

Sovereign AI Data Center: the nine building blocks

Everything stays in-country, under your control.

01 Cooling

See 1.1–1.11

Direct liquid cooling: cold plates and CDUs, plus rear-door heat exchangers. ~130 kW per GB200 NVL72 rack vs 30–40 kW for air.

02 GPU compute

See 2.1–2.10

Training: NVIDIA Blackwell HGX B200/B300, GB200/GB300 NVL72. Inference: RTX PRO / L40S-class servers, plus edge nodes.

03 Networking and storage

InfiniBand or Spectrum-X fabric, BlueField DPUs, parallel storage with GPUDirect Storage.

04 Data platform

On-prem lakehouse (Iceberg on object storage), streaming ingest, catalog and lineage, vector database for RAG.

05 Platform software

Red Hat OpenShift and OpenShift AI, NVIDIA AI Enterprise (NIM, NeMo), air-gapped registry.

06 Models

Open-weight models on-prem as NIM (Nemotron, Llama, Mistral), fine-tuned for Bangla and mission data.

07 Security

Zero trust, HSM with in-country keys, data diodes between classifications, SOC/SIEM, signed images and SBOM.

08 Governance

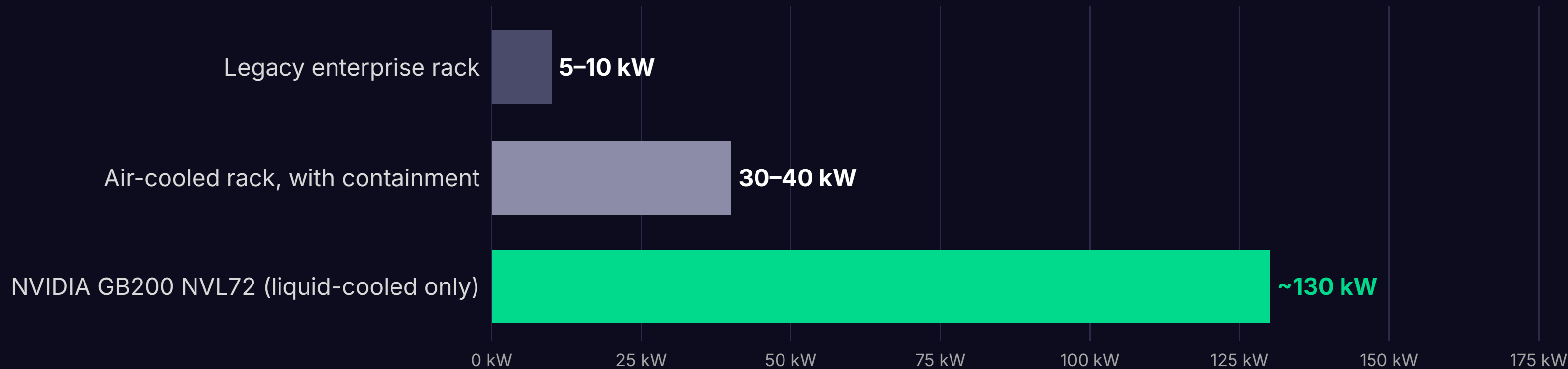
AI policy and ethics, model risk and evaluation, data classification, audit trail, lifecycle refresh.

09 Use Cases

See 9.1–9.4

e.g., intelligence analysis, document and translation assistants, ISR imagery analysis, cybersecurity, logistics and maintenance.

POWER PER RACK, BY RACK TYPE



Sources: CompTech, Liquid cooling vs air cooling (2026); NVIDIA GB200 NVL72 product page. Ranges are typical values.

→ WHY COOLING MATTERS

Cooling decides the AI data center

The next GPU purchase forces the cooling decision.

~130 kW

Power drawn by one NVIDIA GB200 NVL72 rack, which ships liquid-cooled only. Air cooling tops out near 30 to 40 kW per rack, so the heat has to leave as liquid.

→ AI SERVER POWER

How much power AI servers need

A 40 KW AIR RACK HOLDS ONLY TWO DGX B200 OR B300 SERVERS. ONE NVL72 RACK DRAWS OVER 3X THAT, ON LIQUID.

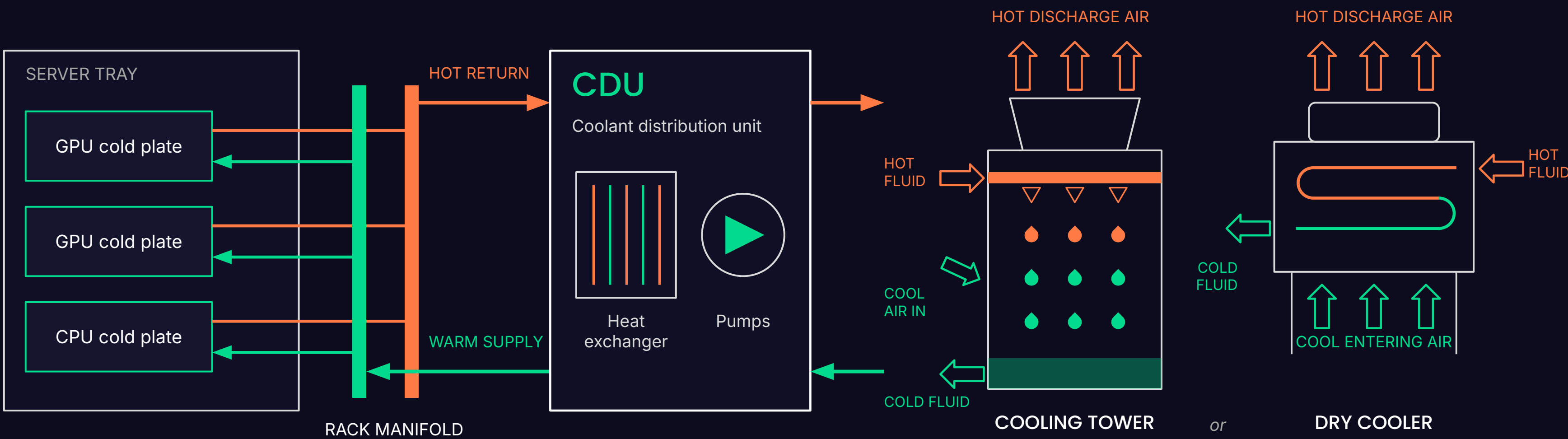
System	GPUs	Max power	Cooling	Fit in a 40 kW air rack
NVIDIA DGX H100	8x H100	10.2 kW per server	Air	3 servers (30.6 kW)
NVIDIA DGX H200	8x H200	10.2 kW per server	Air	3 servers (30.6 kW)
NVIDIA DGX B200	8x B200	~14.3 kW per server	Air	2 servers (28.6 kW)
NVIDIA DGX B300	8x B300 (Blackwell Ultra)	~14 kW per server	Air	2 servers (28 kW)
Supermicro SYS-522GA-NRT (RTX PRO 6000 Server Edition)	8x RTX PRO 6000	9.1 kW per server	Air	4 servers (36.4 kW)
NVIDIA GB200 NVL72 (Supermicro rack)	72x B200	125–135 kW per rack	Liquid only	None: liquid-cooled rack only
NVIDIA GB300 NVL72 (Supermicro rack)	72x B300	132–140 kW per rack	Liquid only	None: liquid-cooled rack only

