groups. Native IPv4/IPv6 dual-stack support combines with built-in L3/L4/L7

DDoS protection. Managed Kubernetes built on Cluster API provides **99.99% SLA** with GPU-enabled worker nodes, auto-scaling, and CSI driver support for NFS storage persistence.

Orchestration and multi-tenancy

Gcore AI Cloud Stack exposes a comprehensive REST API with permanent API tokens for programmatic access. A HashiCorp-verified **Terraform provider** enables declarative infrastructure management, including VMs, networks, Kubernetes clusters, load balancers, CDN resources, and DNS zones. Heat orchestration support enables stack-based deployments for complex topologies.

Multi-tenant isolation operates through project-based resource containers with five levels of role-based access control: Client Administrator, Project Administrator, User, Limited User, and Observer. Each role provides granular permissions over cloud resources, cost reports, audit logs, and quota management. Security certifications include **ISO/IEC 27001:2013, PCI DSS Level 1**, and GDPR compliance.

White-label and reseller capabilities enable partners to brand the entire platform—from CDN to AI infrastructure—with custom domains, local language support, and integration with external billing systems. The Gcore Admin Portal provides account management, product visibility configuration, and API access for reseller operations.

Nokia DCF provides automation-first networking

Nokia's Data Center Fabric solution centers on **SR Linux**, a network operating system designed with programmability in mind. The architecture comprises an unmodified Linux kernel, a modular operating system layer, and containerized applications, with each protocol (BGP, IS-IS, EVPN) running as an isolated process. Applications communicate via gRPC/protobuf using Nokia's Impart Database (IDB) for high-performance pub/sub messaging.

SR Linux delivers 100% YANG coverage

Every configuration and state element in SR Linux is modeled in YANG, enabling true model-driven management without translation layers. The CLI itself is derived from YANG models, ensuring consistency between interactive and programmatic interfaces. Management options include:

- **gNMI** (port 57400): Primary interface for streaming telemetry and configuration with TLS authentication
- **JSON-RPC** (ports 80/443): REST-like HTTP interface for web-based automation
- **NETCONF** (port 830): Standards-compliant XML management
- **SSH CLI:** Python-based advanced CLI with transaction mode and JSON/YAML outputs

The **NetOps Development Kit (NDK)** enables custom applications written in any gRPC-supported language to integrate deeply with system components, including the FIB, LLDP, and BFD. These applications appear as native SR Linux processes with full streaming telemetry support. This includes extension of the Yang extensions, to make sure configuration is integrated in the stack and directly provide same automation capability's ad native applications.

The 7220 IXR switch portfolio scales from 3.2 to 51.2 Tb/s

Nokia's 7220 IXR series provides purpose-built switches for leaf and spine roles

Model	Throughput	Ports	Buffer	Form factor	Role
IXR-D3L	3.2 Tb/s	32×100GE QSFP28 + 2×SFP+	32 MB	1U	Frontend/storage leaf
IXR-D4	6.0 Tb/s	28×100GE + 8×400GE QSFP-DD	82 MB	1U	Mixed leaf with 400G uplinks
IXR-D5	12.8 Tb/s	32×400GE QSFP-DD + 2×SFP+	132 MB	1U	High-density 400GE leaf/spine
IXR-H4	25.6 Tb/s	64×400GE QSFP-DD + 2×SFP+	113.5 MB	2U	Spine for large fabrics
IXR-H5-64O	51.2 Tb/s	64×800GE OSFP + 2×SFP+	165.2 MB	2U	AI-optimized spine and leave
IXS-A1	88 Gb/s	48xRJ45 + 4xSFP+		1U	Out-of-Band mgmt

The H5 series introduces **800GE native ports**, supports **DDR5 memory**, and TPM security modules. The upcoming **IXR-H6** support Linear Pluggable Optics (LPO) for power-efficient transceiver connectivity and extends capacity to **102.4 Tb/s** with 1.6TE interfaces for next-generation AI factories.

All platforms support extensive breakout options: 400GE ports break to 4×100GE or 2×200GE; 100GE ports break to 4×25GE or 4×10GE. Power supplies are 1+1 redundant with both AC and DC options; fan modules provide N+1 redundancy with front-to-back or back-to-front airflow configurations.

