FOR NETWORK, SYSTEMS AND PARTNER SOLUTION ARCHITECTS

Architecture planning guide

Turn workload requirements into a fabric design you can build, power, cool and grow.

The fabric decides how much useful work your accelerators and CPUs finish each hour. It also shapes deployment risk, power, floor space and how far the cluster can grow without a rebuild. Use this guide to make those design choices on purpose, before your partner finalizes the system bill of materials.

Key takeaways

- **Design from the workload out:** Communication patterns set the SuperNIC, the topology and how much oversubscription you can accept.
- **Size for the final cluster on day one:** Spine count and tier count set your ceiling. Changing them later means recabling.
- **Treat cabling, power and cooling as design inputs:** Rack layout, cable runs and liquid cooling all shape the rack plan.
- **Verify before production:** A cable plan you can check against the running fabric cuts bring-up time and risk.

Ten design decisions in this guide

1. Workloads to design
2. SuperNIC choice
3. Size and topology
4. Multi-plane and multi-rail
5. Worked examples
6. System integration
7. Cabling
8. Power and cooling
9. Growth and migration
10. Deploy and accept

Where this guide fits

The Technical evaluation worksheet sets your goals. The RFI/RFP checklist sets targets and tests. The Buyer brief builds support inside your company. This guide comes last: it turns all of that into a design your OEM or integrator can quote and build.

How to read the numbers

Specifications come from published Cornelis product pages, product briefs and technical documentation, cited on each page. Fabric sizing rules come from the CN5000 documentation. Confirm CN6000 details with Cornelis or your partner, since product documentation is still growing.

BEFORE YOU START, HAVE THESE READY

- **Workload profile:** How your main jobs communicate, from the Technical evaluation worksheet.
- **Scale targets:** Node count at first build and at full build, and bandwidth per node.
- **Site limits:** Rack power, cooling type (air or liquid) and the floor layout.
- **Partner and platform:** The OEM or integrator, server model and accelerators you plan to use.

1 Turn workloads into design choices: start with how jobs talk

Every fabric choice traces back to how your jobs communicate. Small, frequent messages reward high message rate and few hops. Large collectives reward bandwidth and a non-blocking core. Use the table to map each workload you run to a design starting point. Then check the starting point against the rules in sections 2 to 4.

If you run	What the traffic looks like	Design implication	Starting point
MPI and SHMEM codes (CFD, climate, life sciences)	Many small messages, frequent synchronization	Keep hop count low. Run 1:1 where collectives cross the fabric. Keep cables short.	OMNI-PATH SuperNIC, sub-1 µs MPI latency, 1.6B messages/s[1]
Large-scale AI training	All-reduce and all-to-all collectives in lockstep	Size for bandwidth per node and a non-blocking spine. Contain faults so long jobs keep running.	OMNI-PATH SuperNIC, 800G per node, multi-plane design (section 4)
Disaggregated inference	KV cache handoffs and small activation exchanges on the critical path	Optimize for latency and message rate, not peak bandwidth alone.	OMNI-PATH SuperNIC, sub-1 µs MPI latency and 1.6B messages/s[1]
Fine-tuning on a cluster that must stay on Ethernet	Bursty RDMA traffic between GPUs	Plan lossless RoCEv2 on the Ethernet switches you run.	Cornelis also offers the Ethernet SuperNIC (section 2)
Storage, checkpoint and restart	Large transfers in bursts that compete with compute traffic	Separate traffic by plane or rail, or size uplinks for the burst.	Dual-plane or dual-rail design (section 4)[2]
Containerized jobs on Kubernetes	Pod-to-pod RDMA plus IP traffic	Plan a separate Ethernet network for the Kubernetes control plane.	CN5000 fabric support for Kubernetes, with RDMA device and CNI plugins[3]

DECISION RULE: MIXED WORKLOADS

Design for the most demanding workload that must share the fabric. If one workload is latency bound and another is bandwidth bound, size the core for bandwidth and keep latency-bound jobs on fewer hops through scheduler placement.