Max system power from vendor spec sheets; actual draw varies by workload. Rack fit = servers whose combined max power stays at or under 40 kW.

→ DIRECT LIQUID COOLING

How direct liquid cooling works

WARM SUPPLY (MINT) FLOWS TO THE CHIPS. HOT RETURN (ORANGE) CARRIES THE HEAT OUT.



SECONDARY LOOP: sealed, treated coolant inside the rack

PRIMARY LOOP: facility water to outdoor heat rejection

Cold plates capture most of the server heat at the chip, about 70 to 75 percent of the rack load. Memory, drives and power supplies shed the rest into room air, so a smaller air system stays. A chiller is only a trim device for the hottest days.

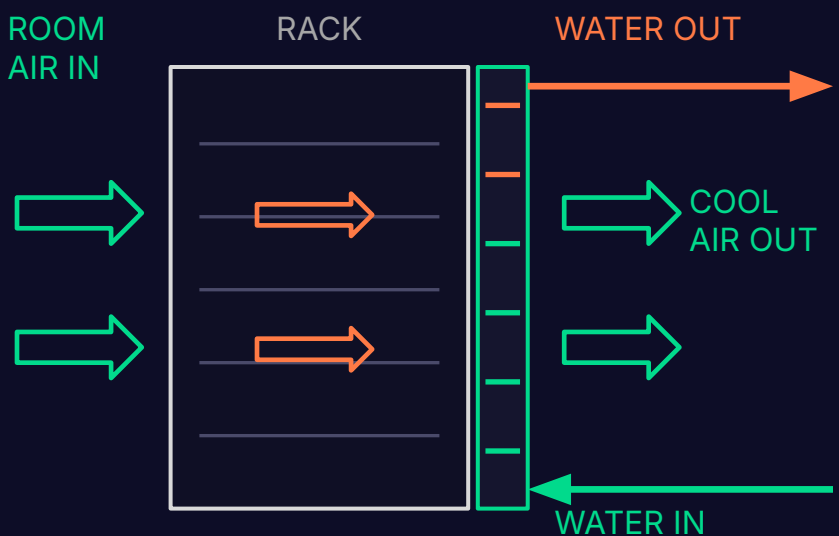
→ OTHER LIQUID OPTIONS

Rear doors and immersion

Two more ways to move heat into liquid. A rear door retrofits racks you already own; immersion suits new, purpose-built halls.

Rear-door heat exchanger

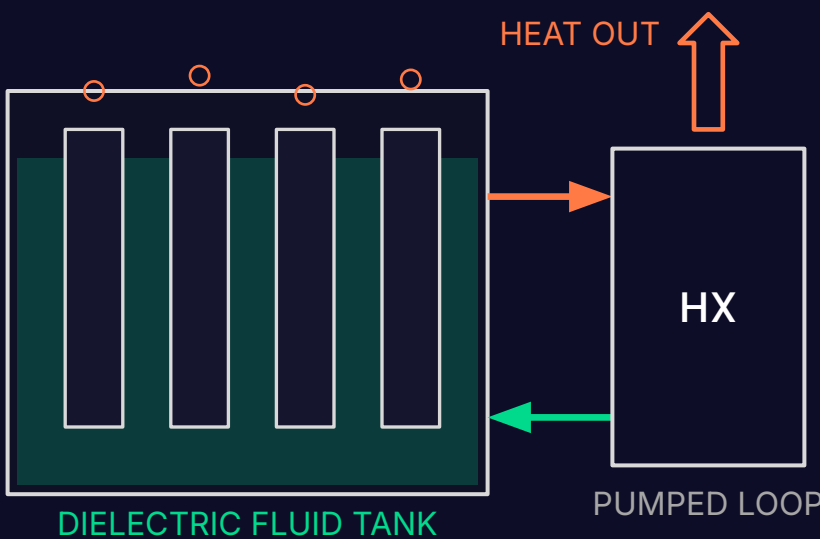
BOLTS ONTO EXISTING RACKS



A water coil in the rear door absorbs heat from server exhaust before it enters the room. Passive doors rely on server fans; active doors add their own. About 50 to 70 kW per rack.

Immersion cooling

SERVERS SUBMERGED IN FLUID



Servers sit in a tank of dielectric fluid. Single-phase: the fluid is pumped to a heat exchanger. Two-phase: it boils on the chips and condenses on a coil in the lid.

