PART 8 · SECOND EDITION · SEPTEMBER 2026 (FIRST PUBLISHED JULY 2026)

Networking, Fabrics & Optics

The fabric is the part of the computer that decides how fast every other part runs. Buy it on link rate and you'll pass every acceptance test — then watch the training job expose the tail the bandwidth test never touched.

THE FLEET-SCALE TRUTH

about 110 s between flaps; about 1 h between hard failures

Ten million optics: a flap about every two minutes, a hard failure about every hour — the job waits every step

Assumed component rates, not fleet telemetry: hard failures about hourly, flaps every couple of minutes — and MTBF specs count only the former. This is where goodput leaks.

PER-OPTIC ASSUMPTIONS · AS-OF 2026-03-04

THE REFRAME

The network is not plumbing between computers. It is the part of the computer that decides how fast every other part is allowed to run.

A synchronous job moves at the speed of its slowest collective and stalls on its longest tail. Design for delivered goodput, not raw link rate — only one of them shows up in wall-clock time.

THE NORTH-STAR METRIC

90% vs 96% scenario

Illustrative 90%-versus-96% goodput sensitivity — not an industry baseline

Use these points only to test economics; replace them with the named fleet's measured goodput.

GUIDE ANALYSIS · AS-OF 2026-09-08

ACT I

Three networks, one computer

Scale-up, scale-out, scale-across. Bandwidth drops by orders of magnitude at each boundary — so where you draw the boundaries decides which traffic runs at memory speed and which crawls.

THE TERRITORY

Bandwidth falls off a cliff at each tier boundary

Three network tiers: put traffic where it fits

Per-GPU bandwidth per direction on GB300 NVL72, to scale. The WAN tier has no fixed rate.

Scale-up
inside the NVLink rack

~0.9 TB/s per direction · NVLink 5

Carries tensor- and expert-parallel collectives, every layer

Scale-out
back-end fabric across racks

800G NIC per GPU · ConnectX-8 on GB300 (~9x below NV

Carries data-parallel all-reduce and pipeline sends

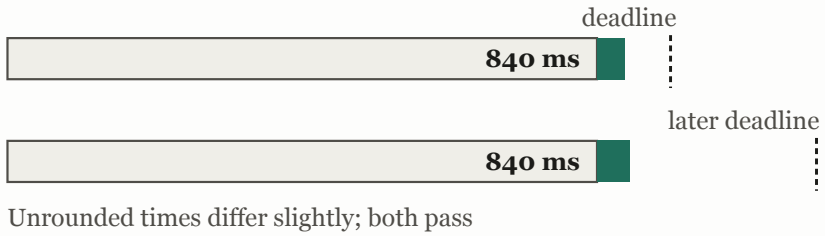
Scale-across
WAN between campuses

No fixed rate: measured RTT and leased capacity decide

Carries hierarchical or relaxed-sync exchange over long paths

Before a group crosses a tier, test its step healthy and degraded

other work · exposed communication



Unrounded times differ slightly; both pass

Fits in both states → cross the tier

Misses in either → keep local or reschedule

Guide scenario · Source: NVIDIA + Guide derivation · as of October 2026

Tightly-coupled parallelism stays inside the scale-up domain unless a hierarchical exchange still makes the step deadline with a link down. Every other Part 8 decision is downstream of that boundary.

THE CLIFF, QUANTIFIED

1.8 TB/s bidirectional (~0.9 TB/s per direction)

NVLink-5 per-GPU scale-up bandwidth — roughly eighteen times what a 400G scale-out NIC delivers per direction

Push a tensor-parallel all-reduce that wanted this fabric onto the NIC and model-FLOPS-utilization collapses. The parallelism mapping is not optional.

NVIDIA GB200 NVL72 PRODUCT PAGE · AS-OF 2026-10-16

THE PLACEMENT RULE

Map the parallelism to the tier — or starve the collective

PARALLELISM	DOMINANT COLLECTIVE	TARGET TIER	WHY IT LANDS THERE
Tensor (TP)	All-gather / reduce-scatter, every layer	Scale-up (NVLink)	Highest-frequency and bandwidth-bound; only the memory-speed fabric keeps up
Expert (EP, MoE)	All-to-all token routing, every MoE layer	Scale-up; spills to scale-out as EP grows	Bisection-stressing; wider domains raise the EP ceiling
Pipeline (PP)	Point-to-point at micro-batch boundaries	Scale-out	Sparse and latency-tolerant; a cheap packet fabric suffices
Data (DP / FSDP)	DP all-reduces gradients; FSDP all-gathers parameters and reduce-scatters gradients by wrapped module	Scale-out; scale-across at multi-campus	Wrapping, prefetch, buckets, model and runtime set frequency and overlap — measure before oversubscribing

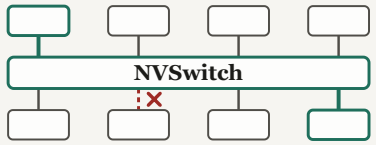
THE MOST CONTESTED DECISION IN AI NETWORKING

How large can the scale-up domain get — and over what medium?