DECISION RULE: UNKNOWN PATTERNS

If you have not profiled your jobs, plan a 1:1 core. Oversubscription saves money only when most traffic stays inside a leaf switch, and you need data to prove that.[4]

Not sure how your jobs communicate yet? Go back to the Technical evaluation worksheet, area 1, and profile them before you fix the design.

WHY IT MATTERS

Design mistakes here cost the most to fix. A fabric sized for the wrong traffic pattern leaves accelerators waiting, and the fix later is more switches, more cables and a maintenance window. Starting from the workload protects utilization and time to solution for the life of the cluster.

Sources: 1. CN6000 SuperNIC page · 2. Multi-rail topologies · 3. Running Kubernetes on a CN5000 fabric · 4. Oversubscription

2 Choose your SuperNIC: lead with Omni-Path

Start with the CN6000 Omni-Path SuperNIC. It pairs the industry's highest message rate with 800G per node on a lossless fabric, and it works with CN6000 switching or your existing CN5000 infrastructure.[1]

Choose the Omni-Path SuperNIC when

- **Jobs wait on the network:** HPC codes, large-scale training and disaggregated inference are bound by latency, message rate or collectives.
- **You run Omni-Path today:** CN6000 is backward compatible with CN5000 infrastructure, so you grow on the fabric your teams know.[1]
- **You need bigger, more resilient fabrics:** Two 400G links per SuperNIC support multi-plane designs (section 4).

Design factor	CN6000 Omni-Path SuperNIC
Switching	CN6000 Omni-Path Switch; backward compatible with CN5000 infrastructure[1,2]
Bandwidth and host link	800G per SuperNIC (dual x4-lane QSFP-DD, 112 Gb/s per lane) on PCIe 6.0 x16[1]
Message rate and latency	1.6B messages/s; under 1 µs MPI latency[1]
Lossless method	Credit-based flow control and link-level retransmission[2]
Routing and congestion	Fine-grained adaptive and static dispersive routing; hardware congestion control[1,3]
Software	OPX libfabric provider; MPI, OpenSHMEM, NCCL and RCCL[4]; GPUDirect for NVIDIA and AMD[1]
Security and management	Post-quantum Root of Trust (SPDM 1.4); SR-IOV; PLDM, MCTP and Redfish[1]

IF YOUR DESIGN MUST STAY ON ETHERNET

Cornelis also offers the Ethernet SuperNIC

It runs hardware RoCEv2 with standard Verbs on standards-based Ethernet switches, and it is Ultra Ethernet compliant (Preview).[1,5] It is part of the same CN6000 SuperNIC family and brings Omni-Path design to Ethernet.

- **Plan for:** Priority Flow Control on the standards-based switches you choose.
- **One fabric per SuperNIC:** Each SuperNIC runs one fabric, so plan each fabric for one SuperNIC type.

WHY IT MATTERS

The SuperNIC sets the fabric your teams operate for years. Choosing the fabric built for message rate and lossless delivery protects time to solution and accelerator utilization, and open software keeps you free to choose accelerators.

Sources: 1. CN6000 SuperNIC page · 2. CN6000 Omni-Path Switch product brief · 3. Congestion control architecture · 4. CN5000 family page · 5. CN6000 family page

3 Size the fabric: pick tiers, ratio and hops

Omni-Path fabrics use fat-tree topologies.[1] Your endpoint count at full build picks the tier count, and your traffic pattern picks the subscription ratio. The table uses CN5000 building blocks: the 48-port, 400G CN5000 Switch and the 576-port Director Class Switch.

CN5000 design	Max endpoints at 400G	Notes
1 tier: one CN5000 Switch	48	Lowest latency; one hop[2]
1 tier: one Director Class Switch	576	Internal 2-tier fat tree; minimal cabling[1]
2 tiers, 1:1 non-blocking	1,152	24 down and 24 up per leaf[3]
2 tiers, 2:1 oversubscribed	1,536	32 down and 16 up per leaf[4]
3 tiers, 1:1, directors above the edge	13,824	Fewer discrete switches and cables[5]

Planning on CN6000 switching? The CN6000 Omni-Path Switch offers 48 ports at 800G or 96 ports at 400G in 1U.[7] Ask Cornelis for CN6000 sizing and for designs beyond this table.