→ OPTIONS COMPARED

Four cooling options compared

DENSITY FROM PUBLISHED INDUSTRY AND VENDOR RANGES. RELATIVE COST IS INDICATIVE CAPEX, NOT A QUOTE.

Option	Density per rack	Retrofit effort	Relative cost	Best fit
Air (containment)	10–40 kW	None: native to the hall	Low	General IT, storage and network racks under ~20 kW
Rear-door HX	Up to ~50–70 kW	Low: bolts onto existing racks	Medium	Hot racks in an existing air-cooled hall
Direct liquid (cold plate)	80–250+ kW	Medium: rack by rack, adds CDUs	Medium to high	GPU racks such as GB200 NVL72; new AI halls
Immersion	80–250+ kW	High: difficult in an existing hall	High	Greenfield, edge and HPC sites

→ COST AND TCO

Where liquid saves money

Liquid cooling costs more per rack to install, then saves on energy, space and plant every year the hall runs.

01

ENERGY AND PUE

Typical PUE: air 1.3–1.8, direct liquid 1.05–1.25, immersion 1.03–1.10. Global average: 1.54.

02

FLOOR SPACE

About 130 kW per liquid rack versus 30–40 kW per air rack: fewer racks and a smaller hall.

03

CHILLERS

Water at 30–45 °C is cooled by dry coolers most of the year. The chiller becomes a trim device.

04

GENERATOR AND UPS

Less cooling power means less load to back up. A Vertiv–NVIDIA study cut total facility power by 10.2%.

Figures are industry/vendor ranges; size your own TCO in the design phase.

→ RETROFIT PATH

Four steps from air to liquid

A phased path that keeps the hall running. Most operators add liquid to an existing air-cooled hall one suite or corner at a time.

STEP 1 · AIR HALL

Contain hot and cold aisles and add monitoring. Air carries about 30–40 kW per rack.

STEP 2 · REAR DOORS

Bolt water-coil rear doors onto the hottest racks: up to about 50–70 kW, servers unchanged.

STEP 3 · GPU POD

Add a liquid-cooled GPU pod: cold-plate racks, in-row CDUs and a facility water loop.

STEP 4 · LIQUID HALL

Warm-water liquid becomes the default, with dry coolers. Air and chillers only trim.

→ BANGLADESH FIT AND RESILIENCE

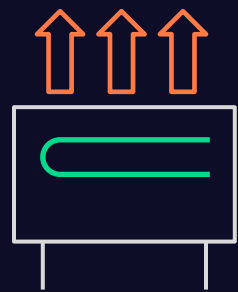
Built for Bangladesh

Dhaka is hot and humid year round, grid power is constrained and water is not free. Warm-water liquid cooling answers all three.

1

BUILT FOR DHAKA'S HEAT AND HUMIDITY

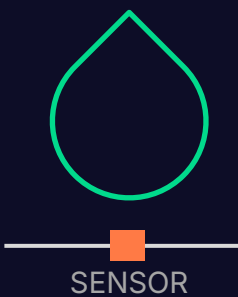
Warm water at 30–45 °C lets dry coolers run compressor-free most of the year; a trim chiller covers the hottest afternoons. Dry coolers use little or no water, unlike cooling towers.



2

LEAK-SAFE BY DESIGN

Leak detection in the CDU, piping and racks; dripless quick-disconnects on every server tray; small coolant volumes and drip pans under fittings. Pressure-tested at commissioning.



3

REDUNDANT AND MONITORED

N+1 CDUs with dual hot-swap pumps and redundant power; flow, temperature and leak alarms feed DCIM/BMS. Keep CDU pumps on UPS and generator power.



SUPPLY CHAIN

Who makes liquid cooling

COLD PLATES At the chip ✓ CoolIT Systems ✓ Boyd ✓ Motivair (Schneider) ✓ JetCool	CDUs Loop control ✓ Vertiv (Liebert XDU) ✓ Motivair (Schneider) ✓ CoolIT Systems ✓ nVent ✓ Delta ✓ LiquidStack	REAR-DOOR HX Retrofit ✓ Vertiv ✓ Motivair ChilledDoor ✓ nVent ✓ USystems	IMMERSION Submerged ✓ GRC ✓ Submer ✓ Asperitas	SERVERS AND RACKS NVIDIA NVL72 ✓ Dell (XE9712) ✓ HPE ✓ Supermicro
--	--	--	--	---

Each vendor checked on its own website for that product category, October 2026. Names only, not an endorsement.

