

Datacentres for AI

How the LLM revolution reshaped the data centre and network landscape



Andy Long

Enterprise Sales Director / Verda

From web-scale halls to AI factories — what changed

- 01 The AI inflection point
- 02 Roadmap
- 03 Building around the network
- 04 The Intra-DC AI fabric
- 05 The WAN dimension
- 06 Building for a New Physics
- 07 The Forced Migration to Liquid
- 08 Traffic Spikes & Utilization
- 09 What Verda is doing

What the industry looked like then:

Cloud-first



5–15 kW racks
– the whole industry.

CPU-centric



Air cooling, everywhere.

Steady growth



Predictable demand you could
plan around.

The AI Inflection Point

The catalyst moment (ChatGPT, end of 2022) and the resulting land rush.

Training vs. inference

two different workloads, two
different infrastructure profiles

Scale shift: datacenter silicon 3-4x the cost of DC construction

- / A GPU cluster is only as fast as its collective communication (all-reduce traffic)
- / Compute scaling $\rightarrow$ communication scaling ($O(n)$ GPUs = nontrivial traffic growth)

Training vs. inference traffic profiles:

- / Training: synchronized, bursty, east-west, extremely loss-sensitive
- / Inference: north-south + east-west mix, spiky, latency-critical

The density jump: 5–15 kW racks to 600 kW Kyber — and climbing

Grid connection queues are measured in years in many markets — power is the gating resource.

5–15 kW

2015 — the whole industry

150–230+ kW

Grace Blackwell / Vera Rubin 2026/7

600 kW

Rubin Ultra / Kyber rack / 2027

Hopper

Grace Blackwell

Vera Rubin • 2026

Rubin Ultra • 2027

Feynman • 2028

H100/H200 — air still survives at the low end.

GB200/GB300 NVL72 — liquid becomes mandatory.

VR200 NVL144 — 150–230+ kW per rack.

The 600 kW Kyber rack.

Co-packaged optics — transceivers leave the equation.

Cheap, green, and available: the Nordic answer to the power problem

Power

Abundant hydro + wind

Low-carbon power at low cost — while other markets queue for grid connections measured in years.

Climate

Cool by default

Free-air cooling most of the year; coastal access as a structural PUE advantage.

Grid

Stable and available

Strong grids built for heavy industry — capacity where the queue is short.

