Optical Layer Design for AI Data Center Fabric

Author: Akash Patel

1) Scope & Assumptions

This document focusses only on **Backend Fabric**. The storage Fabric would be very similar to this scope and Frontend Fabric would be much smaller in terms of # of switches ,cables and optics.

- Scale: ≈2000 GPUs (2 pods × 128 nodes/pod × 8 GPUs/node).
- Racks: 4 nodes/rack ⇒ 8 racks/SU ⇒ 32 racks/pod.
- Per pod: 4 SUs × 32 nodes/SU = 128 nodes.
- Rails: rail0 and rail1 per pod (independent, end-to-end).
- NIC: NVIDIA BlueField-3 SuperNIC; 2×400G per node (1 per rail).
- Switching: NVIDIA Spectrum — Leaf/Spine at 400G; Core at 800G with 2×400G breakout.
- Goal: Non-blocking leaf↔spine within pods; inter-pod core sized with scalable 800G uplinks.
- ** This document is very high level document and doesn't take any Power, HVAC and other logistic considerations in terms of Rack design and hence all cabling. This would potentially change overall distance and hence type of cables and optics

2) Standard Row & Distance Assumptions (NVIDIA-style)

- Rows / pod: e.g., 4 rows × 8 racks/row per SU; keep SUs contiguous.
- Rack spacing: 0.6–0.8 m center-to-center; aisles 1.2–1.5 m.
- Distance bands (route lengths):
- - Same rack: DAC ≤3 m (active DAC to 5–7 m).
- - Same SU inter-rack: AOC ≤30 m.
- - Intra-pod long: DR4 ≤500 m, FR4 ≤2 km.
- - Core uplinks: 800G 2×FR4.

3) Per-Rack Patching Map (per SU / per rail) + SU/POD/Fabric Breakdown

R1 (Leaf Rack) — 4 nodes via DAC400G-3M to leaf.

R2–R4 — Active DAC400G-5M to leaf.

R5–R8 — AOC400G-15M/30M to leaf.

Leaf uplinks to spines: AOC400G-10/20/30M when ≤30 m; else 400G-DR4.

Rack	Cable Type & SKU	Length	Dest
R1	Corning DAC400G-3M	3 m	Leaf
R2–R4	Corning DAC400G-5M	5 m	Leaf
R5–R8	Corning AOC400G-15M/30M	15–30 m	Leaf

3.1 Per SU (32 nodes; both rails)

Group	Assemblies / Links	Optics (both ends)	Notes
Node→Leaf DAC	8	—	Leaf rack per rail (4 nodes) × 2 rails
Node→Leaf AOC	56	—	Remaining 7 racks per rail × 2 rails
Leaf↔Spine AOC	51	—	From leaf P33–P56 (approx.)
Leaf↔Spine DR4	13 links	26× 400G-DR4	From leaf P57–P64 (approx.)
Spine→Core 800G 2×FR4	8 links	16× 800G 2×FR4	Even allocation (pod total / 4 SUs)

3.2 Per POD (128 nodes; both rails)

Group	Assemblies / Links	Optics (both ends)	Notes
Node→Leaf DAC	32	—	4 SUs × 8
Node→Leaf AOC	224	—	4 SUs × 56
Leaf↔Spine AOC	204	—	4 SUs × 51
Leaf↔Spine DR4	52 links	104× 400G-DR4	4 SUs × 13
Spine→Core 800G 2×FR4	32 links	64× 800G 2×FR4	4 SUs × 8

3.3 Total Fabric (2 pods)

	Assemblies / Links	**Optics (both ends)**	**Notes**
Node → Leaf (DAC)	64	—	2 PODs × (4 SUs × 8)
Node → Leaf (AOC)	448	—	2 PODs × (4 SUs × 56)
Leaf ↔ Spine (AOC)	408	—	2 PODs × (4 SUs × 51)
Leaf ↔ Spine (DR4)	104 links	208 × 400G-DR4	2 PODs × (4 SUs × 13)
Spine → Core (800G 2×FR4)	64 links	128 × 800G 2×FR4	2 PODs × (4 SUs × 8)

4) Leaf↔Spine Patching (per SU / per rail)

From Leaf Port Range	Cable & SKU	Length	Dest Spine
P33–P40	Corning AOC400G-10M	10 m	S0
P41–P48	Corning AOC400G-20M	20 m	S1
P49–P56	Corning AOC400G-30M	30 m	S2
P57–P64	Corning EDGE8-DR4	SMF	S3

5) Spine↔Core Patching (per rail)

Spine	Ports	Optic	Breakout	Dest Core
S0–S3	C1–C4	800G 2×FR4	2×400G per port	Core-R Switch-A..D

6) Per-Pod BOM (Corning + NVIDIA-qualified; counts are per pod, both rails)

Media	Qty	Representative SKU

Passive DAC	32	Corning DAC400G-3M
Active DAC	48	Corning DAC400G-5M
AOC (Node→Leaf)	176	Corning AOC400G-15M/30M
AOC (Leaf↔Spine)	204	Corning AOC400G-10/20/30M
DR4 (Leaf↔Spine)	104 optics	400G-DR4 QSFP-DD + Corning MTP jumpers
800G 2×FR4 (Spine→Core)	64 optics	800G 2×FR4 + LC-LC jumpers

7) Fabric-Wide Totals (2 pods)

Cable Assemblies (per 2 PODs)

- Passive DAC → **64 assemblies**
- Active DAC → **96 assemblies**
- AOC (Node→Leaf) → **352 assemblies**
- AOC (Leaf↔Spine) → **408 assemblies**
- Total DAC + AOC **= 920 assemblies**

Optics (per-end, ×2 per link for 2 Pods)

- DR4 (Leaf↔Spine) → 104 links × 2 x 2 = **416 optics**
- 800G 2×FR4 (Spine→Core) → 64 links × 2 x2 = **256 optics**
- Total DR4 + FR4 optics **= 672**

Jumpers (patch cables)

- MTP-12 jumpers (for DR4) **→ 104 links × 2 pods =** 208 MTP-12 jumpers
- LC-LC jumpers (for FR4) **→ 64 links × 2 pods** = **128 LC-LC jumpers**

8) Structured Cabling — Corning EDGE8/EDGE16

- Trunks: EDGE8 OS2 MTP-12 Type-B (Leaf↔Spine DR4); EDGE16 OS2 MTP-16 Type-B (Spine↔Core).
- Modules: EDGE8 LC-duplex (MTP-12→LC) for DR4; EDGE16 MTP-16 pass-through for DR8 future.
- Housings: EDGE/EDGE8 HD panels (1U/2U/4U).
- Jumpers: OS2 LC-LC (1–5 m); OS2 MTP-MTP low-loss for panelized routes.

9) Ops Notes

- Keep rail0/rail1 color-coded and path-separated end-to-end.
- Inspect→clean→inspect→connect; record DOM baselines; IL/polarity tests on SMF.