REAL-WORLD EXAMPLES

Liquid cooling in the field



COOLANT PIPING

Frontier, Oak Ridge: coolant supply and return run overhead

Photo: Oak Ridge National Laboratory, CC BY 2.0



GB200 NVL72 RACK

Liquid-cooled trays beside a liquid-to-air sidecar

Photo: Geekerwan, CC BY 3.0, cropped



IMMERSION TANK

Lifting a server out of a Submer SmartPodX tank

Photo: Submer Immersion Cooling, CC BY-SA 4.0



DIRECT LIQUID COOLING

JUPITER, Jülich: BullSequana XH3000 racks

Photo: FZ Jülich / Sascha Kreklau, CC BY 4.0

PRODUCT CLOSE-UPS

Rear doors and cold plates up close



COOLLOOP RDHx ACTIVE

Vertiv CoolLoop RDHx DCD35 rear door with fan module

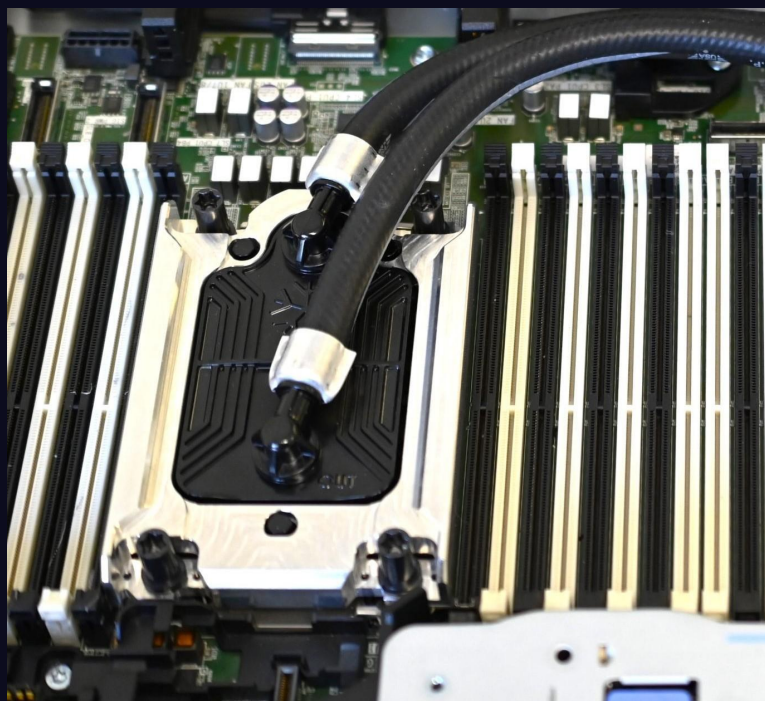
Image: Vertiv (product page)



VERTIV COOLLOOP RDHx

Chilled-water rear door, up to 80 kW per rack

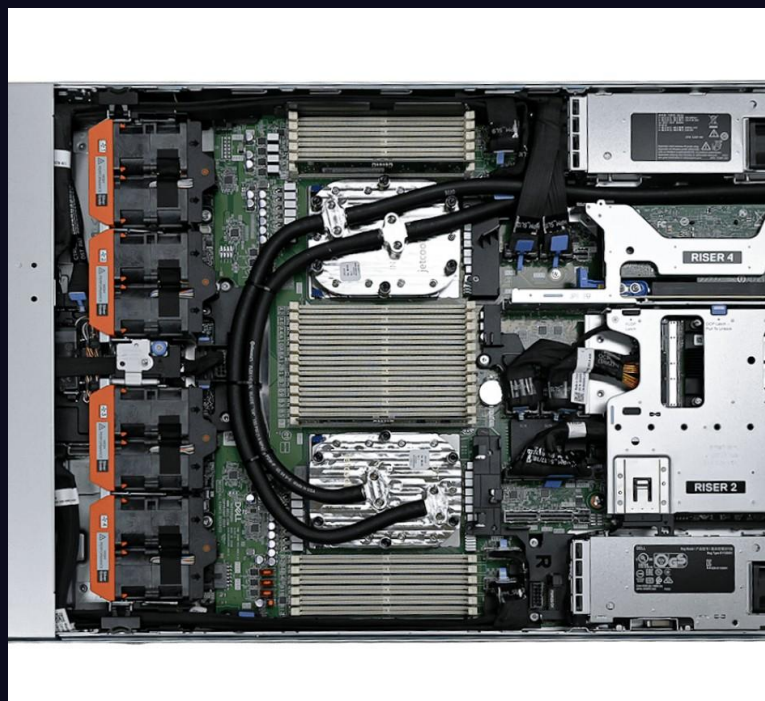
Image: Vertiv (newsroom)



JETCOOL SMARTPLATE

Cold plate on a CPU, coolant hoses clipped above

Image: JetCool Technologies



SMARTPLATE, TOP VIEW

Dual AMD EPYC server, both CPUs on cold plates

Image: JetCool Technologies

→ GPU SELECTION

Which GPU for which job

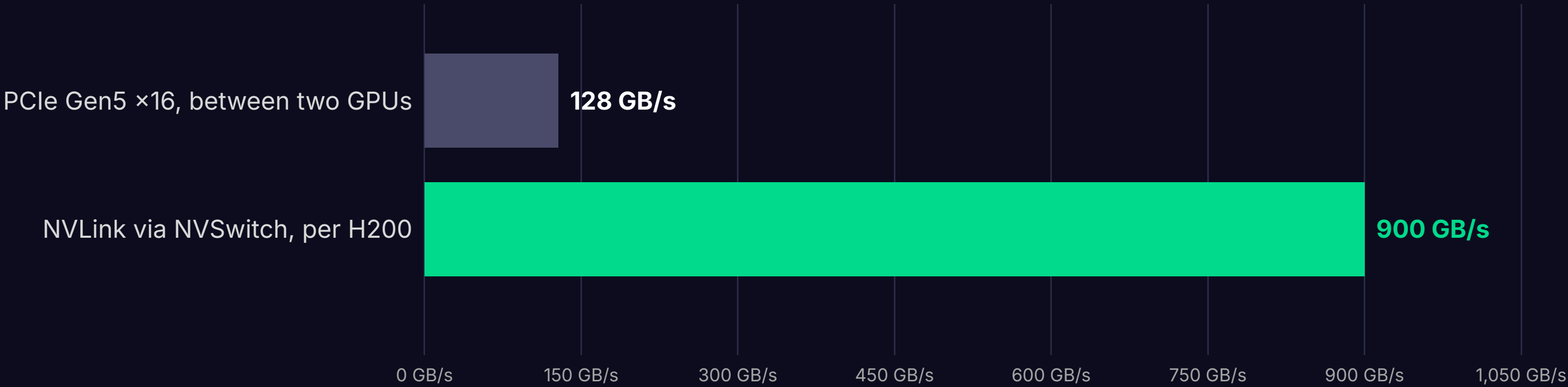
FIVE NVIDIA DATA CENTER GPUS, FROM AIR-COOLED PCIE CARDS TO LIQUID-READY HGX SYSTEMS