Choose a subscription ratio

- **Run 1:1:** When MPI collectives or all-reduce span the whole fabric, or when you can't yet prove traffic stays local.
- **Consider 2:1:** When most traffic stays inside a leaf. The main gain is fewer switches and cables, especially next to a small 3-tier design. Each leaf also serves about 33% more endpoints.[4]
- **Avoid deep ratios for latency-bound jobs:** Queuing at the leaf adds delay that slows every rank waiting at a sync point.[4]

Count the hops jobs really take

- **Know the worst case:** In a symmetric fat tree, traffic between leaves crosses three switches in 2 tiers and five in 3 tiers.
- **Know switch latency:** Under 100 ns for CN5000 switches[6]; as low as 105 ns for the CN6000 Omni-Path Switch.[7]
- **Count the links too:** Every cable adds latency on top of the switches. Ask Cornelis for measured link latency for your design.

DECISION RULE: SCHEDULE FOR LOCALITY

Worst-case hop counts apply only to traffic that crosses the spine. Many HPC and AI jobs never take those paths when the scheduler places ranks that talk most under the same leaf. Pair your topology with topology-aware scheduling before you pay for more tiers or a lower ratio.

WHY IT MATTERS

Tier count drives switch count, cables, power and latency. Picking the smallest design that fits your full-build size, at the ratio your traffic needs, keeps cost per endpoint down without trading away time to solution.

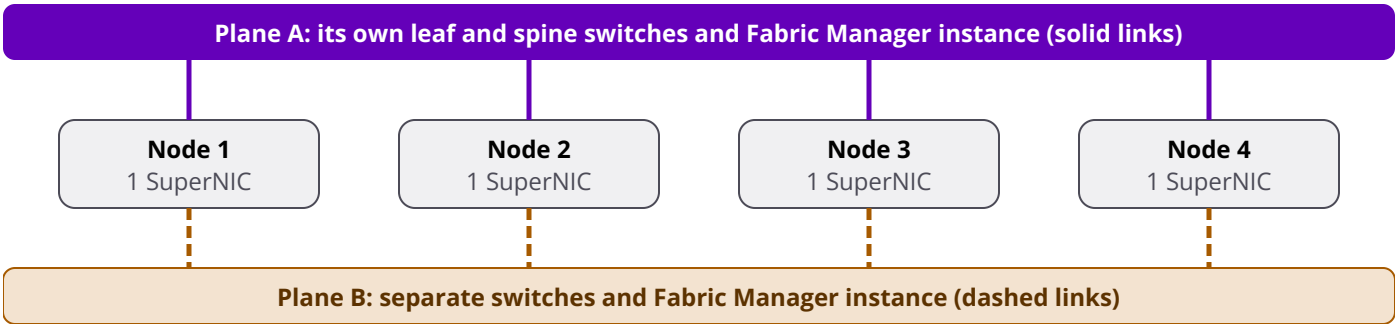
Sources: 1. Supported topologies · 2. 1-tier trees · 3. 2-tier fat trees · 4. Oversubscription · 5. 3-tier fat trees · 6. Switching, hops and latency · 7. CN6000 Omni-Path Switch product brief

4 Design for 800G: multi-plane and multi-rail

Multi-plane designs shrink fault domains and keep each plane at full 400G radix. Use them when long jobs must survive a switch fault, or to keep storage traffic off the compute path.

- **Multi-plane:** Physical design. Each node connects to two or more independent switch fabrics, called planes.
- **Multi-rail:** Software use. The host drives more than one link at once for load balancing and failover.[1]
- **Rail-optimized:** Placement. Accelerator *n* in every node connects to rail *n*, so collectives between same-index accelerators stay on one rail.