Four scale-up domain shapes

What the memory-speed domain looks like, and what a fault takes out

HGX B200 node, eight GPUs

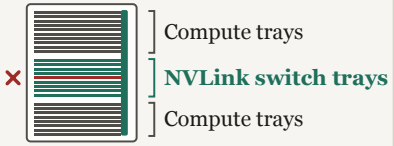


1.8 TB/s per GPU

0.9 TB/s each way, any pair (green)

Link fault: one path lost

GB200 / GB300 NVL72, 72 GPUs



~130 TB/s aggregate

Nine switch trays on a copper spine

GB200: a switch-tray fault removes NVLink from every partition

AMD MI350-series, eight GPUs

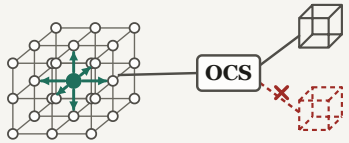
No switch: each GPU wires straight to the **other seven**

1,075.2 GB/s per GPU

Seven links of 153.6 GB/s each

A GPU pair gets one link: a seventh of the total

Google TPU ICI, Ironwood pod



9,216 chips per Ironwood pod

Six neighbors per chip, 3D torus

OCS reroutes around a failed cube

Outside the domain, the NIC is an order of magnitude slower

Per GPU: scale-up link against the scale-out NIC, bars drawn to scale

GB200, NVLink 5	GB300, NVLink 5	Vera Rubin, NVLink 6
1.8 TB/s	1.8 TB/s	3.0 TB/s
NIC ~18× slower	NIC ~9× slower	NIC ~7.5× slower

Measured/sourced · Source: NVIDIA + AMD + Google · as of September 2026

The domain boundary is the bandwidth cliff. Its size caps tensor and expert parallelism, sets MoE inference economics — and sets your largest blast radius.

THE LOCK-IN FORK

Buy the domain integrated, or assemble it open?

NVLink / NVSwitch — integrated

— Deepest co-design; the largest shipping domains

— In-fabric reduction (SHARP) built in

— Vendor sets the roadmap, the margin, and the parts list

— NVLink Fusion opens a licensed crack for third-party silicon

UALink / Scale-Up Ethernet — open

— Clean-room standards backed by AMD, Intel, Broadcom, hyperscalers

— Commoditize the switch and the accelerator port

— One Ethernet silicon can serve scale-up and scale-out

— Shipping ecosystem still maturing in 2026

SO WHAT

Every gain from a bigger domain is paid in blast radius: one switch-tray fault can now stall the whole scale-up domain, not eight GPUs. Addressability and coherence semantics remain platform-specific.

DOMAIN SIZES ON THE TABLE

How big the scale-up domain gets, by camp

up to 1,152 XPUs (future roadmap)

NVLink domain roadmap: NVL72 shipping; NVL576 and NVL1152 announced, not proven

NVIDIA NVLINK FUSION PAGE · AS-OF 2026-09

1,024

UALink 200G 1.0 domain ceiling — sub-microsecond RTT at short reach

UALINK CONSORTIUM RATIFICATION ANNOUNCEMENT. · AS-OF 2025-04-08

9,216 chips

Google's Ironwood superpod: optical circuit switching instead of a bigger rack

GOOGLE IRONWOOD ANNOUNCEMENT, 2025-04-09. · AS-OF 2025-04-09

ACT II

The scale-out fabric: silicon, protocol, topology

Everything above the rack is a chain of bets: the SerDes gates the ASIC, the ASIC gates the optics and the topology, and the protocol decides how much of the link rate survives as goodput.

THE GATING SPEC

SerDes generation — not marketed Tbps — gates everything downstream

Matching port speed is not enough

Every layer of the port path has to line up, not just the headline rate

Server

NIC
endpoint ops

Switch, from faceplate to queues

Cage	SerDes	FEC	MAC	Queues
breakout	lanes	error coding	port rate	buffer, NOS

ConnectX-8 runs Ethernet as 2x400GbE
Rate matches, but no single full-rate link

What each vendor spec states, by layer (— = not stated)

Switch	Cage	SerDes	FEC	MAC	Queues
Tomahawk 6	—	lane options	—	102.4 Tb/s	—
Cisco G300	—	224G-class	—	102.4 Tb/s 64×1.6T	—
Spectrum-6	—	—	—	102.4 Tb/s	160 MB

