AGORA

The hidden half of a GPU cluster

A walk through the networks behind an AI cluster, and what the choice of fabric, optics, cooling and power means for the rack, the hall and the data centre that holds it all.

Summary

Most of the attention on AI infrastructure goes to the GPU, however a large part of a working cluster is made up of the networks, optics, cooling and power around it, and these determine whether the cluster performs as intended, what it costs to run and what kind of data centre is required to house it. This paper acts as guide to building a cluster: it goes through each of these in turn, starting with the compute tray and working out through the rack, the networks, the optics, the cooling, the power and the hall, and for each, sets out the main configurations and how they affect the design of the building. We've used one worked example throughout, a 10 MW cluster of the standard 72-GPU racks of 2026 at a typical US site, and have then built the same design at other sizes, from a single rack up to a 100 MW campus, to show how it changes as a cluster grows.

In a nutshell, the choices made around the GPU have a much larger effect on the building than the GPU itself, as in the example the networks alone require 739 switches and over 38,000 optical transceivers and draw 1.3 MW that's often overlooked in a rack's power budget. Therefore a cluster specified as 10 MW requires 11.3 MW once its networks are added. The choice of fabric moves this network load between about 8 and 13 percent of the IT load, as InfiniBand's larger switches avoid the third tier of switching that the Ethernet design requires, whilst co-packaged optics remove that third tier on Ethernet as well and bring the load down to around 7 percent. At 135 kW, each rack must be liquid cooled, and the temperature of that water determines whether the heat is removed from the building by compressors or by dry coolers, whilst the power chain feeding each rack is already being redesigned around 800 volts DC for the next generation of hardware.

The rack that holds all of this weighs 1.5 tonnes and loads the floor at about 1.7 times what a typical hall slab is designed for, therefore the building needs to be designed with the rack in mind from the start. This paper end with what this means for anyone buying or financing a cluster.

1. Introduction

AGORA designs, deploys and finances GPU clusters across the US, and the main thing we've learnt from that work is that the GPU is the easy part of a cluster. It arrives with a datasheet, a price, a power figure and a lead time, whereas the vital components required to make a working cluster are chosen through decisions that no-one outside the engineering team sees, such as which of four networks a cable belongs to, whether the fabric is InfiniBand or Ethernet, where the leaf switches can be physically located, how many fibre cores run under the floor, whether the floor can hold the rack and how the heat leaves the building. These decisions account for most of the non-GPU cost of a cluster, most of the schedule and most of the problems during deployment, however they are rarely discussed outside the engineering teams that make them.

This paper is an introduction to those decisions, working from the GPU outwards one layer at a time, and for each layer setting out what the part is, what the options are and what each option does to the rest of the data centre. To keep the numbers consistent, we've used one worked example throughout, a 10 MW cluster of 72-GPU rack-scale systems at a typical US site with its network built to the GPU vendors' published reference architectures, which our engineering team has put together

for this paper as an illustration[1]. It's a standard design for 2026 and the kind of cluster a well-funded operator is specifying this year, however many clusters are bought well below this size, therefore section 10 sets the same design out at other sizes, from a single rack in a colocation cage to a 100 MW campus, so that each reader can find the scale that applies to them.

2. The tray and its four networks

The best place to start is the compute tray, a single rack unit holding four GPUs and two CPUs under cold plates, as this is where the whole cluster connects, with four separate networks leaving the front of the tray, each with its own cables and switches and each carrying a different type of traffic[2]. The first is scale-up, the scale-up domain that joins the 72 GPUs within a rack, and the second is scale-out, the fabric that joins racks to each other so that a job can run on more than 72 GPUs. The third is the frontend and storage network, which carries data in and out of the cluster and to the storage nodes through a DPU on each tray, and carries over one set of cables what older designs ran as separate storage and in-band management networks, whilst the fourth is out-of-band management, a small separate network that lets the operators monitor and control every piece of hardware even when the rest of the cluster is down.

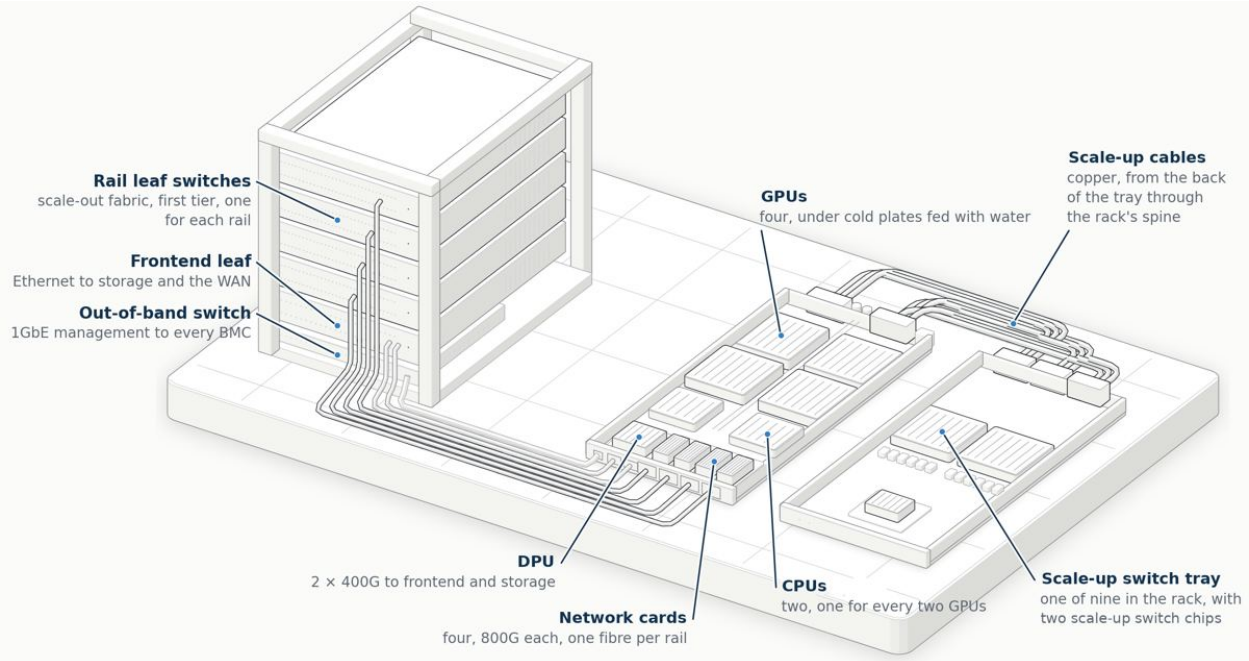

Figure 1. One compute tray with its lid off and the four networks that leave it: the scale-up links in copper to the switch trays, four 800G ports to the four rails' leaf switches (each port split across the two planes, drawn as one fibre), two 400G fibres from the DPU to the frontend leaf, and one 1GbE lead to the management switch.

These four networks are kept separate because the traffic on each one is different. Scale-up carries the heaviest and most constant traffic, as the GPUs that share a model exchange results at almost every step of the calculation, so it is kept inside the rack on short copper links with far more bandwidth than any of the other networks. The scale-out fabric carries the exchanges between racks during a training step, where thousands of GPUs send at once and every one of them waits for the slowest message to arrive, so it is built for bandwidth and for a predictable tail latency, as a step only

[1] AGORA worked example, October 2026, for illustration: Tier III on a temperate US site, 34.5 kV supply with 13.8 kV distribution and 480 V to the rows, and direct-to-chip cooling on 32 °C facility water. The network follows the GPU vendors' published scale-out reference architectures, with switch counts checked against 21 published configurations.

[2] Vendor product documentation for 2026 rack-scale systems: 72 GPUs and 36 CPUs in 18 compute trays and 9 scale-up switch trays, 1.8 TB/s of scale-up bandwidth per GPU and 130 TB/s across the domain; an 800 Gb/s (100 GB/s each way) network card per GPU; a DPU at 2 × 400G per tray in the example's frontend design.

finishes when the last packet lands and one slow link holds up every GPU in the job. The frontend carries data loading, checkpoints and user traffic, which arrive in large and unpredictable bursts and would stall a training step if they shared a network, whilst the management network has to keep working when all of the others have failed.

Network	Per compute tray	Medium	Where it goes	What it is for
Scale-up	4 GPUs on the spine	Copper, inside the rack	9 scale-up switch trays	One 72-GPU domain at 1.8 TB/s per GPU
Scale-out (compute fabric)	4 × 800G network cards, 8 × 400G links	Optical	Leaf of each rail, both planes	Jobs larger than one rack
Frontend and storage	2 × 400G DPU	Optical	Frontend leaf, storage, WAN	Data in and out, checkpoints
Out-of-band	1 × 1GbE to the BMC	Copper	Management switch	Control when all else fails

Table 1. The four networks of a compute tray in a 72-GPU rack.