GPU	Architecture	Memory	Bandwidth	FP8	FP4	Max power	Form factor	Best for
L40S	Ada Lovelace	48 GB GDDR6	864 GB/s	1,466 TFLOPS	—	350 W	PCIe, air	Inference, vision & graphics, small-model fine-tuning
H200 (SXM / NVL)	Hopper	141 GB HBM3e	4.8 TB/s	3,958 / 3,341 TFLOPS	—	700 W / 600 W	SXM (HGX) / PCIe	Large-model inference and training
B200	Blackwell	180 GB HBM3E	7.7 TB/s	9 PFLOPS	18 PFLOPS	1,000 W	SXM, HGX 8-GPU	Frontier training, high-throughput inference
B300	Blackwell Ultra	Up to 288 GB HBM3E	8 TB/s	9 PFLOPS‡	18 PFLOPS‡ (13.5 dense)	Up to 1,400 W	SXM, HGX 8-GPU	Frontier training, reasoning inference at scale
RTX PRO 6000 Server Ed.	Blackwell	96 GB GDDR7	1,597 GB/s	2 PFLOPS	4 PFLOPS	Up to 600 W	PCIe, air or liquid	Enterprise inference, visual computing, agentic AI in standard servers

Takeaway: Start inference on air-cooled PCIe servers (RTX PRO 6000, L40S); keep HGX B200/B300 for training and large-scale serving, with liquid-ready power.

Peak tensor performance with sparsity. ‡ B300 per-GPU = NVIDIA HGX B300 8-GPU figures ÷ 8. Sources: nvidia.com product pages and datasheets.

GPU-TO-GPU BANDWIDTH, BY LINK (BOTH DIRECTIONS)



Sources: CompTech proposal deck, pp. 9–10; NVIDIA H200 product page. Figures are total bandwidth, both directions.

→ WHY NVLINK MATTERS

NVLink is ~7× PCIe for large models

Our rule: serve in FP8, fine-tune in BF16, and keep an FP16 copy of each model.

~7×

NVLink via NVSwitch gives each H200 900 GB/s to every other GPU in an HGX node, against 128 GB/s over PCIe Gen5. Models of 70B and up, long contexts and fine-tuning all cross that link.

→ MODEL FIT

Which models run on which GPU

GPU COUNT AND PRECISION FOR EACH MODEL CLASS. H200 IS THE RECOMMENDED CARD.

Model class	L40S · 48 GB	RTX PRO 6000 · 96 GB, PCIe	H200 · 141 GB · recommended
Small, 8 to 12B Llama 3.1 8B, Nemotron Nano	1 GPU, FP16 or FP8	1 GPU, limited by memory bandwidth	1 GPU, hundreds of users
Mid, about 49B Llama Nemotron Super	2 GPUs, FP8	1 GPU, FP8, few users	1 GPU, FP16, full quality
Large, 70B Llama 3.3 70B	4 GPUs, FP8	2 GPUs over PCIe, slow at scale	1 GPU FP8, or 2 GPUs FP16
Frontier, 253B to 405B Nemotron Ultra, Llama 3.1 405B	Not recommended	Not practical without NVLink	4 to 8 GPUs with NVLink, FP8
Embedding, reranking, speech, vision NeMo Retriever, Riva, vision NIM	1 GPU serves several	Over-specified; use L40S	Over-specified; use L40S

Rule of thumb: model weights in GB $\approx$ parameters in billions $\times$ 2 for FP16 and $\times$ 1 for FP8. Add 20 to 30% for the KV cache and runtime, more for long contexts and many simultaneous users.

→ PHASED GROWTH

Start with one H200, grow to two HGX servers

Each step adds H200 GPUs on the same software, so the design never changes. H200 is the phase-one entry; HGX B200/B300 (slide 2.1) is the later scale-up for frontier training.



→ PRE-VALIDATED AI STACKS

Cisco AI PODs and Dell AI Factory with NVIDIA

Two full-stack designs that pair compute, network, storage and software into one validated, vendor-supported AI stack.

Cisco AI PODs

VALIDATED DESIGN + NVIDIA ERA

COMPUTE

UCS C885A M8 (8RU HGX, training); UCS C845A M8 (4RU MGX, 2–8 PCIe GPUs); X-Series X210c/X215c; C240/C245 M8

NETWORK

Nexus 9000, 400G/800G RoCEv2 lossless

STORAGE

VAST, NetApp AFF, Pure FlashArray, Nutanix

SOFTWARE

NVIDIA AI Enterprise, Red Hat OpenShift, RHEL, Ubuntu, Rancher

OPS AND SECURITY

Intersight, Nexus Dashboard; AI Defense, Hypershield, Isovalent

SIZING

Scale units of 32/64/128+ GPUs; POD1 inference/edge, POD2 training/fine-tuning

Dell AI Factory with NVIDIA

FULL STACK, SOVEREIGN-READY

COMPUTE

PowerEdge XE9780 (8x HGX B200/B300, air) or XE9780L (liquid); XE7740/XE7745 (up to 8x RTX PRO 6000 or H200 NVL); XE9680; R770/R7725/R7715 for light inference

NETWORK

NVIDIA Spectrum-X SN5610/SN2201 + Dell PowerSwitch

STORAGE

PowerScale F710

SOFTWARE

NVIDIA AI Enterprise 8.0, OpenShift AI 3.3, NIM, Run:ai, Ubuntu 24.04

OPS

Dell Automation Platform 2.0; PowerRack rack-scale

SERVICES

Advisory, ProDeploy, ProSupport; modular from pilot to training

→ SIDE BY SIDE

Cisco AI PODs vs Dell AI Factory

FROM VENDOR DATA SHEETS AND REFERENCE ARCHITECTURES. DEPLOYMENT SPEED FIGURES ARE VENDOR CLAIMS.