Every layer lines up
Release silicon and port map

A cage, queue or NIC mode fails
Change the supported system

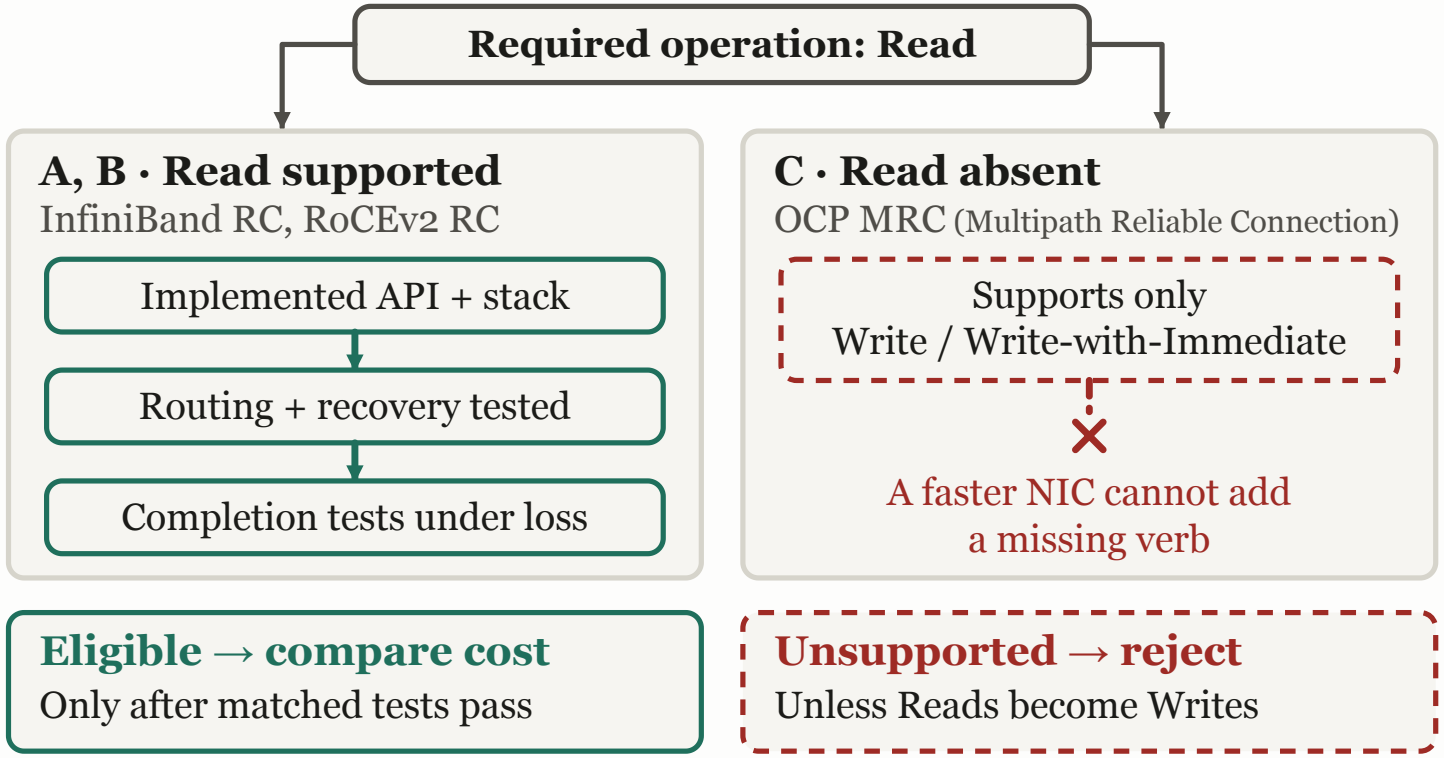
Schematic · Source: Broadcom + Cisco + NVIDIA · as of Sep 2026

Pick the lane rate before you draw a topology: it sets port speed, copper reach, and the optics bill. The topology is downstream of the silicon.

REQUIRED OPERATIONS FIRST **No RDMA Read on the stack, no Read workload — check operations before latency**

Exclude unsupported operations first

Illustrative case: the workload needs RDMA Read



Schematic · Source: OCP MRC 1.0 specification + Guide derivation · as of Sep 2026

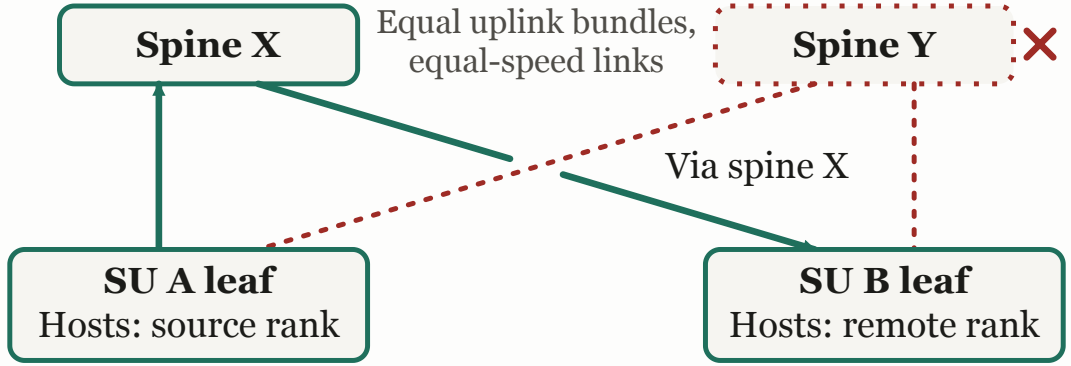
A transport missing the operation the workload issues fails before latency matters. Qualify routing and recovery, then compare matched payload and completion tests; MPI figures rank nothing.