3. The rack: why 72 GPUs share one cabinet

A 72-GPU rack-scale system is a single machine that has the dimensions of an entire rack, rather than a rack made up of servers. Inside it are 18 compute trays, nine scale-up switch trays, eight power shelves and two small management switches, with the GPUs joined into one scale-up domain of 72 by a copper spine at the back of the rack that carries about 5,000 cables and two miles of wire[3]. The rack is built this way because of how much bandwidth the scale-up links carry. Each GPU has about nine times the bandwidth to the other GPUs in its rack as its network card has to the rest of the cluster, and the only economic way to carry that much is over copper. Copper at 200 gigabits a lane reaches a couple of metres at most, so every GPU in the domain has to sit within that distance of the scale-up switches, which is what fixes the domain at one rack. Each GPU has 18 scale-up ports and the nine switch trays hold 18 switch chips between them, so every GPU has one link to every chip and any two GPUs in the rack are a single switch hop apart with no oversubscription, which is what lets the software treat the 72 GPUs as one pool of memory. This is why 72 GPUs and all their switching are packed into one cabinet, and why that cabinet draws as much power as a row of servers used to and needs liquid to cool it. The workloads that need the domain are the ones that split a single model across many GPUs, as tensor parallelism and the expert routing of mixture-of-experts models exchange data at every layer, and that traffic stays inside the rack, whilst the slower and less time-critical traffic between copies of the model goes out over the scale-out fabric.

[3] The spine's figures (about 5,000 cables and about two miles of copper) are the vendor's published figures for the previous rack generation, March 2024. OEM planning guides for 2026 72-GPU racks give the rack power (about 135 kW nominal, about 155 kW peak), weight and coolant conditions. The 1.8 TB/s of scale-up bandwidth is 900 GB/s each way, nine times the network card's 100 GB/s each way.

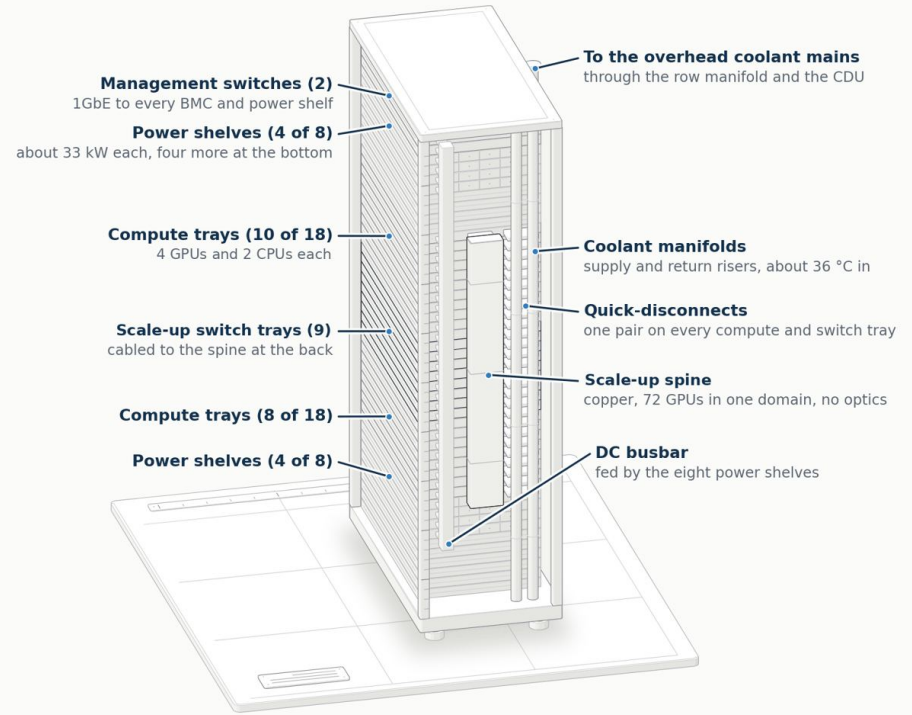

Figure 2. A 72-GPU rack from behind with its rear door and side panel off: the compute and switch trays, the copper scale-up spine, the coolant manifolds and the DC busbar.

For the data centre around it, the rack's most important properties are its power, its weight and its heat. It draws about 135 kW in normal running and can peak at about 155 kW, which is more than an entire row of enterprise racks, and a design normally sizes the electrical plant on the nominal figure and uses the power shelves' capping to hold the peak. It weighs about 1.5 tonnes (3,300 lb), which over its small footprint is 20 kPa (about 425 psf) against the 12 kPa (250 psf) a typical hall slab is designed for, so the slab is also required to be designed with this in mind. Ninety percent of its heat goes into water, with the coolant entering the cold plates at 36 °C and leaving at about 55 °C, whilst the remaining 10 percent still goes to air and still needs air handling in the hall[4]. Unlike a rack of servers, its power also varies with the workload, as a large training job moves the load of the whole cluster in step and most of the 10 MW can change within a second. Some of this swing is absorbed by the energy storage built into the rack's power shelves and by a ramp limit on the GPUs themselves, however the UPS, the generators and the grid connection all have to be sized for what gets past the shelves.

[4] AGORA worked example: 121.5 kW to water and 13.5 kW to air per rack; 3,485 litres (921 US gallons) a minute of rack coolant per hall at 36 to 55 °C; 1,500 kg over 0.72 m² is 20.4 kPa (426 psf). On synchronised power swings, vendor power-smoothing guidance for 72-GPU racks.

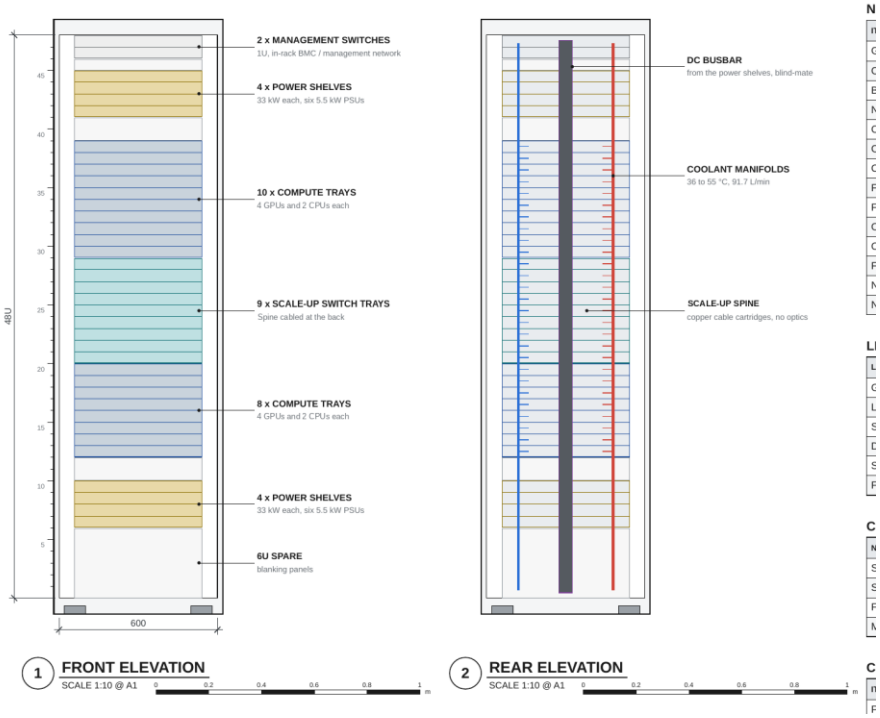

Figure 3. Front and rear elevations of the rack: the tray stack with its rack-unit positions, and behind it the busbar, the manifolds and the spine.

4. The fabric: joining the racks

The scale-out fabric is where most of the cluster's network equipment sits, and it is built as a rail-optimised fat tree, which means that the GPUs in every tray are numbered one to four, and all of the number ones connect to the leaf switches of rail one, all of the number twos to rail two and so on. This is done because of the pattern of the traffic, as the collective operations of a training job, the all-reduce of gradients and the all-to-all of expert routing, mostly pair a GPU with the same-numbered GPU on other trays, so rail alignment lets those exchanges meet on a single leaf switch wherever they can and only go up to the spines when required[5]. Each GPU has an 800-gigabit network card, and because the example's Ethernet switch has 128 ports of 400 gigabits, each card is split across two identical and independent fabrics, called planes, so that every GPU has a link into both. This split doubles the number of GPUs each tier of 400-gigabit switches can reach, and it also gives the cluster two separate failure domains, so that a fault in one plane halves a GPU's scale-out bandwidth rather than cutting it off altogether. Two tiers of these switches can join up to 4,608 GPUs, so the example's 5,472 GPUs need a third, which brings the compute fabric to 694 switches joined by nearly 33,000 links and adds two switch hops to the longest path between GPUs.

[5] Rail-optimised topology as described in the vendors' published scale-out reference architectures: GPU n of every node connects to the leaf switches of rail n; in the example the leaf takes two racks (36 ports) per rail at 1:1, a scalable unit is 8 racks (576 GPUs), and the example's 9.5 SUs need 152 leaves per plane, more than the 128 that one tier of spines can join, which forces a third tier. Ethernet switches at 128 × 400G, so two planes for an 800G card.