The diagram shows four nodes on two planes. Each node has one SuperNIC with two links: one up to plane A and one down to plane B. A fault in one plane leaves the other running.



Pattern	When to use it	What you gain
Each link to its own plane	Default for fabrics larger than one switch	A fault stays in one plane, and each plane keeps 400G radix and hop count
Both links to one switch	Small single-switch fabrics or sparse trees	Simple cabling; avoid it in larger trees
Two SuperNICs per server	Nodes that need more links than one SuperNIC provides	Failover and load balancing across rails[1,2]

Build it with CN6000

- **SuperNIC:** One dual x4-lane QSFP-DD port carries 800G per node.[3] Each 400G link can go to its own plane.
- **Switch:** The CN6000 Omni-Path Switch offers 48 ports at 800G or 96 ports at 400G in 1U.[4]

Map switches and managers to planes

- **One switch, two planes:** The CN6000 Omni-Path Switch holds two independent 48-port 400G planes, and one cable per node reaches both.[4]
- **One Fabric Manager instance per plane:** Give each a unique subnet prefix. One management server can host several instances.[1,5]

WHY IT MATTERS

A single plane is a single failure domain: one switch or optics fault can stall a synchronized job across the cluster. Two planes contain the fault and keep the second plane working, while each plane keeps the radix and hop count of a 400G fabric.

Sources: 1. Multi-rail topologies · 2. Dual rails for dual planes · 3. CN6000 SuperNIC page · 4. CN6000 Omni-Path Switch product brief · 5. Configuring more rails

5 Worked examples: two designs, with the math

These examples apply the rules in section 3 to the two most common sizing questions. Use them as a pattern, then confirm the design with your partner. Counts follow the CN5000 sizing rules; power ranges come from the product briefs.

Example 1: up to 576 endpoints, one director or leaf and spine

COMMON SIZE

Many designs land at 576 endpoints or fewer. In that range, choose between one Director Class Switch and a leaf-and-spine fabric of CN5000 Switches. Both run non-blocking at 400G.[1,2] Up to 48 endpoints, one CN5000 Switch is enough.

Endpoints	Director	Director cables	Leaf and spine switches	Leaf-spine cables
96	2 leaf modules	96	4 leaves + 2 spines = 6	192
192	4 leaf modules	192	8 leaves + 4 spines = 12	384
288	6 leaf modules	288	12 leaves + 6 spines = 18	576
384	8 leaf modules	384	16 leaves + 8 spines = 24	768
576	12 leaf modules	576	24 leaves + 12 spines = 36	1,152

Pick the director when

- **Nodes sit near one core:** One chassis, half the cables and no external inter-switch links.[2]
- **Growth stays under 576:** Add ports in 48-port steps.[3]
- **The site fits it:** Plan a 30-inch-wide rack, port-to-fan airflow and 544 lb (air-cooled).[4,5,6] Up to 21.9 kW typical, fully configured.[4]

Pick leaf and spine when

- **Racks spread across rows:** Leaves sit in the compute racks, so node links stay short.
- **You may pass 576:** Size spines for the end state; 24 spines reach 1,152.[1]
- **Power at 576 endpoints:** 36 switches draw 25.6 to 40.1 kW typical, air-cooled.[7]

Example 2: 2,048 endpoints, directors above the edge

3-TIER RANGE

Past 1,152 endpoints at 1:1, a fabric needs a third tier. Directors as the upper tiers keep the switch and cable count down.[3]

Non-blocking design at 400G

- **Layout:** 86 edge switches with 24 endpoints and 24 uplinks each. Four directors take 6 uplinks per edge (516 ports each).[3]
- **Counts:** 86 switches, 4 directors, 2,048 node cables, 2,064 inter-switch links.
- **Typical power:** 61.1 to 95.9 kW for edge switches, plus up to 87.6 kW for four directors.[4,7]
- **Ceiling:** Up to 13,824 endpoints with directors above the edge.[3]

Lower-cost option if traffic stays local

- **Layout:** Two tiers at 2:1 with 200G endpoints: 32 leaves of 64 endpoints and 16 spines.[8]
- **Counts:** 48 switches and 512 inter-switch links.
- **Typical power:** 34.1 to 53.5 kW, air-cooled.[7]
- **Placement:** Endpoints must sit close to their leaf. Ask Cornelis for the limits.