THE COUNTED RAIL, THREE CURRENCIES

Topology trades three currencies: bisection, optics, and hops

Count paths before ports

Illustrative rail: losing one of two spines halves the uplinks



Healthy rail
840 ms step
All uplinks in service
Meets the healthy deadline

Spine Y lost
1,030 ms step
Half the uplinks remain
Still meets the degraded deadline

Rail-local: test the failed state
Count every spine crossing

Cross-rail: needs a general path
Budget additional paths

Full eight-rail build: 32 switches / 2,048 modules (optical)

Guide scenario · Source: Guide derivation · as of Sep 2026

No family wins all three. Rail-optimized fat-tree is the default because same-rail collectives in an SU skip the spine — every cross-SU exchange still crosses one, so count that path and its optics.

THE LEVER THAT PRICES THE FABRIC

1:1 vs 2:1–3:1

Set oversubscription from measured traffic and SLOs — workload labels are only inputs

Normalize the traffic matrix, collective/request mix, topology, placement, failure headroom and tail-SLO before pricing the ratio; prove congestion and recovery on a representative test fabric.

SEMIANALYSIS NEOCLOUD PLAYBOOK · AS-OF 2025

CONGESTION: THREE FIGHTS AT ONCE

Symptom, cause, and the right lever — never 'add bandwidth'

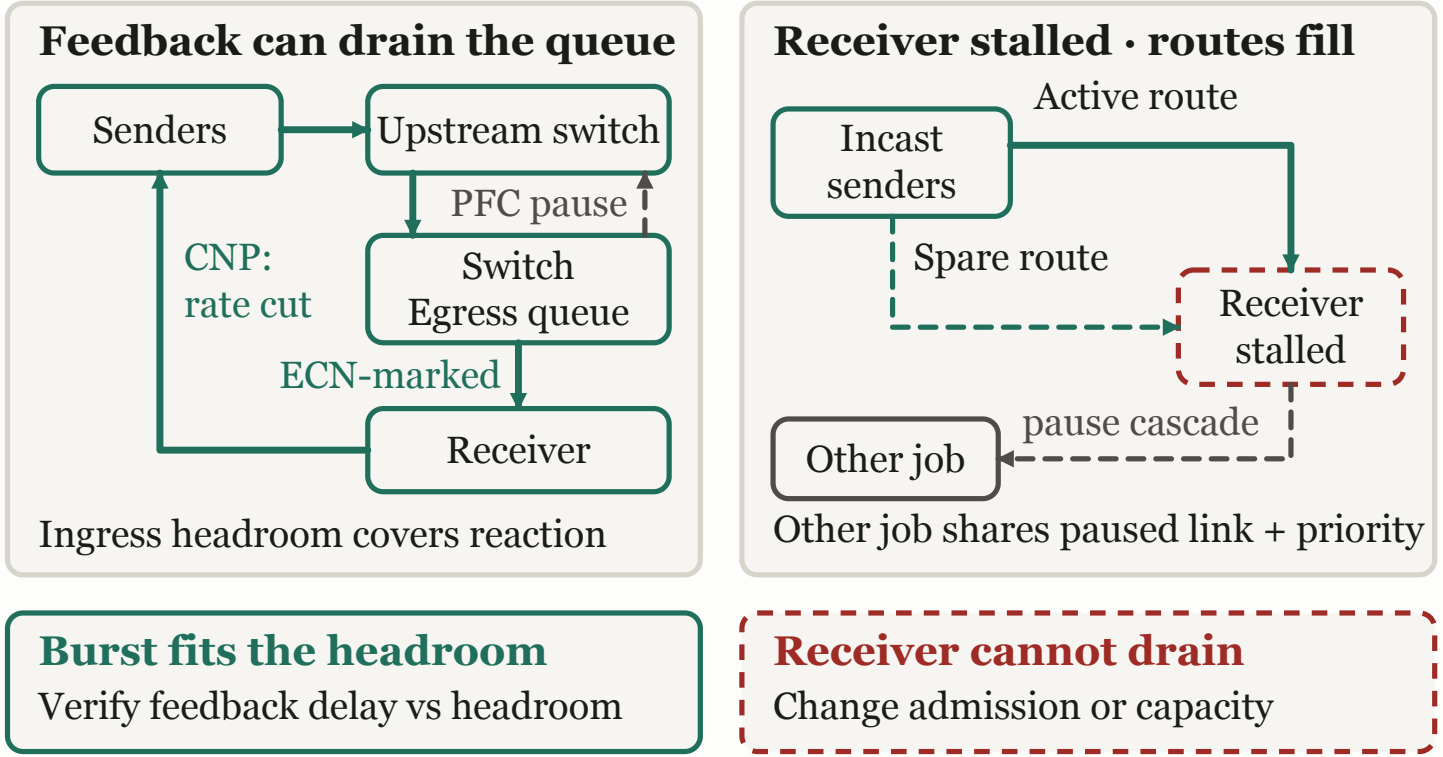
SYMPTOM	UNDERLYING CAUSE	RIGHT LEVER
Fabric-wide slowdown from one hot port	PFC pause propagating upstream	Tune DCQCN to stay out of pause; PFC watchdog
One spine link saturated, parallels idle	Elephant-flow ECMP hash collision	Adaptive routing, or packet spray + NIC reorder
All-reduce slower than link rate implies	Reduction overhead on GPUs and wire	In-network reduction (SHARP)
Healthy cluster, job time creeping up	Untuned congestion loop leaving bandwidth idle	Re-tune ECN thresholds; instrument telemetry