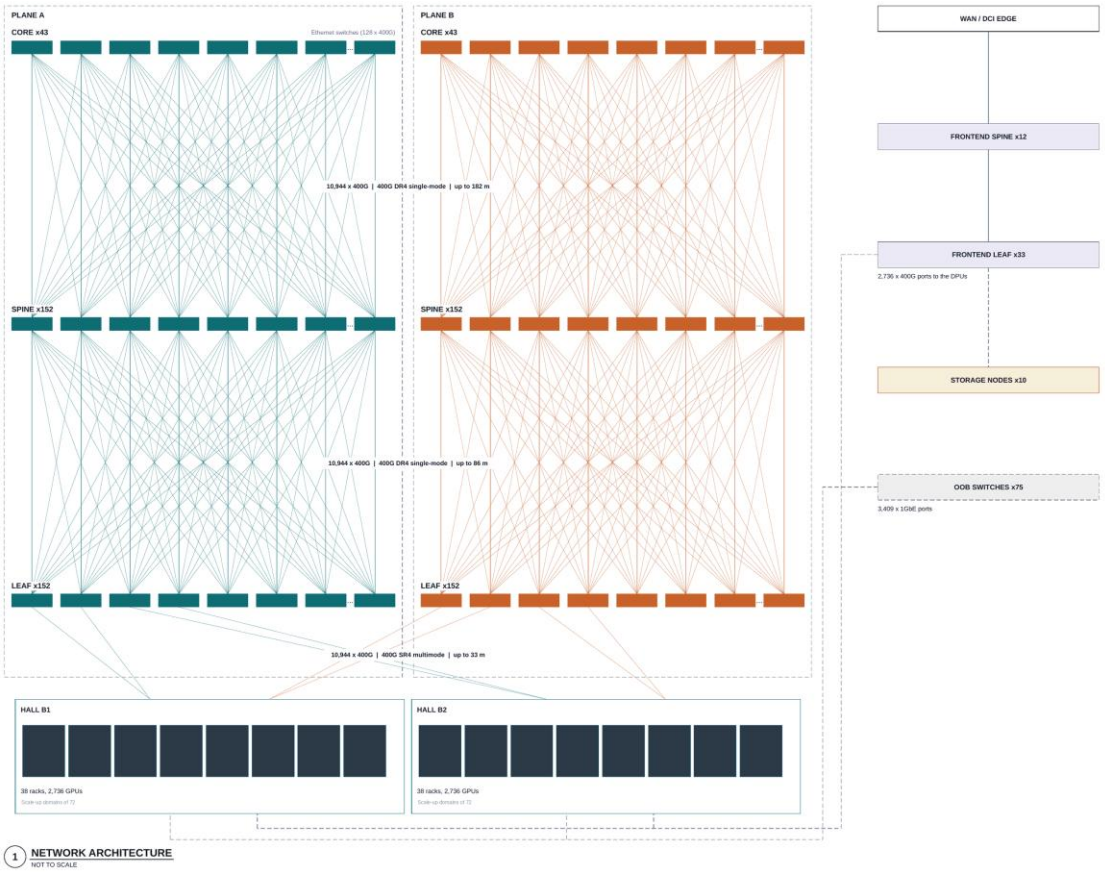

Figure 4. The network architecture of the 10 MW example: two planes of leaf, spine and core switches over the two halls, with the frontend, storage and management networks beside them.

Once the frontend and management networks are added, the cluster has 739 data switches and 75 small management switches, and the data switches alone fill 42 racks, which is 21 racks of networking in every hall of 38 compute racks. The switches and their optics draw about 1.3 MW, which is 12.7 percent of the IT load, so a cluster specified as 10 MW is in practice an 11.3 MW cluster once its networks are added[6], and how large that load becomes depends mainly on the choice of fabric and the choice of optics.

4.1 InfiniBand or Ethernet

InfiniBand and Ethernet are usually compared on the rules each uses to move data around a network, however for a GPU cluster the difference that matters is how the two behave when more traffic arrives at a switch than it can pass on. InfiniBand is lossless by design, as a switch only sends a packet when the next hop has told it there is buffer space (credit-based flow control), so nothing is dropped and each packet can be routed around a busy link, and its switches can also do part of a collective operation themselves, summing the gradients from many GPUs inside the network, which can roughly halve the traffic of the largest collective in a training step. Ethernet drops packets when its buffers fill, and RDMA over Converged Ethernet, the usual way to run GPU traffic over it, copes by pausing upstream links (priority flow control) and marking congested packets (ECN) so that senders slow down, which works well when the buffer thresholds and the senders' rate control are tuned together and badly when they are not, as a pause on one link can back up through the fabric. AI-tuned Ethernet platforms add adaptive routing and congestion control in the switch and the

[6] AGORA worked example: 746 kW of switches and 560 kW of optics, 1,306 kW in all against 10.3 MW of IT. Switch power is indicative (1.0 kW per 128 × 400G Ethernet switch, 2.5 kW per 144 × 800G InfiniBand switch excluding optics) and module power by bandwidth (9 W at 400G, 15 W at 800G, 25 W at 1.6T).

network card to close much of that gap, and the new Ultra Ethernet standard goes further by spraying the packets of a single message across every available path and resending only the ones that are lost, rather than everything after the lost one as most network cards do today.[7] In practice this means that InfiniBand's performance comes largely out of the box, whereas Ethernet's depends on how well it is configured and on the people who do the configuring, which is why we test every fabric we deploy under a full training load before it is handed over.

In the market Ethernet is winning on volume, taking about two thirds of data-centre switch sales into AI clusters in the first quarter of 2026 by Dell'Oro's count, however InfiniBand sales more than tripled in the same quarter as 800-gigabit switches shipped with the latest GPU generation[8]. For any given cluster the choice is first of all a choice about physical quantities, such as how many switches, transceivers and racks the fabric needs and how much power it draws, and Table 2 sets these out by building the same 5,472 GPUs four different ways, with the storage and frontend networks held constant.

Fabric	Switch	Planes	Tiers	Leaf / spine / core	Transceivers	Network power
Ethernet, 2 × 400G (default)	128 × 400G	2	3	304 / 304 / 86	38,415	1,306 kW, 12.7%
Ethernet, 800G	64 × 800G	1	3	304 / 228 / 86	38,415	1,230 kW, 12.0%
InfiniBand XDR	144 × 800G	1	2	76 / 72 / –	19,263	776 kW, 7.6%
Ethernet, 2:1 oversubscribed	128 × 400G	2	2	152 / 72 / –	21,999	589 kW, 5.7%

Table 2. The 10 MW example with four compute fabrics. Transceivers and network power include the frontend and management networks; the frontend and storage are the same in all four.

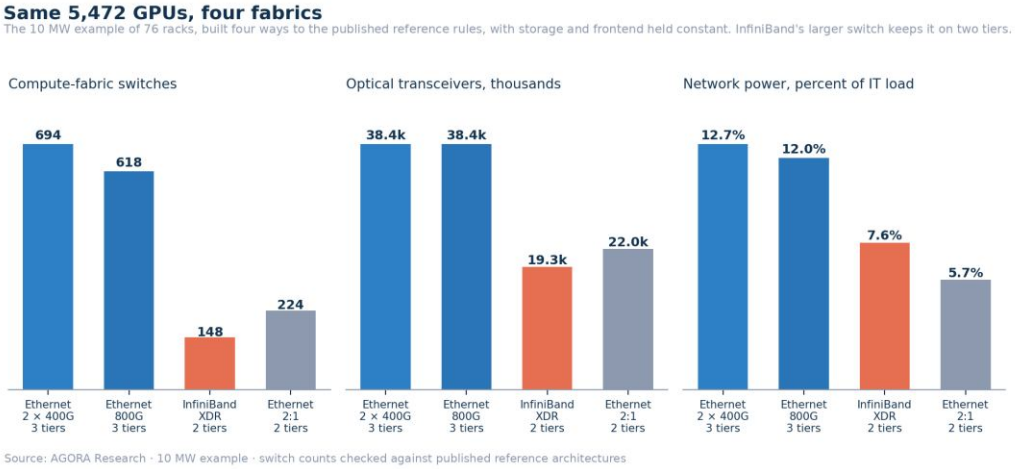

Figure 5. Switches, transceivers and network power for the four fabrics of Table 2.

The first thing Table 2 shows is the effect of switch size, as the InfiniBand switch has 144 ports of 800 gigabits against the Ethernet switch's 128 of 400, so the same fabric needs 148 InfiniBand switches against 694 Ethernet ones, which is nearly five times as many boxes, racks and points of failure, and most of that difference comes from the InfiniBand fabric staying on two tiers at a size

[7] Vendor product documentation for credit-based flow control, adaptive routing and in-network reduction on InfiniBand, and for adaptive routing and congestion control on AI Ethernet platforms. RoCEv2 relies on priority flow control and ECN-based congestion control to approximate a lossless fabric, with go-back-N recovery after a loss on most deployed network cards. Ultra Ethernet Consortium, "Ultra Ethernet Specification 1.0", June 2025; T. Emmons, "Demystifying Ultra Ethernet", Arista, 16 July 2025, on packet spraying, out-of-order delivery and selective retransmission.