HOW TO CHOOSE Pick the non-blocking design when collectives span the whole fabric and every endpoint needs 400G. Pick the 2:1 option when jobs stay within a leaf and 200G per endpoint is enough.

Sources: 1. 2-tier fat trees · 2. Supported topologies · 3. 3-tier fat trees · 4. CN5000 Director Class Switch product brief · 5. Space and clearance requirements · 6. Director Class Switch installation · 7. CN5000 Switch product brief · 8. Oversubscription

6 Fit the fabric into the system: host, software and management

The fabric is one part of the system your OEM or integrator builds. Plan it with the servers, accelerators, software stack and management network, so the partner can validate the whole rack before it ships.

Host and accelerators

- **Slot:** CN6000 SuperNICs need a PCIe 6.0 x16 slot and fit a low-profile bay.[1] CN5000 SuperNICs use PCIe 5.0 x16.[2]
- **GPU data path:** Plan GPUDirect for NVIDIA and AMD accelerators.[1]
- **BIOS and Linux:** Apply the documented settings: turbo on, performance governor, IOMMU passthrough.[3]
- **Same build everywhere:** Install the same OS and kernel version on every host.[4]

Software stack (Omni-Path)

- **Transport:** The OPX libfabric provider runs MPI, OpenSHMEM, NCCL and RCCL.[5]
- **Other interfaces:** Verbs and IPoIB are available for storage and IP traffic.[3]
- **Operating systems:** Red Hat Enterprise Linux, Rocky Linux, SUSE Linux Enterprise Server and Ubuntu. Check the OPX release notes for versions.[4]
- **Containers:** Kubernetes v1.28 or later with the cn-fabric-k8s plugins, plus a separate Ethernet control-plane network.[6]

Fabric management

- **Fabric Manager:** Discovers the fabric, programs routes and collects performance data in band.[7]
- **Redundancy:** Run a standby Fabric Manager on a second management node.[8]
- **Prerequisites:** Plan NTP for all hosts and switches, and passwordless root SSH from the management node.[4]
- **Device management:** Switches and SuperNICs support Redfish and PLDM.[1,9]

Ethernet SuperNIC

- **Switches:** Works with standards-based 400G and 800G Ethernet switches.[11]
- **Interface:** Applications use standard Verbs.[10]
- **Network setup:** Configure the switches for lossless RoCEv2 with Priority Flow Control.
- **Operating systems:** Red Hat Enterprise Linux, with others on request.[11]

WHY IT MATTERS

Settling slots, software and management up front lets your partner validate the full system before it ships. That shortens time to production and lowers deployment risk.

Sources: 1. CN6000 SuperNIC page · 2. CN5000 SuperNIC (docs) · 3. Performance tuning quick start · 4. Software requirements · 5. CN5000 family page · 6. Running Kubernetes on a CN5000 fabric · 7. Fabric Manager overview · 8. CN5000 Fabric Installation Guide · 9. CN6000 Omni-Path Switch product brief · 10. CN6000 family page · 11. CN6000 Ethernet SuperNIC product brief

7 Plan cabling: set the rules with your partner

Your OEM or integrator selects cable types, lengths and routing for your racks. Cornelis shares detailed cabling guidance with partners and customers under NDA. Your job is to make sure the cable plan fits the topology you chose and the site you have.

Agree on these with your partner

- **Who owns the cable plan:** Your partner designs and supplies it, using Cornelis cabling guidance.
- **Rack layout first:** Where leaves, spines and directors sit sets every cable run, so settle the floor layout early.
- **Qualified cables only:** Use cables that Cornelis has qualified for your switches and SuperNICs.