THE FEEDBACK LOOP AND THE TRAP

Tune ECN before PFC — or one hot port can stall the fabric

ECN slows the senders; PFC pauses the neighbor

Illustrative RoCE mechanism · incast at a shared receiver



Guide scenario · Source: Guide derivation · as of Sep 2026

Set ECN to bite as the queue grows and size PFC headroom for bytes in flight during a pause, so pause frames stay a net you rarely trip. Get it wrong and one hot port stalls flows three hops away.

WHAT WINNING THE FIGHT IS WORTH

The congestion dividend, in two numbers

~95%

Data throughput NVIDIA reports for xAI's Colossus on Spectrum-X Ethernet (vendor-reported)

NVIDIA NEWSROOM, 'NVIDIA ETHERNET NETWORKING ACCELERATES WORLD'S LARGEST AI SUPERCOMPUTER, BUILT BY XAI' · AS-OF 2024-10

2(n−1)/n; approaches 2 as ranks increase

Effective all-reduce bandwidth gained by moving the reduction into the switch

NCCL-TESTS PERFORMANCE.MD ON GITHUB: RING BUS-BANDWIDTH NORMALIZATION. GUIDE DERIVATION OF THE IDEAL ENDPOINT-BYTE COMPARISON. · AS-OF 2026-09

ACT III

The physical layer: where the watts and the failures live

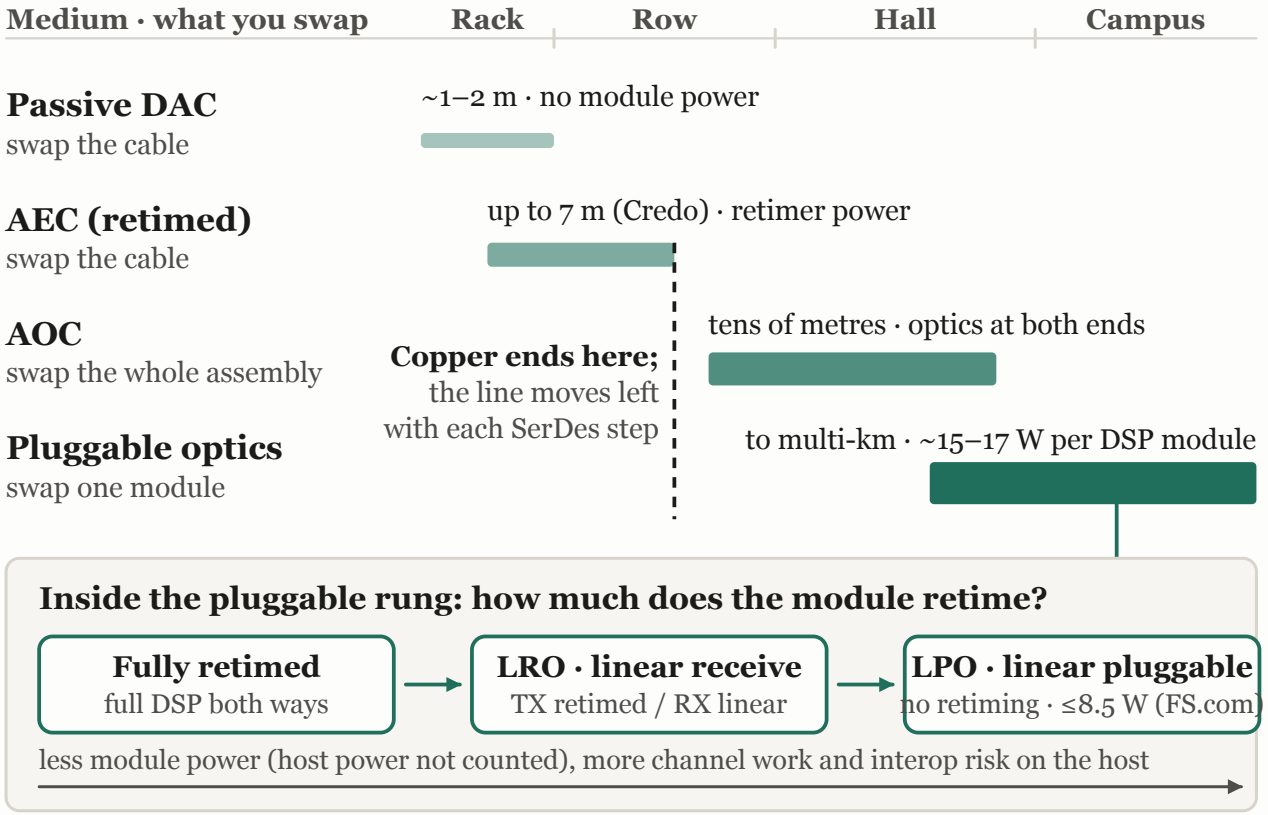
Every link is a bet that a reach can be crossed by the cheapest, lowest-power medium that still closes the error budget. At today's lane rates that bet is lost a metre sooner than one generation ago.