Event-Driven Automation enables intent-based operations

Nokia EDA is a Kubernetes-native automation platform that treats all network changes as events requiring reconciliation toward intended state. Key capabilities include:

- **Digital Twin/Sandbox:** Validate topology changes in a virtual environment before production deployment
- **Flexible Fabric Templates:** Auto-generate leaf-spine topologies from high-level parameters
- **Zero Touch Provisioning:** Bootstrap new nodes automatically upon connection
- **Continuous Reconciliation:** Monitor and correct drift from intended configuration
- **AIOps Integration:** GenAI-powered natural language queries and automated diagnostics

EDA integrates with external systems, including Netbox, VMware, OpenShift, OpenStack, and ITSM platforms via REST APIs. Deployment scales from single-node to 6+ node clusters running on hardened Talos Kubernetes with Rook-Ceph storage.

NVIDIA B300 Blackwell Ultra powers AI compute

The DGX B300 system represents NVIDIA's latest AI supercomputer, featuring **8× B300 GPUs** based on the Blackwell Ultra architecture. Each GPU contains 208 billion transistors across two dies connected via 10 TB/s NV-HBI, delivering up to **15 PFLOPS FP4** and **9 PFLOPS FP8** performance.

Per-GPU specifications

Specification	Value
Architecture	Blackwell Ultra (TSMC N4P)
Memory	288 GB HBM3e (12-high stacks)
Memory Bandwidth	8 TB/s
NVLink 5 Bandwidth	1.8 TB/s bidirectional
TDP	1,200W (HGX) / 1,400W (GB300)

System-level configuration

The DGX B300 ships in a **10U chassis** weighing up to 370 lbs with the following components:

- **CPUs:** 2× Intel Xeon Platinum 6776P (or AMD EPYC 9005 series for HGX)
- **System Memory:** 2 TB DDR5 standard, expandable to 4 TB
- **NVSwitch:** 2× 5th-generation providing 14.4 TB/s total switching capacity
- **Storage:** 8× 3.84 TB E1.S NVMe drives (cache) + 2× 1.92 TB M.2 (boot)
- **Power:** 12× 3.2 kW PSUs (N+N redundancy), 14.5 kW maximum draw

Network connectivity architecture

East-West (GPU-to-GPU via InfiniBand): Each B300 GPU connects to one port of the **ConnectX-8 SuperNIC,** providing **8× 800 Gb/s XDR InfiniBand** ports per server. The CX8 is the first NVIDIA adapter with PCIe Gen6, integrating the PCIe switch fabric to eliminate external switch components. Key features include SHARPv4 in-network computing for collective operations, adaptive routing, and hardware MPI acceleration. Total per-server InfiniBand bandwidth: **6.4 Tb/s.**

North-South (Frontend/Management via Ethernet): Two **BlueField-3 B3240 DPUs** provide **4× 400 Gb/s** Ethernet connectivity (1.6 Tb/s total per server). Each DPU has 2× 400GE ports, for a total of 4 ports per server for redundant leaf-switch connections. Each DPU integrates 16 Arm cores with 32 GB DDR for software-defined networking, storage, and East-West security (offload). The BlueField handles storage access, in-band management, and external connectivity while providing zero-trust security architecture via DOCA software.

BMC Management: On-board 1 Gb/s RJ-45 for out-of-band IPMI connectivity.

Power and cooling requirements

Configuration	Servers/Rack	Racks for 1 SU	Power per Rack	Total Power
Air-cooled	4	18	~56 kW	~1 MW
Liquid-cooled	8	9	~116 kW	~1 MW

Liquid cooling reduces data center cooling load by capturing **92% of heat** at the rack level via in-rack CDUs. DC busbar configurations using MGX racks support power shelves delivering up to 33 kW per shelf.

Reference architecture separates frontend and backend fabrics

The architecture employs a **dual-fabric design** where the Nokia DCF handles north-south Ethernet traffic (frontend network) while existing NVIDIA Quantum InfiniBand switches manage east-west GPU communication (backend network). This separation optimizes each fabric for its specific traffic patterns.