Bring your site rules

- **Cooling:** Immersion or liquid cooling can limit which cables you can use.
- **Jackets and codes:** Low smoke zero halogen (LSZH) jackets and local building codes.
- **Floor:** Raised floors and air plenums can limit cable routes.

ASK FOR CABLING GUIDANCE UNDER NDA

Request the Cornelis cabling guidance through your partner or Cornelis sales before the system bill of materials is final. It covers the cable choices and limits your design must follow.

PLAN THE CABLES IN A FILE, NOT ON A WHITEBOARD

1. Fill in the sample cable-planning spreadsheet installed on the management node, with unique names for every host and switch.[1]
2. Convert it to a topology file with `opaxlattopology`.[2]
3. After install, check the running fabric against that file with `opareport -o verifyall`. Missing or misplaced cables show up before jobs do.[2]

WHY IT MATTERS

Cables are often the most common source of hardware problems, and a bad one can look like a failed switch.[3] Agreeing on the cable plan early lets your partner build the topology you designed, and a checked cable file shortens bring-up.

Sources: 1. Fabric design prerequisites · 2. Verify topology · 3. Verify cables

8 Plan power and cooling: size the whole rack

Fabric switches and SuperNICs draw on the same rack power, cooling loops and floor space as your compute. Use these figures to size circuits and choose air or liquid cooling with your partner.

Device	Typical power	Cooling and input options
CN6000 Omni-Path SuperNIC	45 W with copper; up to 60 W with optics[10]	Air-cooled or conduction-cooled[1]
CN5000 SuperNIC	About 24 W single port; 27.5 W dual port, without optics[2]	Air-cooled or conduction plate; bare board for custom OEM cooling[2]
CN6000 Omni-Path Switch (1U)	856 W copper; 1,663 W optical[3]	Conduction-cooled; 240 to 277 VAC or 48 VDC; N+N hot-swap supplies[3]
CN5000 Switch	Air 710 to 1,115 W; liquid 440 to 840 W[4]	Air or liquid; port-to-fan or fan-to-port; AC or 48 VDC[4]
CN5000 Director Class Switch	21.9 kW air; 21.1 kW liquid, fully configured[5]	Air or liquid; redundant hot-swap supplies and fans[5]

Air-cooled racks

- **One airflow direction per cabinet:** Switch fans are not reversible. A grey tab marks port-to-fan; black marks fan-to-port.[6]
- **Director airflow:** Port-to-fan only, so set the hot aisle to match.[7]
- **Stop recirculation:** Close side gaps, top and bottom transitions and cable passthroughs.[6]

Liquid-cooled racks

- **Flow and temperature:** 1.5 to 3.0 LPM per switch; inlet no higher than 45 °C; 3 to 5 °C rise.[8]
- **Pressure:** No more than 3 bar at the system connection. Bypass switches during high-flow CDU purges.[8]
- **Coolant:** Deionized water, PG25 or ethylene glycol blends. Avoid biocides with amines, chlorides or sulfur.[8]

Environment

- **SuperNICs:** 0 to 55 °C operating.[1]
- **CN5000 switches:** 10 to 35 °C operating, derated above 900 m.[4]
- **Humidity:** Non-condensing; check each product brief for limits.

Racks and floor

- **Racks:** Four-post, 19-inch, grounded, with dual power distribution and anti-tip devices.[7]
- **Directors:** Use a 30-inch-wide rack; 45U for two directors.[9] Plan floor load: 544 lb air, 590 lb liquid.[5]

DECISION RULE: AIR OR LIQUID SWITCHES

If your compute racks already have liquid loops, use liquid-cooled switches. A liquid-cooled CN5000 Switch draws 440 W typical with copper, against 710 W air-cooled.[4] That saves power at the switch and in the room.

WHY IT MATTERS

Power and cooling limits often decide how many nodes fit in a hall. Every watt the fabric saves can go to compute, and matching cooling to the rack avoids late redesigns.