[8] Dell'Oro Group, "Ethernet extends lead in AI scale-out networks despite strong InfiniBand rebound", 2 June 2026: Ethernet about two thirds of data-centre switch sales into AI clusters in 1Q 2026, Ethernet sales more than doubled, InfiniBand more than tripled with the ramp of 800 Gbps switches on the latest GPU generation; a large share tied to brownfield upgrades.

where the Ethernet fabric has had to add a third[9]. The second is that the transceiver count follows the number of tiers and the bandwidth of each module rather than which of the two is used, as the InfiniBand design has one tier fewer and each of its modules carries twice the bandwidth of one on the Ethernet switch, which together account for the 19,000 fewer transceivers in that design. The third is that oversubscription has the largest effect of the three, as allowing two gigabits of leaf traffic for every one of uplink (2:1 rather than the 1:1 of the reference design) brings the Ethernet fabric back to two tiers, cuts the switches by two thirds and takes the network's power from 12.7 to 5.7 percent of the IT load, however the reference compute fabrics are non-blocking because a training job's collective operations saturate the uplinks at exactly the moment every GPU is waiting on them, so a cluster that is cheaper to cable can lose more in training time than it saved. Our own practice is therefore to design the compute fabric non-blocking, to prefer the larger switch where a cluster is close to a tier boundary, and to treat the choice between InfiniBand and Ethernet as secondary to the questions of who will operate the fabric and what the client's other data centres run, as most of the cost of running a fabric is in the people who understand it.

5. The optics: what carries the bits

The links in the fabric are carried on copper or on optics, and the choice for each one is mostly a trade between the distance the signal has to cover and the power it takes to get it there. Passive copper carries 800 gigabits for two or three metres and draws almost nothing, which is why the scale-up spine is copper, whilst active electrical cables, which retime the signal inside their plugs, increase that to about seven metres, which is enough to join a rack to its neighbours and to a switch at the end of its row[10]. Everything beyond the row is therefore optical, and each pluggable optical module contains a DSP that retimes the eight 100-gigabit electrical lanes coming off the switch, lasers and photodiodes that convert them to light and back, VCSELs for the short multimode links and modulated or silicon-photonics transmitters for single-mode, and a connector for 8 or 16 fibres, and draws about 15 watts at 800 gigabits. Two newer approaches reduce this power, as linear pluggable optics leave out the DSP and roughly halve a module's power, however they require shorter and cleaner links and a switch whose own transmitters are good enough to drive the laser directly, whilst co-packaged optics move the optical engines onto the switch package itself, keep only the lasers in replaceable modules on the front panel and remove the pluggable module at the switch end altogether. Table 3 sets out the reach, power and place of each of these in a cluster.

[9] AGORA worked example: a 144 × 800G InfiniBand switch keeps a cluster of up to 18 SUs (10,368 GPUs, 144 leaves) on two tiers, which covers the example's 9.5 SUs, whereas 128 × 400G Ethernet switches in 400G planes reach only 8 SUs (4,608 GPUs) on two tiers. Table 2's four variants hold storage (10 nodes) and frontend (33 leaves, 12 spines) constant.

[10] Credo, "ZeroFlap SPAN 800G Active Electrical Cable", product brief, September 2025: up to 7 m; passive 800G copper cables run to about two to three metres. Module power: 800G DSP-based modules are typically 12 to 16 W and linear pluggable modules roughly half that (EDGE Optic, "LPO vs DSP optical transceivers", 2026, gives about 12 W and 4 to 6 W). Optic classes: SR (multimode) to about 50 m, DR (single-mode) to 500 m, FR to 2 km; 400G SR4 and DR4 at 8 fibres per link, 800G SR8 and DR8 at 16.

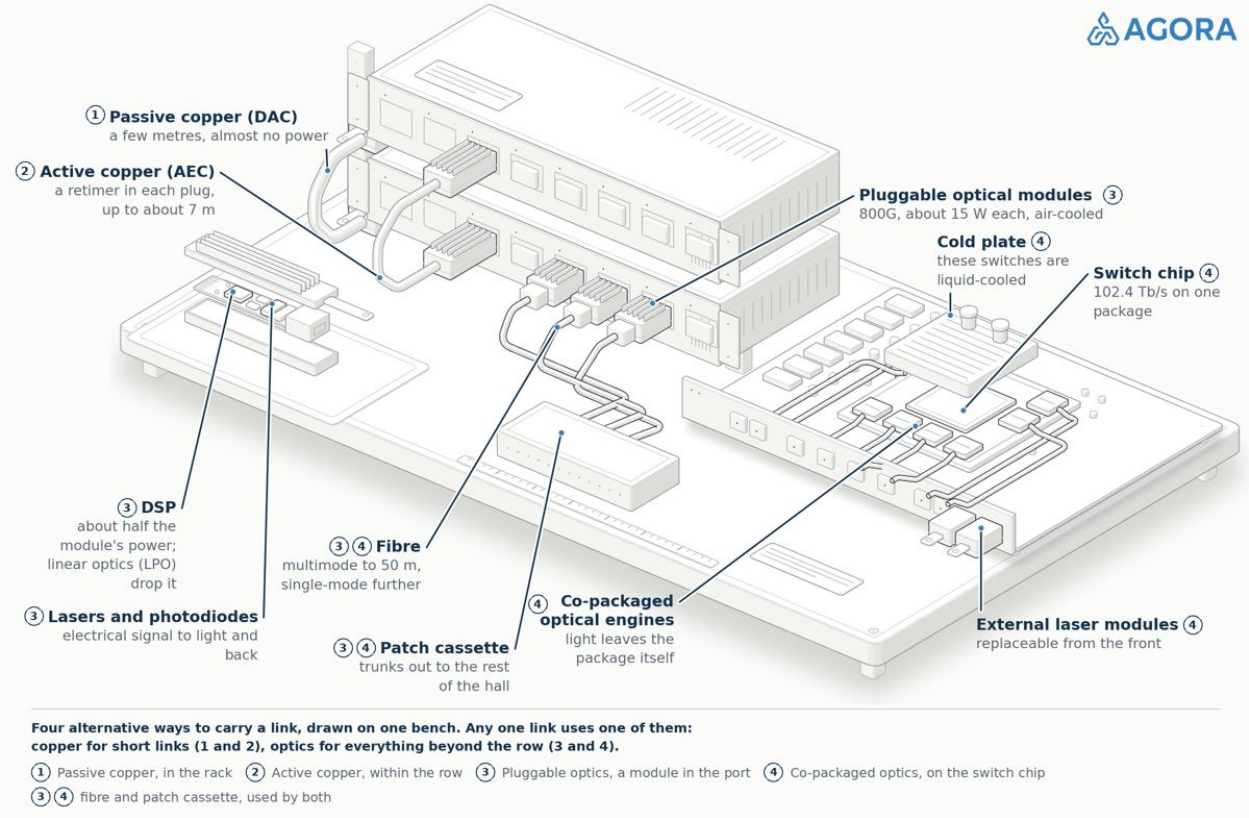

Figure 6. Four alternative ways to carry a link out of a switch, numbered on the drawing and shown together on one bench: two in copper for short links, 1 passive copper and 2 an active electrical cable, and two in optics for everything beyond the row, 3 pluggable optical modules, one opened to show its DSP and lasers, and 4 a co-packaged switch board whose optics sit on the chip package. The fibre and the patch cassette are shared by 3 and 4, and any one link uses one of the four.

Medium	Reach at 800G	Power at the link end	Where it goes in a cluster
Passive copper (DAC)	2 to 3 m (7 to 10 ft)	almost none	inside the rack: the scale-up spine, short in-rack links
Active copper (AEC)	up to about 7 m (23 ft)	a retimer in each plug	rack to rack within a row
Multimode optics (SR)	about 50 m (164 ft)	about 15 W a module	GPU to leaf inside a hall
Single-mode optics (DR, FR)	500 m to 2 km	about 15 W a module	leaf to spine, spine to core, between buildings
Linear pluggable optics (LPO)	short, clean links	about half a DSP module	where both ends can drive a linear link
Co-packaged optics (CPO)	as single-mode	about 3.5 times less (vendor figure)	on the switch package, no switch-end module

Table 3. The media that carry a cluster's links, with their reach, power and place.

The reach of each optic informs the design of the floor plan, as in the example the link from a GPU's network card to its leaf switch is about 33 metres (108 feet) and runs on multimode optics, which only works because the leaves stand inside the hall, whilst the links from leaf to spine and from spine to core are long enough to need single-mode optics, parallel DR8 modules on eight fibre pairs to 500 metres, and a campus of several buildings would push the top tier into the two-kilometre FR class, which multiplexes four wavelengths onto each pair and cuts the fibre count by three quarters. The leaves therefore have to be located in the hall, and in the example the spines and cores sit beside them in 21 network racks along each hall's south wall, with the fibre leaving each hall in trunk cables of 144 cores, nearly 300,000 fibre cores in all. Co-packaged optics change more than the power, as removing the switch-end modules frees up the switch's front panel and lets one package carry far more ports, so the first co-packaged switches carry two to eight times the bandwidth of the

pluggable switch in the example[11], and when the smaller of them is run as 256 ports of 400 gigabits it keeps the example's 5,472 GPUs on two tiers rather than three (Figure 7).