THE REACH LADDER

Use copper where it reaches; buy optical reach only where it cannot

Every rung of reach costs more power per port

Reference reach at current lane rates; thicker bar, more watts per port



Schematic · Source: SemiAnalysis + Credo + FS + Juniper · as of January 2026

Each rung up costs more watts, dollars and failures — climb only when host, channel or repair unit fail the budget below. Faster SerDes keeps pulling copper's reach inward; re-qualify it each refresh.

SHORT-REACH COPPER

~20 kW/rack

Transceiver and retimer power NVIDIA says an optical NVL72 scale-up would have needed

A counterfactual vendor estimate from NVIDIA's GTC 2024 keynote, as reported by SemiAnalysis. Copper NVLink avoids it; compare complete channel power and the required reach.

SEMIANALYSIS, NVIDIA'S OPTICAL BOOGEYMAN, 25 MARCH 2024, QUOTING NVIDIA'S GTC 2024 KEYNOTE: THE TRANSCEIVERS AND RETIMERS ALONE WOULD HAVE COST 20,000 WATTS · AS-OF 2024-03

THE DECADE'S OPTICS FORK

Optics on the faceplate, or on the package?

Pluggable (incl. LPO/LRO)

— Hot-swap in seconds; multi-vendor, second-sourced

— Commodity market disciplines the price

— DSP burns the most power per port

— Signal crosses lossy PCB before it sees a laser

Co-packaged optics (CPO)

— Optical engine millimetres from the switch ASIC

— The biggest per-port power cut on the menu

— Not faceplate-swappable — the FRU is the sub-assembly

— Single vendor's laser, process, and packaging

SO WHAT

CPO trades the worst power problem in the building for the worst serviceability problem in the building — and collapses a multi-vendor optics market into one supplier. Price the lock-in, not just the watts.

WHERE CPO ACTUALLY STANDS

Real savings, tiny footprint — for now

~0.5%

CPO share of AI optical modules in 2026 — top-of-fabric first, spreading outward

TRENDFORCE ORIGINAL FORECAST ANNOUNCEMENT · AS-OF 2026-03-11

~$26B

The AI optical-transceiver market CPO is trying to unseat

TRENDFORCE PRESS RELEASE · AS-OF 2026-04-20

BACK TO THE FLAP CLOCK

about 20–30%

Share of cluster productivity operators report losing to link instability before the bad optics are found

MTBF specs count hard failures; goodput dies by flaps. Commission against the flap: qualify optics at field temperature with real margin, gate on connector cleanliness, instrument every link.