Attribute	Cisco AI PODs	Dell AI Factory with NVIDIA
Compute	UCS C885A M8 (8RU HGX), C845A M8 (4RU MGX), X-Series, C240/C245 M8	PowerEdge XE9780/XE9780L, XE7740/XE7745, XE9680, R770/R7725/R7715
GPUs	H100, H200, RTX PRO 6000, L40S (AI PODs data sheet); C885A also offered with AMD MI350X	8x HGX B200/B300 (XE9780); up to 8x RTX PRO 6000 or H200 NVL (XE7740/XE7745)
Network	Nexus 9000, 400G/800G RoCEv2 lossless	NVIDIA Spectrum-X SN5610/SN2201 + Dell PowerSwitch
Storage	Partners: VAST, NetApp AFF, Pure FlashArray, Nutanix	Dell PowerScale F710
Management	Intersight + Nexus Dashboard	Dell Automation Platform 2.0
Software	NVIDIA AI Enterprise, Red Hat OpenShift, RHEL, Ubuntu, Rancher	NVIDIA AI Enterprise 8.0, OpenShift AI 3.3, NIM, Run:ai, Ubuntu 24.04
Sizing	Scale units of 32/64/128+ GPUs; POD1 inference/edge, POD2 training/fine-tuning	Modular foundations from pilot to training; PowerRack rack-scale
Cooling	Air-cooled (C845A); liquid option not published	Air (XE9780) or liquid (XE9780L)
Deployment speed	Vendor claim: up to 50% faster deployment	Vendor claim: ~1 week faster software deployment

For a sovereign defense build, both give one-vendor support, a validated design and faster deployment. [CompTech integrates either stack with OpenShift and NVIDIA AI Enterprise.](#)

PRODUCT CLOSE-UPS

Cisco and Dell AI hardware up close

CISCO



UCS C885A M8

8RU server built on NVIDIA HGX for AI training

Image: Cisco



UCS C845A M8

4RU NVIDIA MGX server with 2 to 8 PCIe GPUs

Image: Cisco



NEXUS 9364E-SG2

2RU switch: 64 × 800GbE, 51.2 Tbps, Cisco Silicon One

Image: Cisco

DELL TECHNOLOGIES



POWEREDGE XE9780

Air-cooled, 8 GPUs on NVIDIA HGX B300 or B200

Image: Dell Technologies



POWEREDGE XE7745

4U air-cooled, up to 8 double-wide or 16 single-wide PCIe GPUs

Image: Dell Technologies



POWERSCALE F710

1U all-flash NVMe NAS node with 10 drives

Image: Dell Technologies

■ 2.8 · VALIDATED INFERENCE STACK

Inside a Cisco AI POD

for inferencing

Cisco Validated Design (CVD)-based solutions for Edge Inference, RAG and Large-Scale Inferencing, with centralized management and automation.

- ✓ **The problem:** infrastructure needs vary significantly with the type and size of the AI model; point AI solutions are costly and complex.
- ✓ **Right-sized:** size the investment for today, with a view to future scale.
- ✓ **Validated:** CVD-based and performance tested; linear scalability in benchmark tests on real-life model simulation.
- ✓ **Scalable:** independent scalability at each infrastructure layer, for DC or Edge AI deployments.
- ✓ **Four configurations** that vary the amount of CPU and GPUs.
- ✓ **Stack:** AI model · NVIDIA AI Enterprise · Red Hat OpenShift · UCS X-Series · Nexus · Pure Storage / NetApp · Intersight, Ansible, Terraform, APIs · Mint Partners, Cisco CX.