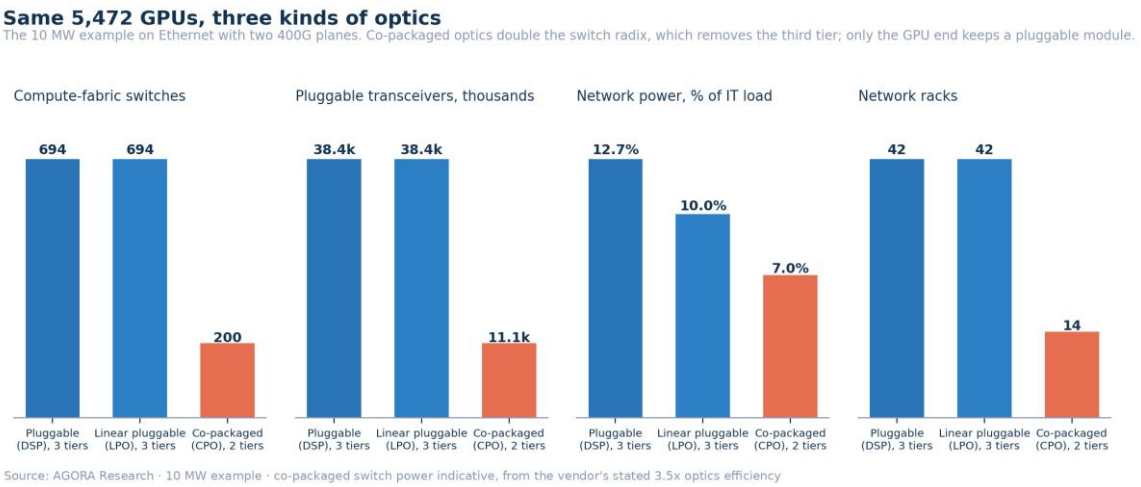

Figure 7. The 10 MW example on Ethernet with pluggable optics, linear pluggable optics and co-packaged optics: compute switches, pluggable transceivers, network power and network racks.

Co-packaged optics have the largest effect of any option in this paper, as the compute fabric falls from 694 switches to 200, the pluggable transceivers fall by more than two thirds to only the GPU ends and the frontend network, the network racks fall from 42 to 14 and the network's power from 12.7 to about 7 percent of the IT load, whilst linear optics keep the same switches and modules and save about 280 kW on the modules alone. There are several caveats to this, as co-packaged switches were only entering production in 2026, our switch power figure is an estimate built on the vendor's stated efficiency, the photonics switches bring liquid cooling into the network racks, and a failed optical engine means replacing a switch rather than a module, which is why the lasers have been moved out to the front panel. The optics are also where the price inflation on the networking side currently is, as the optics and cabling line of our networking price index is up about 25 percent in twelve months whilst switches are up by around 5 percent, and the pressure is in the lasers, as the indium phosphide wafers they are grown on come from two or three suppliers whose output is sold out into 2027, with much of it already tied up in multi-year contracts by the largest buyers. Lead times have followed, as a third-party 800-gigabit transceiver is an 8 to 12 week item on our lead-time series, about the same as the rack itself, whilst modules from the GPU vendors themselves were being quoted at over 50 weeks on some part numbers in September, and 38,000 of them is a large order in its own right, so on the clusters we deploy the optics are specified and ordered with the racks rather than after them.[12]

6. The cooling: getting 135 kW out of a rack

The newest racks have to be liquid-cooled, as water carries about 3,500 times as much heat as the same volume of air for each degree it warms, so taking 135 kW out of a rack by air would mean pushing about 14,000 cubic feet of air a minute through a single cabinet, whereas about 90 litres (24

[11] Co-packaged Ethernet and InfiniBand switches entered production in mid-2026, with general availability listed for the second half of the year: 102.4 Tb/s (128 × 800G) in 2U for the smaller Ethernet switch and 409.6 Tb/s (512 × 800G), liquid-cooled, for the larger, whilst the InfiniBand version carries 144 × 800G, liquid-cooled, with 18 removable light-source modules. The vendor puts the power-efficiency gain of co-packaged optics at 3.5 times. Example variants: the 102.4 Tb/s switch run as 256 × 400G in two planes at an indicative 2.6 kW per switch including its optics; linear optics at half the module power on links up to 500 m.

[12] AGORA pricing and lead-time data, October 2026; Fusion Worldwide, September 2026; LightCounting, August 2026; TrendForce, August 2026.

US gallons) a minute of water carries the 90 percent of it that goes to the cold plates[13]. At lower densities there is still a choice to be made, and Figure 8 sets the main options out side by side. Air with hot-aisle containment takes racks up to about 40 to 60 kW, with every watt going into the room and back through chilled-water air handlers, whilst a rear-door heat exchanger hangs a water coil on the rack's exhaust and takes about 80 percent of the heat at the rack, although it needs chilled water, which in most climates means chillers. Cold plates on the GPUs and CPUs take about 90 percent of the heat straight into a coolant loop, and that loop either rejects its heat to the room through liquid-to-air sidecars beside the racks, which needs no water from the building, or passes it through coolant distribution units (CDUs) to a facility-water loop at 32 °C (90 °F) or warmer, the CDU's heat exchanger keeping the racks' treated coolant separate from the building's water and its pumps holding the supply above the hall's dew point so that nothing condenses on the plates. Single-phase immersion sinks the servers in a dielectric fluid and takes about 95 percent of the heat to warm water, however the servers have to be built for it and it is not offered for 72-GPU racks.

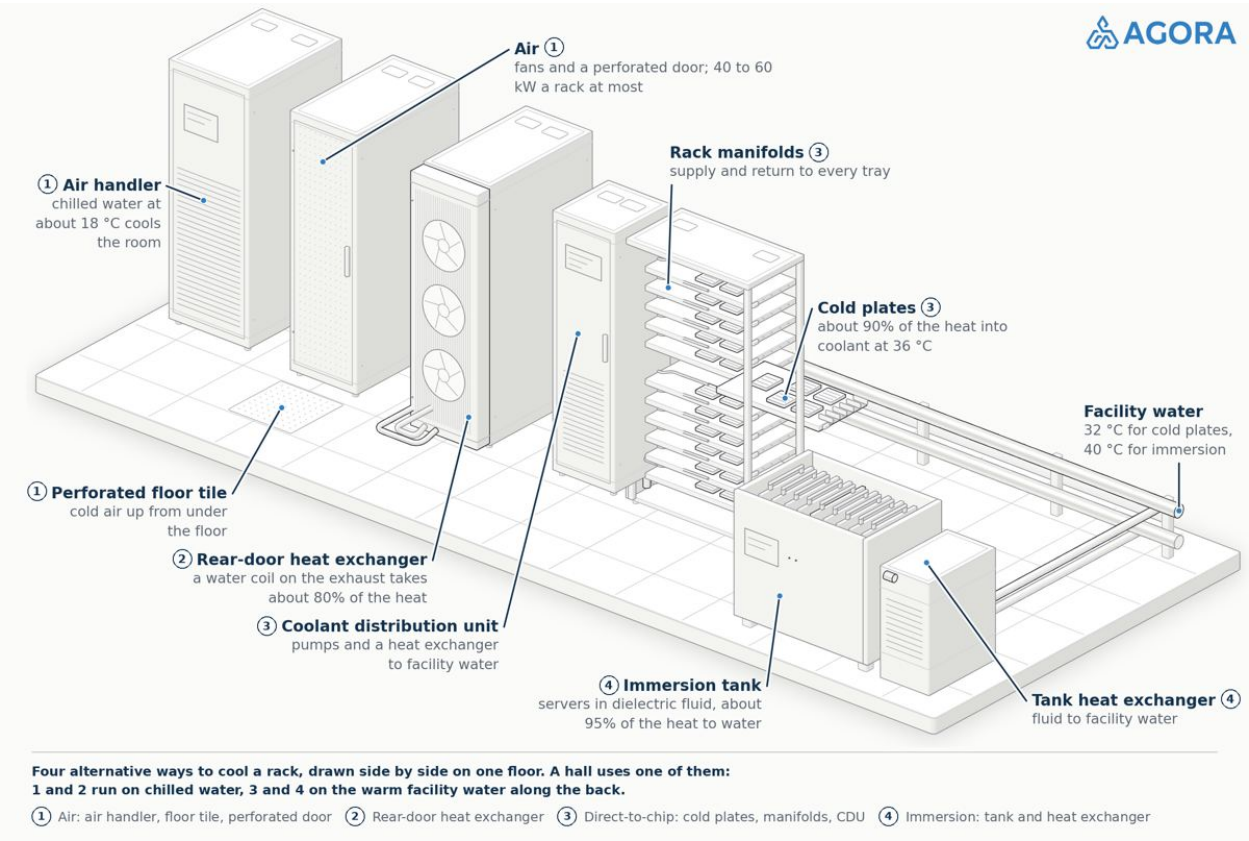

Figure 8. Four alternative ways to cool a rack, numbered on the drawing and shown side by side on one section of floor: 1 an air-cooled rack with its air handler and perforated floor tile, 2 a rear-door heat exchanger, 3 direct-to-chip cold plates with their manifolds and a coolant distribution unit, with one tray pulled out, and 4 an immersion tank with its heat exchanger. A hall uses one of them; 1 and 2 run on chilled water, whilst 3 and 4 use the warm facility water along the back.