CREDO BLOG · AS-OF 2025-12

ACT IV

Beyond the building — and beneath the fabric

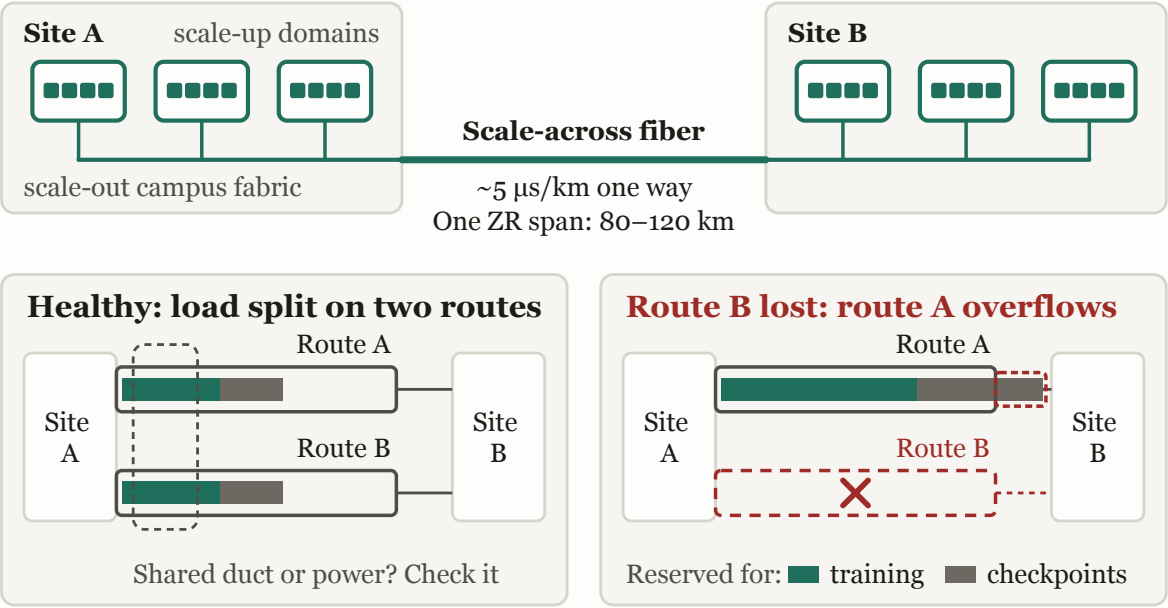
When portfolio capacity spans sites, DCI can carry replication, failover, burst traffic, or qualified cross-site work. A job crosses only when its algorithm, latency, bandwidth, placement, and failure model close.

THE CONDITIONAL HIERARCHY

Portfolio DCI first; only qualified jobs span sites

Lose one route: does the job still fit?

Each tier out is slower, and the last one can lose a route



Exposed delay = max(0, bytes ÷ reserved payload rate + rounds × RTT − valid overlap)

Fits after route loss
Admit; keep a restorable checkpoint

Does not fit
Change placement or sync

Schematic · Source: M2 Optics + OIF-800ZR-01.0 + Guide derivation · as of Sep 2026

Scale-across connects portfolio sites. Keep jobs site-local by default; cross sites only where the algorithm, latency budget, bandwidth, placement, and failure model are qualified.

THE FLOOR ON STANDARD FIBER

~5 µs/km

The one-way latency floor of light in fiber — the radius between sites picks your training algorithm for you

Within a metro, a run can stay synchronous. Across a region it cannot on standard fiber, whatever the fabric spend; hollow-core (~3.4 µs/km) is route-specific. Decide radius and algorithm together.

M2 OPTICS · AS-OF 2025

THE RELAXED-SYNCHRONY ESCAPE HATCH

When the sites are far, change the algorithm, not the optics

~500x

Less inter-site traffic for DiLoCo vs fully-synchronous SGD, at matching convergence

DILOCO, DOUILLARD ET AL. · AS-OF 2024

90–95%

Compute utilization OpenDiLoCo sustained training a 1.1B model across two continents

OPENDILOCO · AS-OF 2024-07

THE INVISIBLE FABRICS

The clock you diagnose with, and the network you rescue with

~10 ns

PTP accuracy in NVIDIA's Spectrum-3 boundary-clock test, the clock behind log correlation

NVIDIA TECHNICAL BLOG, 'CALCULATING AND SYNCHRONIZING TIME WITH THE PRECISION TIMING PROTOCOL ON THE NVIDIA SPECTRUM SWITCH' · AS-OF 2022-09

~1 ms

Typical NTP accuracy on a LAN — orders of magnitude too blurry to chase a microburst

LINUXPTP · AS-OF 2025

THE THROUGH-LINE

You don't buy a network; you buy goodput. Every choice in this Part — domain size, protocol, topology, medium — is the same question asked at a different layer: how much of the link rate survives to the training step?

TAKE IT HOME

What to do Monday

01 Draw the scale-up boundary first — the parallelism mapping and everything downstream inherit it.

02 Characterize the traffic before sizing the fabric: collective profile, message sizes, incast budget — not peak link rate.

03 Pick the transport as a lock-in decision: proprietary-and-finished now, or open-and-maturing — and keep an Ethernet exit.

04 Set oversubscription from measured traffic, request mix, topology, failures and SLOs — then prove it on a test fabric.

05 Keep copper wherever it reaches; when it can't, budget optics as a top capital line and the leading failure source.

06 Commission against the flap, not the MTBF — and stand up the OOB fabric and PTP timing plane before the first incident.

THE DEFINITIVE GUIDE TO

AI Data *Centers*

This deck is the map. The guide is the territory — every number here is dated, sourced, and kept current in the full chapters.