Sources: 1. CN6000 SuperNIC page · 2. CN5000 SuperNIC (docs) · 3. CN6000 Omni-Path Switch product brief · 4. CN5000 Switch product brief · 5. CN5000 Director Class Switch product brief · 6. Air-cooled equipment requirements · 7. Space and clearance requirements · 8. Liquid-cooled equipment requirements · 9. Director Class Switch installation · 10. CN6000 Omni-Path SuperNIC product brief

9 Plan growth and migration: expand cleanly

Clusters grow in phases and inherit older fabrics. Plan the full-build fabric now, install only what phase one needs, and leave clean steps for the next phase.

Growth rules

- **Set the tier count for the final size:** Moving from 2 to 3 tiers later means recabling the core.
- **Size spines for the end state:** Spine ports cap how many leaves you can add (example 1).
- **Grow directors in steps:** Add ports in 48-port steps, up to 576.[1]
- **Grow 3-tier fabrics by pod:** Add pods while the core has free ports.[1]
- **Add a plane:** A second plane adds bandwidth and fault containment.[2]

WHERE THE FABRIC IS GOING (VISION)

Cornelis has announced Active Compute Fabric: an open architecture across scale-up and scale-out that adds lossless transport, in-fabric acceleration and programmable compute.[3] CN7000, the next generation, extends the roadmap to in-network compute.

Roadmap plans can change. Base today's design on shipping CN5000 and CN6000 products, and ask Cornelis how your design carries forward.

Pick your migration path

Moving from	Path	Watch for
CN5000 Omni-Path	Connect CN6000 Omni-Path SuperNICs to your CN5000 infrastructure.[4]	Confirm the OPX release supports both generations. CN5000 boards come in Version I and II, with different port mapping.[5]
Earlier Omni-Path (100G)	Build a new CN5000 or CN6000 fabric and keep your Omni-Path operations practices.	Ask Cornelis about running old and new fabrics side by side.
InfiniBand	Move MPI, OpenSHMEM, NCCL and RCCL jobs to the OPX libfabric provider.[6] Use Verbs and IPoIB for other traffic.[7]	Rebuild libfabric with CUDA or ROCm support for GPU jobs.[8]
Ethernet with RoCE	Cornelis also offers the Ethernet SuperNIC for your standards-based switches; applications use standard Verbs.[9]	Priority Flow Control settings on your switches.

CUT OVER IN PHASES

1. Stand up the new fabric as its own partition beside the old one.
2. Port and run your yardstick jobs, and compare results with the old fabric.
3. Move workloads queue by queue, then retire old segments.
4. Plan each SuperNIC for one fabric. Don't design an adapter to run Omni-Path and Ethernet at the same time.

WHY IT MATTERS

A growth plan lets you buy phase one without stranding it later. Phased migration keeps production running while you prove the new fabric on your own jobs, which lowers risk and protects the schedule.

Sources: 1. 3-tier fat trees · 2. Multi-rail topologies · 3. Active Compute Fabric release (Sept 2026) · 4. CN6000 SuperNIC page · 5. CN5000 SuperNIC (docs) · 6. CN5000 family page · 7. Performance tuning quick start · 8. CN5000 Fabric Installation Guide · 9. CN6000 family page

10 Deploy and accept: from site prep to sign-off

Follow the documented bring-up order, and verify each stage before moving on. Your partner usually leads installation. Use this page to agree on who checks what, and when.

Bring-up sequence

1. **Prepare the site:** Racks, power, cooling and the cable plan.[1]
2. **Install hardware:** SuperNICs, switches and directors.[1]
3. **Install host software:** Same OS and kernel on every host; OPX software from the Customer Center.[2,3]
4. **Start the Fabric Manager:** Run it from a node attached to the newest switch generation; add a standby.[1]
5. **Verify:** Check servers, Fabric Manager sweeps, cables and topology once all ports are active.[4]
6. **Tune:** Apply BIOS, Linux and MPI settings from the tuning guide.[5]
7. **Accept:** Re-run the named tests from your RFI/RFP checklist on the installed system.