Much of the difference between these options comes down to the temperature of the water, as chilled water at 18 °C is colder than the outside air for much of the year, so air, rear doors and sidecars all depend on chillers, whereas facility water at 32 °C (90 °F) is warmer than the outside air for most of the year on a temperate site and can be cooled by dry coolers, which are fans blowing outside air over coils. A dry cooler can only bring the water down to within four or five degrees of the outside air, its approach temperature, so on the hottest days, when the outside air is above about 27

[13]Volumetric heat capacity of water about 4.18 MJ per m³ per K against about 1.2 kJ per m³ per K for air at room conditions, a ratio of about 3,500. Air for 135 kW at a 17 K rise: 135 ÷ (1.2 × 1.005 × 17) ≈ 6.6 m³/s, about 14,000 cubic feet a minute; water in the example about 92 litres (24 US gallons) a minute per rack.

°C (81 °F), the water would come back too warm, and a small trim chiller runs alongside the dry coolers for those hours to take the water the last few degrees down, which on a temperate US site is about 350 hours a year. In the 10 MW example the facility water runs from 32 to 44 °C and the design PUE of 1.23 becomes about 1.08 over a year, with 7.2 MW of heat leaving the building at 44 °C that a district heating network or an industrial user could take[14]. The network has to be cooled as well, as the switches and optics in each hall's 21 network racks give off about 650 kW to the air rather than to the water, which is more than the hall's 38 compute racks still send to air, so the air handlers of a liquid-cooled hall have to be sized for the network as much as for the racks, and the co-packaged switches of section 5 would move much of that load into the water loop. Figures 9 and 10 show the two ends of that range at the scale of a whole hall and its plant.

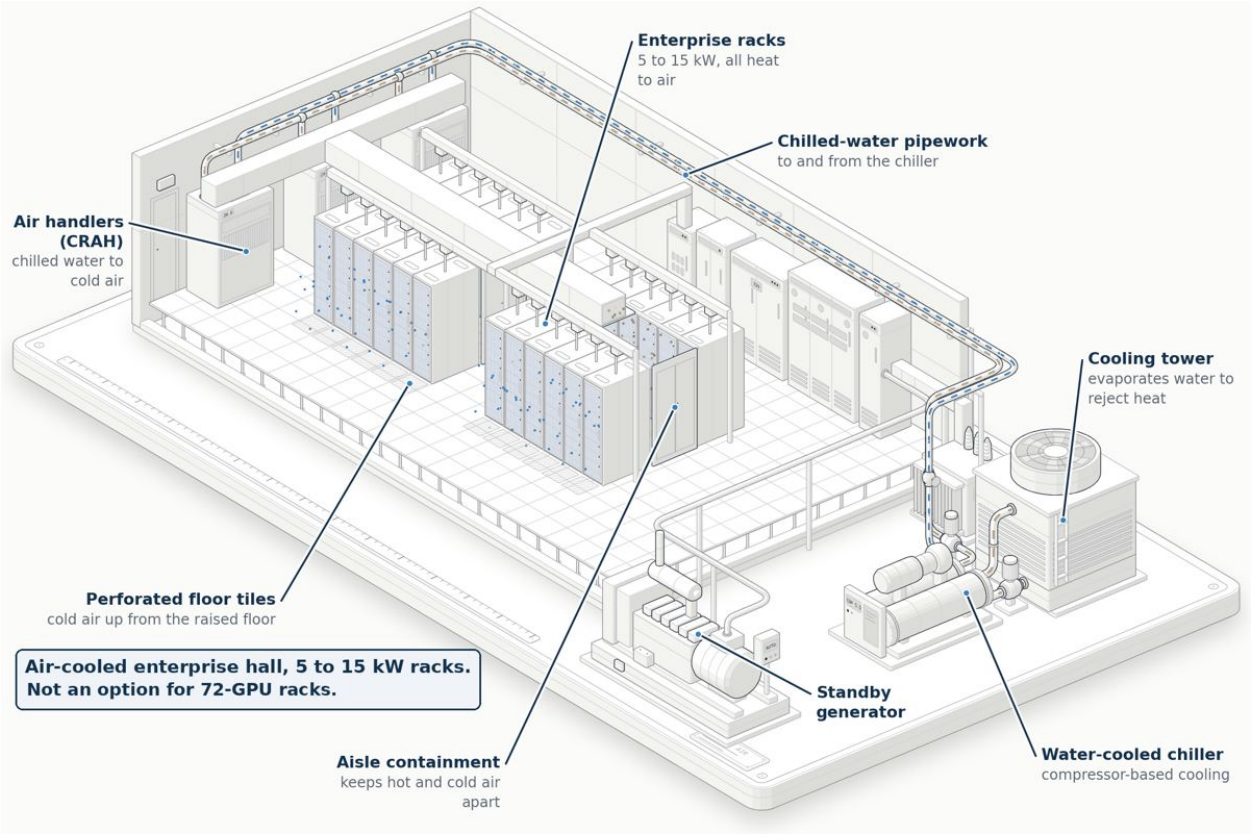

Figure 9. An air-cooled hall of enterprise racks at 5 to 15 kW each, in which every watt leaves the rack as hot air and reaches the outside through chillers and a cooling tower. This is how most existing halls are built, however it cannot cool a 72-GPU rack.

[14]AGORA worked example: 5,515 litres (1,457 US gallons) a minute of facility water per hall at a 12 K rise; dry coolers alone hold 32 °C water up to about 27 °C outside, and about 350 hours a year need trim chillers; 94.2 percent of the year's heat is rejected without compressors; 7.2 MW of reusable heat at 44 °C, about 63 GWh a year.

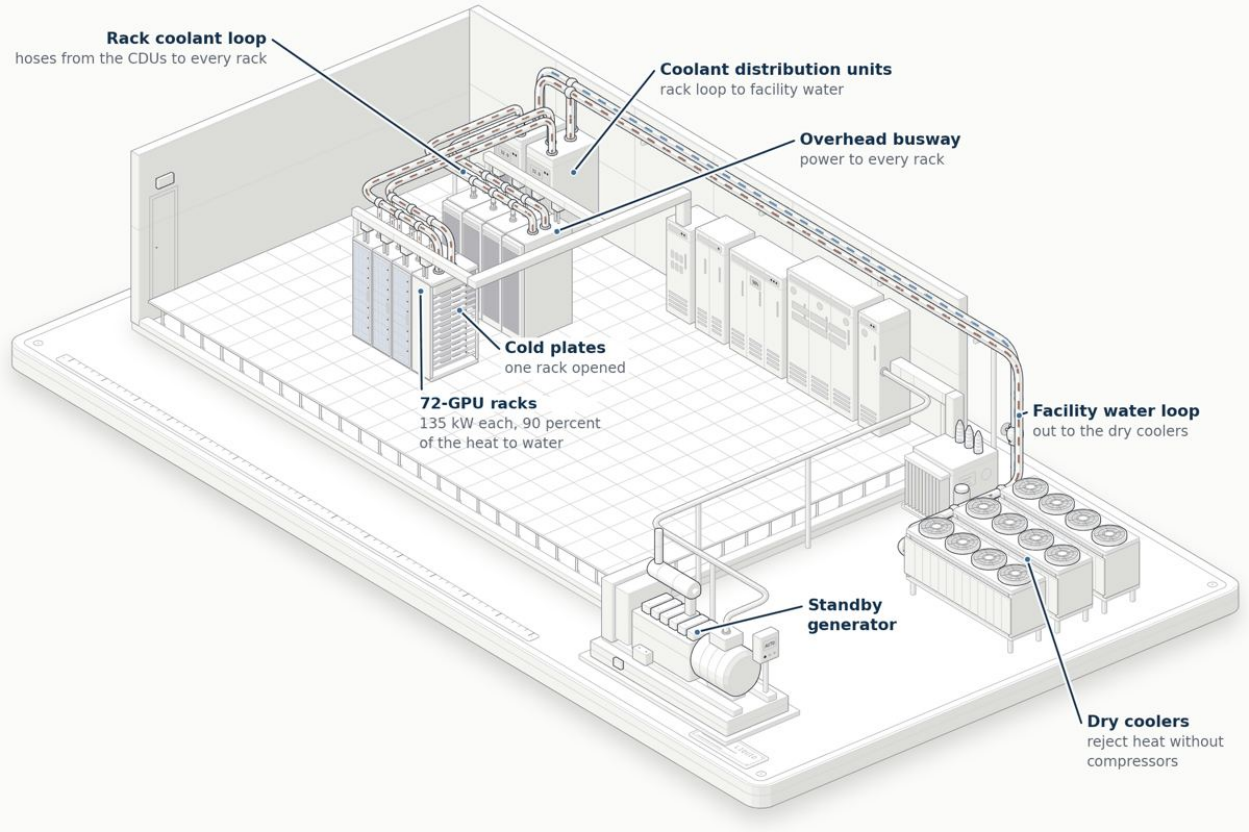

Figure 10. The same hall built for 72-GPU racks: 90 percent of the heat leaves through two water loops and dry coolers, with no compressor running for most of the year.