Economics

Low PUE → low price

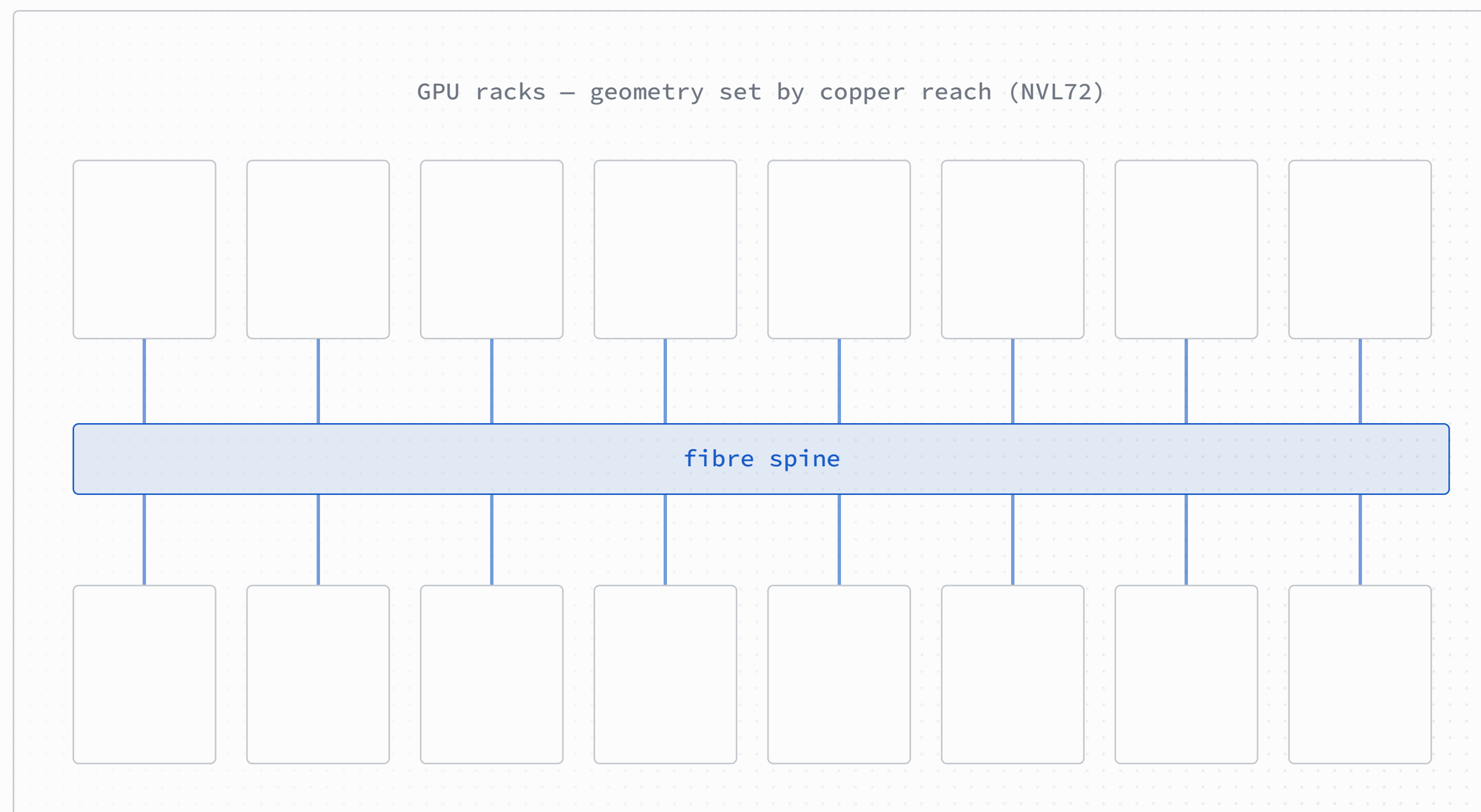
Low cooling cost flows straight through to more competitive prices for the customer.

Fibre is now the critical path of construction — not an afterthought

Orders of magnitude more fibre than legacy halls.

Spine density, patch management, riser space and route diversity designed up-front.

Deterministic, purpose-built cabling — not the generic trays of 2015.



Modular pods, pre-cabling, repeatable GPU-hall reference designs — speed to market.

Copper reach defines pod geometry: rack-and-row layouts follow the cables.

Legacy halls rarely deliver fibre + power + cooling together — greenfield advantage.

InfiniBand won training

— Ethernet is counter-attacking

InfiniBand — why it won training

- Lossless by design; credit-based flow control.
- SHARP: collective offload inside the network.
- RDMA end to end.
- Rail-optimized fat tree / Dragonfly+ topologies.
- The price: cost premium and NVIDIA vertical integration.

Ethernet — the counter-attack

- RoCEv2 viable at scale — Meta and others prove 24k+ GPU clusters.
- Ultra Ethernet Consortium: RDMA, congestion control, multipathing fixes.
- Ecosystem volume economics vs. InfiniBand's performance lead.

What lossless means in practice — and where the fabric goes next

Congestion & loss

One drop stalls 10,000 GPUs

A single dropped packet stalls the whole collective — lossless fabrics, ECN/PFC and adaptive routing are mandatory.

Topology

1:1 rail-optimized back-ends

Non-blocking beats oversubscription where collectives live; the back-end network is never the place to save money.