Frontend Ethernet fabric with Nokia switches

The frontend network serves management connectivity, storage access, external dataset ingestion, and AI inference serving. Traffic represents **15-20%** of total cluster bandwidth with lower latency sensitivity than the backend.

Topology: 2-tier leaf-spine using Nokia IXR switches with eBGP IPv6 unnumbered underlay and EVPN/VXLAN overlay.

Two deployment options based on server connectivity speed:

Option A: 400GE Server Connectivity (Recommended for maximum bandwidth)

Leaf layer (IXR-D5):

- 32 × 400GE QSFP-DD ports per switch
- DGX B300 nodes are connected via BlueField-3 DPUs at 2 × 400GE per server to a pair of redundant leaves
- 4 × IXR-D5 redundant leaf-pairs support 72 servers: ~18 x 400G server connections per leaf
  - Port allocation: 18 ports for servers + 14 ports for spine uplinks
- **Oversubscription:** 18:14 **≈ 4:3** (acceptable for frontend traffic, especially noting this is used redundantly and load-balanced All-Active)

Spine layer (IXR-H4):

- 2 × IXR-H4 spines provide sufficient capacity (64× 400GE each)
- A total of 104x400G ports per spine for leaf connections, leaving 24 ports available
- Remaining ports connect to exit routers and Gcore cloud infrastructure
- 2 spines provide N+1 redundancy with substantial headroom for expansion

Option B: 100GE Server Connectivity (Cost-optimized)

Leaf layer (IXR-D4):

- 28× 100GE QSFP28 + 8× 400GE QSFP-DD ports per switch
- DGX B300 nodes are connected via BlueField-3 DPUs at 2 × 100GE per server to a pair of redundant leaves
- 2 × IXR-D4 redundant leaf-pairs support 72 servers: ~36 servers per leaf pair
  - Port allocation: 36 x 100GE downlinks (servers) + 4 x 400GE uplinks (spines)
- **Oversubscription: ~2,25:1** for frontend traffic

Spine layer (IXR-H4):

- 2 × IXR-H4 spines provide sufficient capacity (64× 400GE each)
- A total of 16x400G ports per spine for leaf connections, leaving 48 ports available
- Remaining ports connect to exit routers and Gcore cloud infrastructure
- 2 spines provide N+1 redundancy with substantial headroom for expansion

Backend InfiniBand fabric remains NVIDIA Quantum

The backend network handles GPU-to-GPU communication during distributed training, requiring **sub-microsecond latency** and **zero packet loss.** Traffic represents **80-85%** of total cluster bandwidth.

Topology: Rail-optimized fat-tree using NVIDIA Quantum-X800 (Q3400) switches.

Rail-optimized design: Each GPU position (GPU 0-7) across all servers connects to the same pair of leaf switches (a "rail"). Intra-rail traffic is always one hop away; inter-rail traffic traverses the spine layer. This optimization reduces average hop count for collective operations.

Switch specifications (Q3400-RA):

- 144 ports at 800 Gb/s XDR InfiniBand
- 4U form factor
- SHARPv4 in-network computing
- Two-level fat-tree supports up to 10,368 NICs

EVPN/VXLAN overlay provides multi-tenancy

The overlay uses MP-BGP EVPN control plane with VXLAN data plane encapsulation:

- **Edge Routed Bridging (ERB):** All leaf switches perform L2 and L3 functions; spines provide IP transport only
- **Distributed Anycast Gateway:** Same gateway IP/MAC on all leaves for optimal first-hop routing
- **L2VNI per VLAN:** Extends broadcast domains across fabric
- **L3VNI per VRF:** Provides inter-subnet routing for tenant isolation
- **ARP suppression:** Reduces broadcast traffic via overlay control plane

VXLAN segments isolate:

- Storage network (VAST Data access)
- In-band management (cluster services, Slurm, Run:ai)
- Out-of-band management (BMC/IPMI)
- Tenant workload networks

Cloud network integrates via ML2 plugins

Nokia provides multiple integration pathways for connecting SR Linux switches to the network management service in Gcore AI Cloud Stack.