7. The power: from the grid to the busbar

Power reaches a GPU through a long chain of conversions, which Figure 11 follows for a single data hall of the example. On a US site of this size the utility typically delivers at 34.5 kV, which the site's own transformers step down to 13.8 kV for distribution, with the standby generators connected to the same medium-voltage switchgear so that they can pick up the whole site. Unit substations then take each hall's supply down to 480 volts, the UPS carries the load on batteries for the few minutes it takes the generators to start, and busways along each row deliver it to tap-off boxes above every rack, from which the rack's power shelves turn 480 volts AC into 50 volts DC on a busbar that runs the height of the rack[15]. At 135 kW that busbar carries about 2,700 amps, and the voltage regulators on each tray then take it down again, in two stages, to the fraction of a volt that the GPU's cores run on, at currents of well over a thousand amps per GPU.

[15] Typical US practice: utility supply at 34.5 kV for a site of this size, 69 or 138 kV and above for larger campuses; 13.8 kV medium-voltage distribution; 480 V three-phase to the UPS and busways. Power shelf datasheets for 2026 72-GPU racks: 346 to 480 V line-to-line AC input at 50/60 Hz, 50 V DC output, 33 kW from six 5.5 kW supplies, eight shelves per rack. 135 kW ÷ 50 V = 2,700 A.

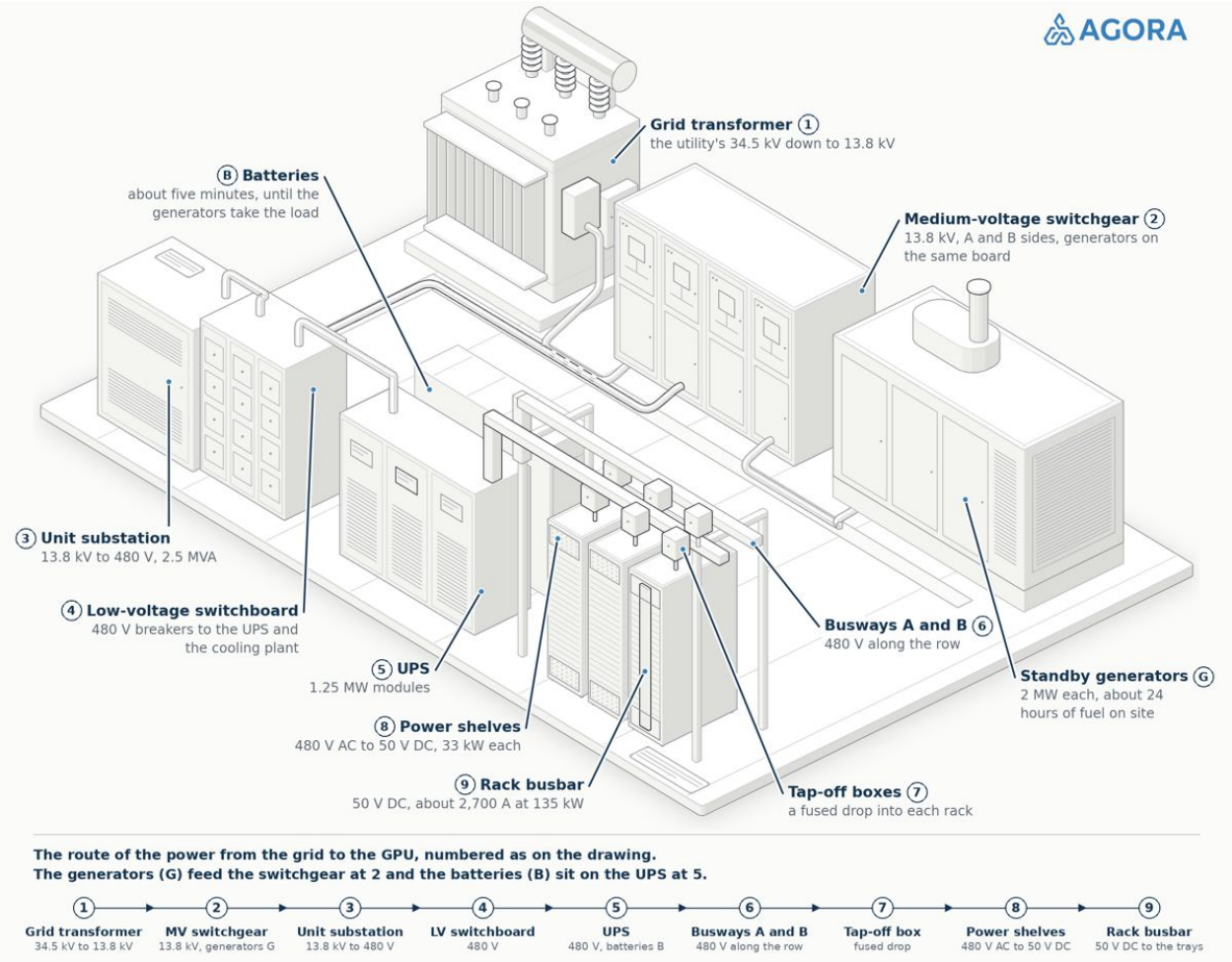

Figure 11. The power chain of one hall, compressed onto one pad and not to scale, with the route of the power numbered from the grid transformer (1) to the rack busbar (9) and set out in the strip below the drawing: medium voltage at the back, low voltage at the front, two busways over three racks and the DC busbar in the last of them.

The example is designed to Tier III, which means that every piece of plant has to be maintainable without stopping the load, so there is a spare UPS module, generator, chiller, dry cooler and CDU in every group, and two independent paths, A and B, run from the medium-voltage board to the rack, so that each rack sits under two busways and each tray draws from both. For the 10 MW example that comes to two grid transformers, eight 2 MW generators with 24 hours of fuel, ten UPS modules and 12 unit transformers in A and B pairs, for a facility that peaks at 12.3 MW, whilst a full row of 19 racks draws 2.6 MW, which takes two busways on each path. The network's 1.3 MW draws through the same chain, as its switches and optics sit on the UPS exactly as the racks do, so a detailed design sizes every element from the grid transformer to the busway for the 11.3 MW of the cabled cluster rather than the 10 MW on the specification.

This chain is also about to change, as the next generation of racks moves to 800 volts DC from 2027, with a single conversion from AC at the edge of the building, the batteries hung directly on the DC bus rather than behind a UPS of their own, DC busways along the rows and DC-to-DC conversion in the rack, which its developers say improves efficiency by up to 5 percent and cuts the copper by almost half, which matters because a 1 MW rack fed at 54 volts would need up to 200 kilograms of

copper busbar[16]. For a hall this means fewer UPS rooms and AC switchboards and more rectifiers and DC protection, so a building designed today for 480 volts AC should leave room for that change.

8. The row and the hall

Every row in a modern AI data centre has a steel frame above it carrying the fibre, the power and the water, and the order they are stacked in decides how tall the hall needs to be, how the rows get serviced and what a leak lands on. The usual arrangement, shown in Figure 12, puts the fibre tray at the front of the racks where the network ports are, with the 144-core trunks from the network racks landing in patch cassettes at the end of each row, the two busways directly over the racks where the power shelves draw from them, and the coolant mains over the contained hot aisle at the back, with an isolation valve on every drop so that one rack can be drained without stopping the row, and all of it sits on a frame that stands on the floor rather than hanging from the roof. Hot-aisle containment is still required in a liquid-cooled hall, as the 10 percent of heat that goes to air is as much as a whole enterprise rack once produced.