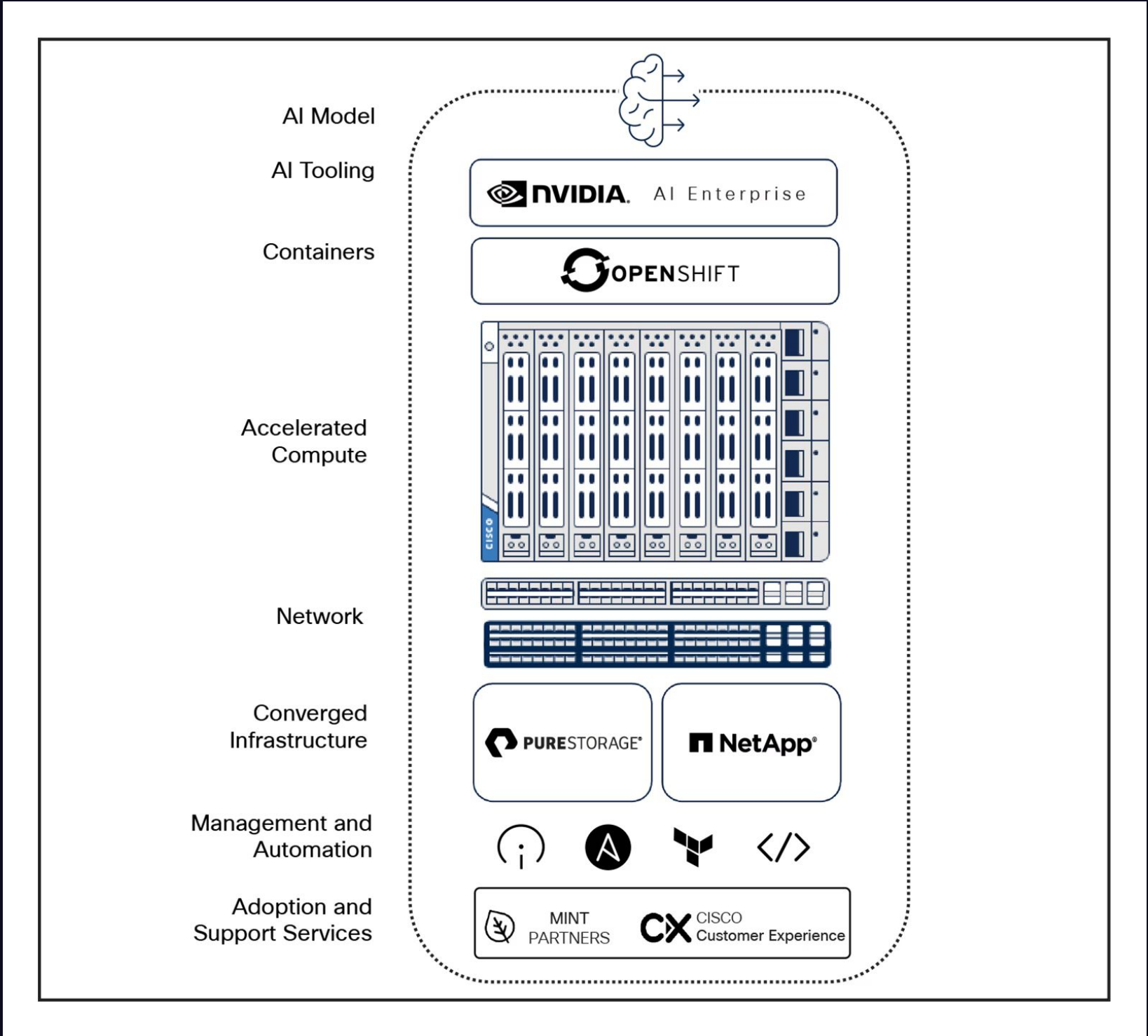


Image: Cisco

What every Cisco AI POD contains

Regardless of the configuration, they all contain:

COMPUTE	FABRIC	MANAGEMENT	SOFTWARE	STORAGE
UCS X-Series	Connectivity	Operations	AI stack	Optional
✓ Cisco UCS X9508 Chassis ✓ UCS X-Series M7 Compute Nodes ✓ UCS X440p PCIe Node with NVIDIA GPUs	✓ UCS 9108 Intelligent Fabric Modules ✓ UCS 6536 Fabric Interconnect or UCS Fabric Interconnect 9108 100G ✓ UCS X9416 X-Fabric Modules	✓ Cisco Intersight ✓ Cisco Services	✓ NVIDIA AI Enterprise subscription ✓ Red Hat OpenShift licensing	✓ NetApp (FlexPod) ✓ Pure Storage (FlashStack) ✓ DataOps toolkits for developers and data scientists

BENEFITS

AI-ready infrastructure with performance assurance and seamless scalability

Faster time to production-ready inferencing through full-stack validation of infrastructure, software and AI toolsets