Nokia Connect ML2 plugin with Event-Driven Automation (EDA)

The preferred enterprise approach uses Nokia's **Connect ML2 plugin**, which integrates directly with EDA:

Workflow:

1. Neutron creates network segment → ML2 plugin creates matching BridgeDomain in EDA
2. Neutron creates VM port → ML2 plugin identifies compute node and configures sub-interfaces on connected switches
3. EDA pushes configuration to SR Linux fabric via gNMI
4. Streaming telemetry confirms state reconciliation

Storage network design supports VAST Data integration

The architecture dedicates network infrastructure for high-performance storage access, separating storage traffic from management and compute fabrics.

VAST Data deployment pattern

VAST Data's DASE (Distributed And Shared Everything) architecture requires:

- **Backend network:** Single L2 VLAN across NIC0 interfaces for NVMe over RDMA between storage nodes
- **Frontend network:** NFS/S3/SMB access from GPU servers via Nokia leaf switches
- **Bandwidth:** Minimum 40 GBps per DGX node (80 GBps for B300)

Network configuration:

- Standalone interfaces (no port channels on backend)
- MTU 9216 (jumbo frames)
- DSCP marking 26 for QoS prioritization
- Lossless Ethernet via PFC and ECN for RoCEv2

Storage fabric topology

Dedicated InfiniBand option (NVIDIA pattern):

- MQM9700-NS2F switches for storage fabric
- 1:1 port-to-uplink ratio for storage devices
- DGX connections at 4:3 oversubscription

Converged Ethernet option (Nokia pattern):

- Storage traffic on dedicated VLANs across Nokia leaf-spine
- DCB (Data Center Bridging) enabled for lossless behavior
- DCQCN congestion control for RoCEv2

Hardware bill of materials for one Scalable Unit

A single Scalable Unit (SU) comprises 72 B300 servers delivering 576 GPUs with 166 TB aggregate HBM3e memory. The following BOM covers the complete infrastructure stack.

GPU rack components

Air-cooled configuration (18 racks):

Component	Quantity per Rack	Total (18 racks)
DGX B300 servers	4	72
In-rack power shelves	2	36
Per-rack power	~56 kW	~1 MW total

Liquid-cooled configuration (9 racks):

Component	Quantity per Rack	Total (9 racks)
DGX B300 servers	8	72
In-rack CDU	1	9
Per-rack power	~116 kW	~1 MW total

Liquid cooling enables 2× server density per rack while reducing data center cooling requirements by capturing 92% of heat at the rack level.

Frontend Ethernet fabric (Nokia)

Option A: 400GE connectivity (recommended)

Component	Role	Quantity	Rationale
IXR-D5	Server leaf	8	18×400GE downlinks + 14x400GE uplinks = 18 servers at 400G per leaf-pair with dual-homing
IXR-H4	Spine	2	N+1 redundancy, 52 ports per spine for leaves + headroom

Note: at higher SU counts, 12 uplinks are used on the server leaf to allow better scaling to higher number of spines

Port allocation per IXR-D5 leaf: ~18 server ports (400GE) + 14 spine uplinks (400GE)

Option B: 100GE connectivity (cost-optimized)

Component	Role	Quantity	Rationale
IXR-D4	Server leaf	4	36×100GE downlinks (26xQSFP28 and 2 QDD ports break out) + 4x400GE uplinks =18 servers per leaf-pair with dual-homing
IXR-H4	Spine	2	Full mesh connectivity to all leaves

Port allocation per IXR-D4 leaf: 36 server ports (100GE) + 4 × 400GE uplinks (spines)

Backend InfiniBand fabric (NVIDIA)

Component	Role	Quantity	Rationale
Q3400-RA	Compute leaf	8	1 per rail (GPU position)
Q3400-RA	Compute spine	4	Non-blocking fabric
MQM9700-NS2F	Storage leaf	2	Optional dedicated storage fabric
MQM9700-NS2F	Storage spine	2	Optional dedicated storage fabric

Rail-optimized topology: Each Q3400 leaf connects 72 GPUs (one from each server at same position).

Cloud infrastructure rack

Component	Quantity	Role
Cloud controllers	3	HA control plane (Nova, Neutron, Keystone, Glance, Cinder)
Network nodes	2	Neutron agents, load balancers
Ceph OSD nodes	6	Distributed block storage (minimum 3 for replication)
Ceph MON/MGR nodes	3	Cluster management
Management servers	2	Base Command Manager, Slurm, UFM

Storage leaf connectivity: IXR-D3L with 25GE to each Ceph node, 100GE to VAST Data.