Four networks per hall

Not one fabric — four

Back-end GPU fabric, front-end, storage, and OOB management — each with its own design rules.

Optics explosion

400G → 800G → 1.6T

Module power and thermals are now material; co-packaged optics is the endgame.

No single site has enough power — clusters now span buildings and sites

Training jobs spanning campuses need hundreds of Tb/s between sites. Physics decides what is possible.

$\sim 5 \mu\text{s}/\text{km}$

Round-trip latency — the speed of light is the budget

$< 10 \text{ km}$

Campus: active training fabrics over dark fibre — feasible

$< 50 \text{ km}$

Metro: borderline for synchronous training

100s km

Regional: pipeline parallelism and replicated training only

Customer expectation

Dedicated links into the hall — Direct Connect / ExpressRoute-style access to assets inside the data centre.

From gigabit rings to terabit waves: the WAN grew up

AI traffic is elephant flows and PB-scale checkpoint bursts — not cacheable user traffic.

Coherent pluggables	100G/400G ZR/ZR+ collapse DWDM into the routers.
Dark fibre backbones	Long-haul coherent for AI backbones; fibre pair counts exploding.
Terabit waves	Wave capacity planned for cluster-to-cluster flows, not consumer peak.
Reliability	WAN flaps kill checkpoint integrity — diverse paths, fast failover, deep telemetry.
The Nordic angle	Sweden → Europe routes with multiple diverse carriers; low-latency paths to Frankfurt and Amsterdam. Training lives remotely and still connects.

Site selection is power first, location second — grid access drives everything

Site selection

Power first

Grid access decides where the hall goes; everything else negotiates around it.

Structural

Heavier, denser, wetter

Heavier floor loads, liquid infrastructure, smaller footprints, tightly interconnected clusters.

Speed

Modular and prefab

Modular construction beats traditional build to market — and repeats.

Retrofit vs greenfield

Can legacy DCs survive?

Most legacy halls cannot deliver the power, cooling and fibre density together.

Takeaway

GPU fleets switch on in waves — power draw is turbulent, and neither DCs nor grid operators were built for that.

The forced migration to liquid

Air’s realistic ceiling is ~80 kW per rack with rear doors — the roadmap left that behind two generations ago.

Direct-to-chip today

DLC is the current state of affairs — with its own risks and operational challenges.

Immersion tomorrow?

best potential from a physics standpoint — but too many practical challenges

Facility changes

CDUs, warm-water loops — the mechanical plant becomes part of the compute design.

Heat reuse

District heating from waste heat — a Nordic advantage.

Free cooling

Nordic ambient air and coastal access as a structural PUE and cost advantage.

GPUs are too expensive to idle — and each workload idles differently

Training

Near-100%, but fragile

Checkpointing and failure domains spanning hundreds or thousands of GPUs — reliability of the interconnect is the whole game.

Inference

Spiky and unpredictable

Autoscaling, disaggregated serving, Kubernetes + GPU schedulers, capacity mixing with spot and batch underlays.

Stranded capacity

Who absorbs the idle cost?

The economics of underused clusters decide who survives this market.

From minimum viable clusters to 4,000-GPU builds

500+

Nodes per build — and scaling

4,000

GPUs per cluster build

NVL72

GB300 rack-scale, homogeneous fleets

NVIDIA reference architecture

Building to the reference designs — first in line for each generation.

IB and RoCE — and what’s next

Running both today; watching Ultra Ethernet closely.

Co-packaged optics

NVIDIA CPO roadmap (Quantum-X800 InfiniBand switch) — real power savings from removing transceivers from the equation.

Three takeaways

01

We are still early

The scale of DC build and GPU clusters for training and inference is still accelerating.

02

In-cluster traffic grows exponentially

Disaggregation of the inference chain multiplies east-west traffic further.

03

The spiral continues

Interconnect speed and rack kW keep climbing — multi-site distributed training is the real frontier.

Thank you!