Single support for the AI deployment architecture: CVDs backed by Cisco TAC

Inside the Dell AI Factory with NVIDIA

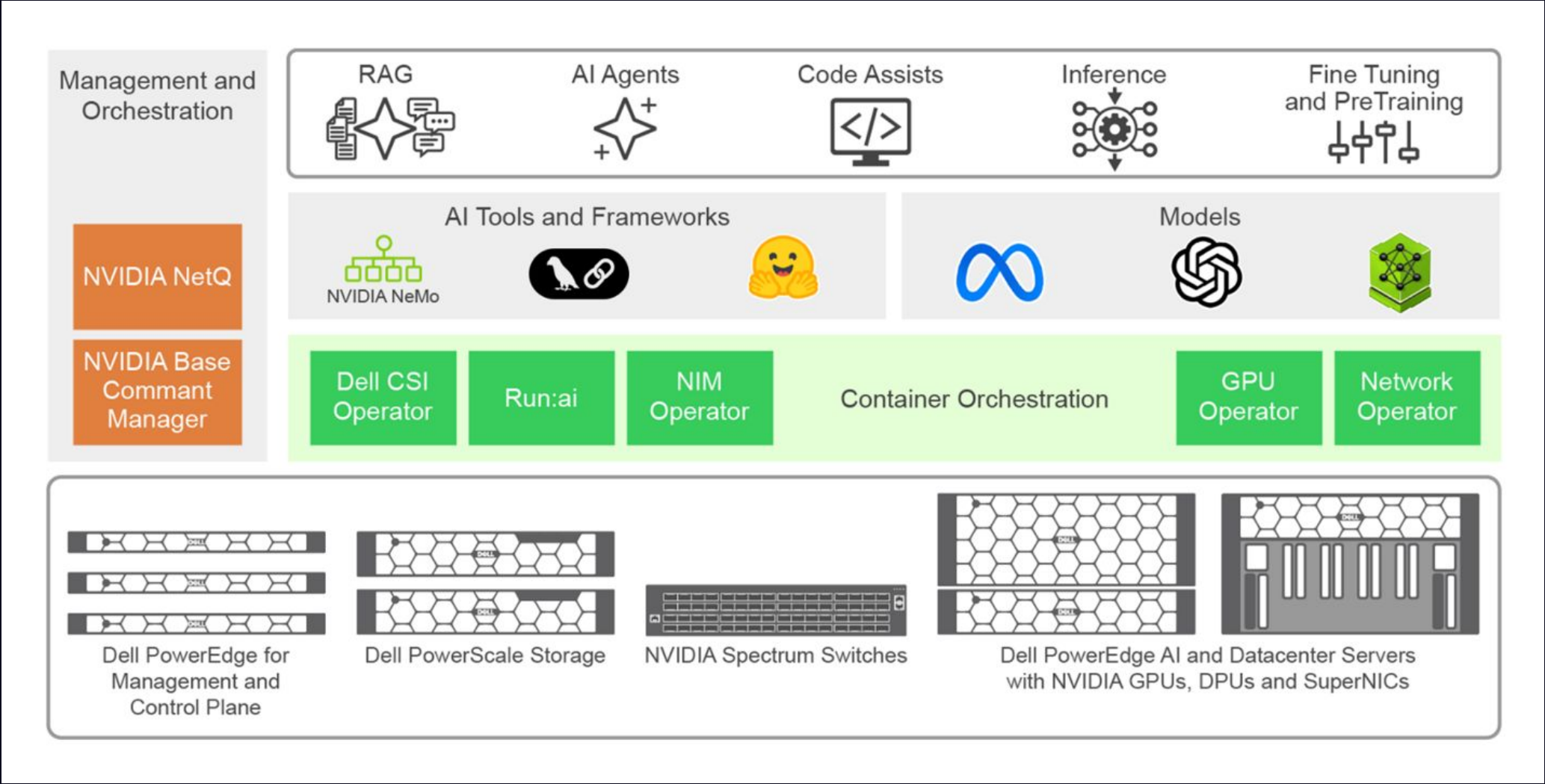


Image: Dell Technologies

UNIFIED PLATFORM

Compute, networking, storage, orchestration and AI software in one platform for enterprise generative AI: RAG, AI agents, code assistants, fine-tuning and inference.

COMPUTE

High-density AI-optimized servers for training and fine-tuning. Data-center-class servers as a cost-efficient option for inference only. Management and control plane on separate general-purpose servers.

NETWORK

NVIDIA Spectrum-X (Spectrum-4 switches and SuperNICs): one fabric for frontend (EVPN-VXLAN) and backend GPU traffic. NVIDIA Cumulus on all switches.

OPERATIONS

NVIDIA Base Command Manager Essentials: provisioning, bring-up, lifecycle. NetQ: network visibility and troubleshooting. Kubernetes with GPU, Network, Dell CSI and NIM Operators, plus Run:ai.

→ USE CASE MAP

16 use cases across eight mission areas

Two use cases in each area, all on the same in-country platform. Detail on 9.2 and 9.3.

01 Surveillance and ISR

Imagery analysis of drone and satellite images ·
Video search and summaries.

02 Border and maritime

Border monitoring alerts from camera, radar and
sensors · Maritime domain awareness.

03 Logistics and maintenance

Predictive maintenance of vehicles, aircraft and
vessels · Supply and route planning.

04 Cyber defence

Network threat detection · Phishing and
data-leak alerts.

05 Intelligence

Document exploitation and translation ·
Open-source analysis.

06 Command and knowledge

Secure doctrine assistant · Staff briefing
support.

07 Training and simulation

Digital twins of terrain · Synthetic training data.

08 Disaster response

Flood and cyclone mapping · Relief route
planning.

→ Start here

Doctrine assistant, staff briefings and document
translation: no new sensors, and they run on the
first H200 server.

MISSION AREAS 1-4

See, watch, sustain, defend

SURVEILLANCE AND ISR

See more, faster

- ✓ **Imagery analysis**
Finds vehicles, vessels, new structures and changes in drone and satellite images.
- ✓ **Video search and summaries**
Searches hours of drone and CCTV video in plain language.

Built with: Metropolis · vision NIM · VSS Blueprint

BORDER AND MARITIME

Watch border and coast

- ✓ **Border monitoring alerts**
Fuses camera, radar and sensor feeds; alerts on unusual movement.
- ✓ **Maritime domain awareness**
Flags vessels that behave unusually or hide their identity.

Built with: Metropolis · DeepStream · RAPIDS

LOGISTICS AND MAINTENANCE

Keep equipment ready

- ✓ **Predictive maintenance**
Predicts which vehicles, aircraft or vessels are likely to fail.
- ✓ **Supply and route planning**
Plans routes, loads and schedules across depots and units.

Built with: RAPIDS · cuOpt · NIM

CYBER DEFENCE

Defend the network

- ✓ **Network threat detection**
Watches traffic and logs to spot intrusions early.
- ✓ **Phishing and data-leak alerts**
Flags phishing and sensitive data leaving the network.

Built with: Morpheus · RAPIDS · NeMo Guardrails

Two use cases per mission area. Examples, benefits and full build details are in the speaker notes.

→ MISSION AREAS 5–8

Understand, command, train, respond

Four more mission areas, two use cases each, on the same in-country platform.

01

INTELLIGENCE

Document exploitation: reads, translates and summarises seized documents.

Open-source analysis: spots threats and disinformation.

02

COMMAND AND KNOWLEDGE

Doctrine assistant: answers from manuals and orders, with sources.

Staff briefings: unit reports become short daily summaries.

03

TRAINING AND SIMULATION

Terrain digital twins: 3D copies of sites for planning and rehearsal.

Synthetic data: images to train detectors when real ones are rare.

04

DISASTER RESPONSE

Flood and cyclone mapping: maps floods, cut roads and people in need.

Relief routes: fastest routes for boats, trucks and helicopters.

→ WORKLOAD PLACEMENT

Which use case runs on which GPU

L40S CARRIES THE SMALL, PARALLEL MODELS; H200 CARRIES THE LARGE MODELS AND FINE-TUNING.

Use case	Model or SDK	GPU and precision	Why
Chat assistant	Llama 3.3 70B or Llama Nemotron Super NIM	H200, FP8, 1 GPU	141 GB of HBM holds a 70B model plus a large KV cache
Document search	NeMo Retriever embedding and reranking NIM	L40S, FP16	Small models, many parallel requests
OCR and PDF extraction	NeMo Retriever extraction NIM	L40S, FP16	Vision models, modest memory, runs for days
Data processing	RAPIDS, Spark RAPIDS, cuML	L40S or a MIG slice	Memory bandwidth and storage matter more than size
Speech and translation	Riva ASR, TTS and NMT	L40S, FP16	Low latency, small models
Fine-tuning on Army data	NeMo Customizer	H200 with NVLink, BF16	Gradients cross GPUs on every step
Frontier model, 253B and up	Llama Nemotron Ultra NIM	8× H200, FP8	Needs NVLink and HBM bandwidth

Precision follows the rule on 2.2: FP8 to serve, BF16 to fine-tune. GPU memory figures as on 2.1.