Who does what

- **Your OEM or integrator:** Builds, validates and supports the system, fabric included.[6]
- **Cornelis:** Reviews the fabric design and provides software, firmware and documentation.[3]
- **Your team:** Owns the yardstick jobs and the acceptance results.

Test before you buy

Run your yardstick jobs on partner-hosted CN5000 clusters before the order is final.

Head off common bring-up problems

Symptom	What to do
Missing or misplaced links	Compare the fabric with your topology file using `opareport -o verifyall`.[7]
Slow Fabric Manager sweeps on large fat trees	Switch the routing algorithm from shortestpath to fattree.[8]
Inconsistent benchmark results	Find overloaded hosts with `opacheckload` before you blame the fabric.[4]
A link running below full width	Dynamic lane scaling keeps a three-lane link up by default. Find and replace the faulty cable or part.[9]
Changing fabric settings after go-live	Stop the Fabric Managers and running jobs before you change routing or buffer settings.[10]

WHY IT MATTERS

Each problem caught during bring-up is one that doesn't stop a production job. A checked, documented install shortens time to first useful work and gives your team a baseline to compare against later.

Sources: 1. CN5000 Fabric Installation Guide · 2. Software requirements · 3. Cornelis Customer Center · 4. Fabric verification · 5. Performance tuning quick start · 6. Blog: OEM-validated fabrics reduce deployment risk · 7. Verify topology · 8. Verify the fabric · 9. Dynamic lane scaling · 10. Multi-rail parameters and impacts

Your design at a glance: ten rules to act on

Use this page as the summary for your design review. Each rule points back to the section with the detail.

Decision	Rule of thumb
1. Workloads	Design for the most demanding workload that shares the fabric.
2. SuperNIC	Lead with the Omni-Path SuperNIC. Cornelis also offers the Ethernet SuperNIC when a design must stay on Ethernet.
3. Size and topology	Pick tiers for the final size, run 1:1 unless traffic stays local, and schedule for locality.
4. Multi-plane	Send each CN6000 link to its own plane to contain faults and keep 400G radix.
5. Examples	Up to 576 endpoints, weigh one director against leaf and spine; beyond that, put directors above the edge.
6. Integration	Settle slots, OS, software and management before the partner quotes.
7. Cabling	Let your partner own the cable plan, ask for cabling guidance under NDA, and keep a checked cable file.
8. Power and cooling	Match switch cooling to the rack; liquid switches cut typical power.
9. Growth and migration	Install phase one, but cable and size the core for the end state.
10. Deploy and accept	Verify each stage, then re-run your named tests on the installed system.

Choose your next step

IF YOUR DESIGN IS DRAFTED

Request a design consult

A Cornelis architect reviews your design with you and your partner.

IF YOU WANT PROOF FIRST

Reserve a test drive

Run your yardstick jobs on partner-hosted CN5000 clusters.

IF A PARTNER BUILDS YOUR SYSTEM

Share this guide

Give your OEM or integrator the design rules and your choices.

Further reading

- **Topology and routing:** CN5000 Topologies and Routing Guide
- **Installation and site preparation:** CN5000 Fabric Installation Guide
- **Congestion:** White paper: Congestion-free networking
- **Utilization:** White paper: Faster pipes, stalled clusters
- **Open standards:** White paper: Open standards and the cost of a locked-in network
- **Cluster design:** Blog: A network-centric approach to cluster architecture
- **Deployment risk:** Blog: OEM-validated fabrics reduce deployment risk
- **GPU networking:** Blog: Benchmark GPU-to-GPU networking with CN5000 and MI300X
- **Standards:** Ultra Ethernet Consortium and OpenFabrics libfabric
- **Market research:** Intersect360 Research: Networking and Interconnects

Bring your design to a Cornelis architect

Email **sales@cornelisnetworks.com** with your workloads, scale and constraints. We will review the fabric design with you and your partner.

Request a design consult