Ch 8.1 Traffic characterization and the goodput north-star

Ch 8.2 The scale-up domain: size, medium, and the lock-in fork

Ch 8.4 The protocol war: IB, RoCE, Spectrum-X, Ultra Ethernet

Ch 8.9 The copper-optics boundary and commissioning against the flap

READ THE FULL PART, WITH EVERY SOURCE AND DERIVATION, AT AIDATACENTERGUIDE.COM

SOURCES

Every figure in this deck traces to the guide's dated register (1/3)

Per-optic assumptions (10^7 h hard MTBF, 3x10^5 h mean time to flap) from an unattributed OCP working-session slide (2025-11-06), as reproduced in Borrill, arXiv 2603.03736 v2 (2026-09-09), which does not validate them; the guide recomputes the aggregate for exactly 10M modules. · as-of 2026-03-04 · https://arxiv.org/html/2603.03736v2

Guide analysis — stipulated sensitivity scenario; no claim of an industry measurement. · as-of 2026-09-08

NVIDIA GB200 NVL72 product page (undated; checked 2026-10-16) · as-of 2026-10-16 · https://www.nvidia.com/en-us/data-center/gb200-nvl72/

NVIDIA NVLink Fusion page (checked 2026-09-29); no delivery date given. · as-of 2026-09 · https://www.nvidia.com/en-us/data-center/nvlink-fusion/

UALink Consortium ratification announcement. · as-of 2025-04-08 · https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-Specification-PR_FINAL.pdf

Google Ironwood announcement, 2025-04-09. · as-of 2025-04-09 · https://blog.google/products/google-cloud/ironwood-tpu-age-of-inference/

SemiAnalysis Neocloud Playbook (newsletter.semianalysis.com/p/ai-neocloud-playbook-and-anatomy); Juniper AI-cluster design; Meta, The Llama 3 Herd of Models (arXiv 2407.21783, 2024-07-31) for the 7:1 case · as-of 2025 · https://arxiv.org/html/2407.21783

SOURCES

Every figure in this deck traces to the guide’s dated register (2/3)

NVIDIA Newsroom, 'NVIDIA Ethernet Networking Accelerates World's Largest AI Supercomputer, Built by xAI' (Oct 2024) · as-of 2024-10 · https://nvidianews.nvidia.com/news/spectrum-x-ethernet-networking-xai-colossus

NCCL-tests PERFORMANCE.md on GitHub: ring bus-bandwidth normalization. Guide derivation of the ideal endpoint-byte comparison. · as-of 2026-09 · https://github.com/NVIDIA/nccl-tests/blob/master/doc/PERFORMANCE.md

SemiAnalysis, Nvidia’s Optical Boogeyman, 25 March 2024, quoting NVIDIA’s GTC 2024 keynote: the transceivers and retimers alone would have cost 20,000 watts — 20 kilowatts — per NVL72 rack had scale-up used optics instead of copper. · as-of 2024-03 · https://newsletter.semianalysis.com/p/nvidias-optical-boogeyman-nvl72-infiniband

TrendForce original forecast announcement; no observed shipment total inferred. · as-of 2026-03-11 · https://www.trendforce.com/presscenter/news/20260311-12962.html

TrendForce press release (2026-04-20) · as-of 2026-04-20 · https://www.trendforce.com/presscenter/news/20260420-13017.html

Credo blog (Bill Brennan, CEO) relaying an unnamed customer's report, 2025-12-10 · as-of 2025-12 · https://credosemi.com/blogs/the-path-to-zero-flap-reinventing-optical-reliability-for-scalable-ai-clusters/

M2 Optics · as-of 2025 · https://www.m2optics.com/blog/bid/70587/Calculating-Optical-Fiber-Latency

SOURCES

Every figure in this deck traces to the guide's dated register (3/3)

DiLoCo, Douillard et al. (arXiv:2311.08105) · as-of 2024 · https://arxiv.org/abs/2311.08105

OpenDiLoCo (Prime Intellect) · as-of 2024-07 · https://www.primeintellect.ai/blog/opendiloco

NVIDIA Technical Blog, 'Calculating and Synchronizing Time with the Precision Timing Protocol on the NVIDIA Spectrum Switch' (developer.nvidia.com, 2022) · as-of 2022-09 · https://developer.nvidia.com/blog/calculating-and-synchronizing-time-with-the-precision-timing-protocol-on-the-nvidia-spectrum-switch/

linuxptp / SUSE & Red Hat PTP tuning guides · as-of 2025 · https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/6/html/deployment_guide/s1-understanding_ntp

SOURCES

About these figures

Figures are resolved by reference from the guide's claim register at build time (Second edition · September 2026 (first published July 2026)) — a number appears here only as it appears, dated and sourced, at aidatacenterguide.com. Newer data may exist: check the live register.