Management and core

Component	Quantity	Role
IXR-D3L	2	Out-of-band management ToR
7750 SR-1	2	Core/exit routers (WAN, Internet)

OOB network: Dedicated VLAN for BMC/IPMI, switch management, PDU monitoring.

Scaling extends from 1 SU to enterprise deployments

The architecture scales linearly by adding Scalable Units, with the Nokia fabric extending through additional spine switches or a super-spine layer.

Scaling patterns

With 400GE connectivity (IXR-D5 leaves):

Configuration	GPU Servers	GPUs	Racks (Air)	Frontend Leaves	Spines
1 SU	72	576	18	8× IXR-D5	2× IXR-H4
2 SU	144	1,152	36	16× IXR-D5	4× IXR-H4
4 SU	288	2,304	72	32× IXR-D5	12× IXR-H4
8 SU	576	4,608	144	64× IXR-D5	12× IXR-H4

Beyond 8 SU, it is advised to use IXR-H5 or IXR-H6 as spine-model as this will further enable growth in a two tier configuration. Alternatively, a multi-POD design could be deployed using a super-spine concept.

With 100GE connectivity (IXR-D4 leaves):

Configuration	GPU Servers	GPUs	Racks (Air)	Frontend Leaves	Spines
1 SU	72	576	18	4× IXR-D4	2× IXR-H4
2 SU	144	1,152	36	8× IXR-D4	2× IXR-H4
4 SU	288	2,304	72	16× IXR-D4	2× IXR-H4
8 SU	576	4,608	144	32× IXR-D4	4× IXR-H4

Network expansion principles

- **Add leaves:** Requires connection to all existing spines (maintains oversubscription ratio)
- **Add spines:** Requires connection to all existing leaves (increases bisection bandwidth)
- **Pre-cable fabric:** Leave unused rack positions cabled to enable hot expansion

The InfiniBand backend scales similarly, with Q3400 spines supporting up to 10,368 ConnectX-8 NICs in a two-level fat-tree topology.

Conclusion

This reference architecture delivers a complete foundation for enterprise AI infrastructure by combining Gcore's proven OpenStack-based cloud orchestration with Nokia's automation-first data center networking. The dual-fabric approach optimizes each network for its specific traffic patterns: **6.4 Tb/s InfiniBand** per server for GPU communication and **Nokia IXR switches** for management, storage, and external connectivity.

Key architectural specifications for 1 Scalable Unit:

Component	Recommended (400GE)	Cost-Optimized (100GE)
GPU Servers	72× DGX B300 (576 GPUs)	72× DGX B300 (576 GPUs)
Racks	18 (air) or 9 (liquid)	18 (air) or 9 (liquid)
Nokia Spine	2× IXR-H4	2× IXR-H4
Nokia Leaf	8× IXR-D5	4× IXR-D4
Server Ethernet	2× 400GE (800 Gb/s)	2× 100GE (200 Gb/s)
IB Spine	4× Q3400	4× Q3400
IB Rails	8× Q3400	8× Q3400
Server InfiniBand	8× 800G (6.4 Tb/s)	8× 800G (6.4 Tb/s)

Key architectural decisions include eBGP IPv6 unnumbered underlay for simplified automation, EVPN/VXLAN overlay for multi-tenant isolation, and rail-optimized InfiniBand topology for efficient collective operations. The infrastructure-as-code approach using Terraform, Ansible, and containerlab enables repeatable deployments from single Scalable Units to enterprise-scale AI factories.

Organizations adopting this architecture gain the operational benefits of Gcore's multi-tenant cloud platform—including white-label capabilities, per-minute billing, and 99.99% SLA Kubernetes—combined with Nokia's model-driven network automation and streaming telemetry. The result is an AI infrastructure that scales predictably while maintaining the operational simplicity that enterprise IT teams require.

Nokia OYJ
Karakaari 7
02610 Espoo
Finland
Tel. +358 (0) 10 44 88 000

CID:215349