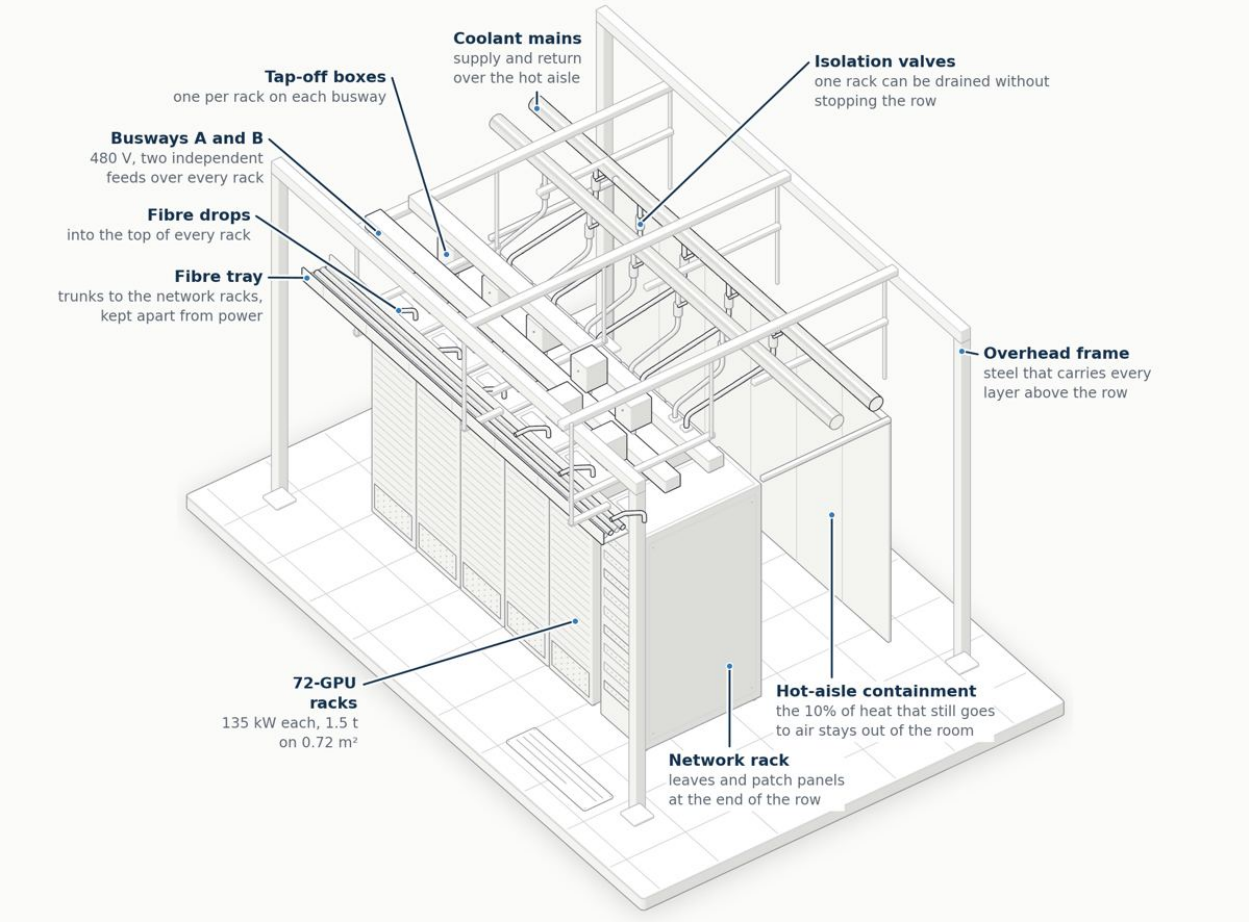

Figure 12. A short row of 72-GPU racks and the layers above it: fibre at the front, power over the racks, water over the hot aisle.

The same logic is used to design the hall itself, as each 5 MW hall of the example holds 38 racks in two rows of 19 in a room of 24.8 by 16 metres (81 by 52 feet), which is short enough for every GPU to reach a leaf switch on multimode fibre, with the CDUs along the north wall, the air handlers on the

[16] Industry 800 V DC architecture announcements, 2025: 800 V DC from 2027 with the next generation of rack-scale systems; up to 5 percent better end-to-end efficiency than 54 V systems; 85 percent more power through the same conductor and 45 percent less copper than 415 V AC; up to 200 kg of copper busbar for a 1 MW rack at 54 V. AGORA worked example for the plant counts: two 20 MVA grid transformers, 8 × 2 MW generators, 76,750 litres of fuel in two bulk tanks, 10 × 1.25 MW UPS modules, 12 × 2.5 MVA unit substations; a full row of 19 racks at 2,565 kW.

side walls, the 21 network racks along the south wall beside the door, and the UPS and switchrooms outside the white space altogether (Figure 13). The hall is small because the power density is so high, so most of the building is taken up by plant rather than racks.

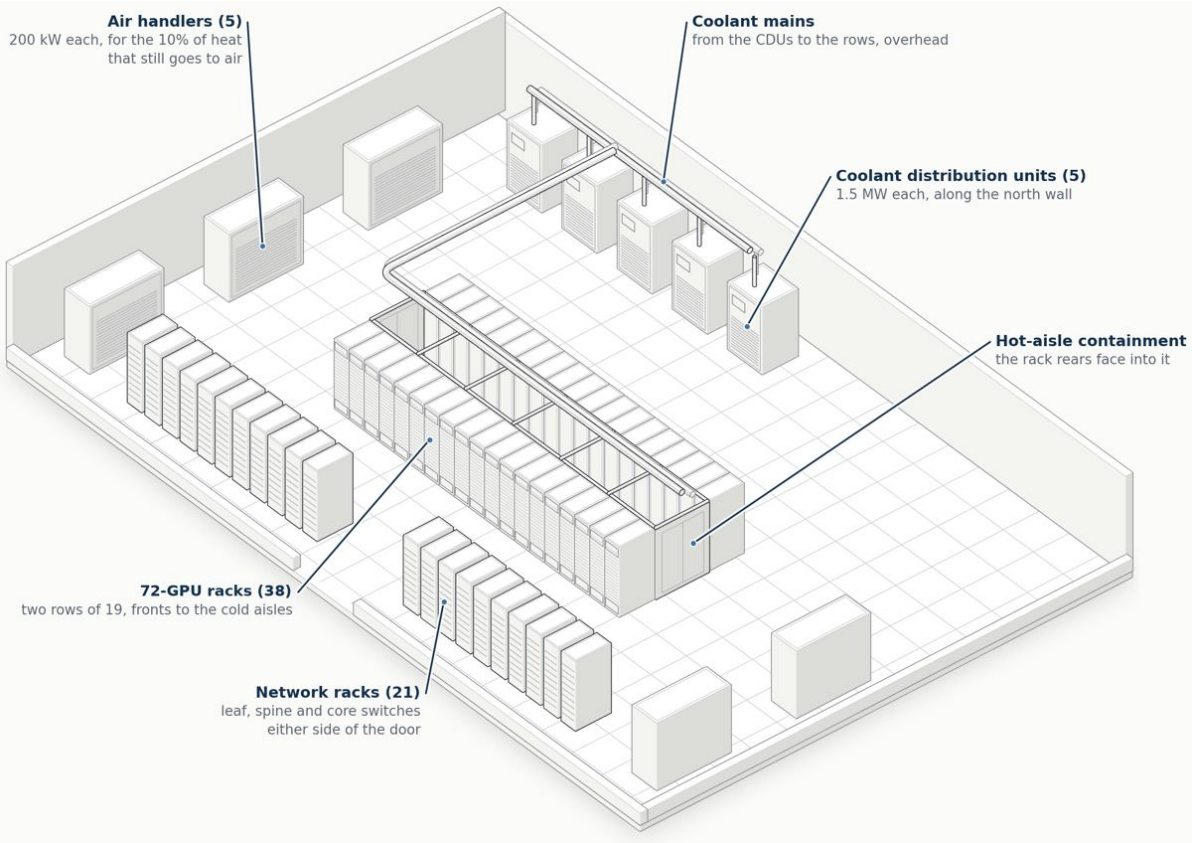

Figure 13. One 5 MW hall of the example, 24.8 by 16 metres (81 by 52 feet), with the near walls cut away: two rows of 19 racks either side of a contained hot aisle, five coolant distribution units along the north wall feeding the coolant mains over the aisle, five air handlers on the side walls, and 21 network racks along the south wall either side of the door.

9. Putting it together: the 10 MW data centre

The sections above have taken the 10 MW example apart one layer at a time, and put back together it comprises two data halls in a building 70 by 38 metres (228 by 125 feet) holding 76 racks and 5,472 GPUs, with the generator yard, the cooling yard and the medium-voltage intake around it on a site of under four acres. It would cost about $167 million to build before any IT hardware is bought, which is $16.7 million per megawatt using a mid-range US construction cost, and as with any estimate at this early stage the final cost could be around 30 percent higher or lower[17], whilst Table 4 sets out its main specification.

[17] The construction cost uses Turner & Townsend's Data Centre Construction Cost Index 2025, in which US markets range from US$9.5 per watt in Charlotte to US$13.3 per watt in Silicon Valley. The example takes US$10.5 per watt, adjusted for the size of the building and increased by 8.5 percent for liquid cooling, which gives US$128m for the building itself. Site works add 8 percent, then design fees and contingency add 10 percent each on the running total, bringing it to US$167m. The estimate excludes land, the utility connection and all IT equipment.



Figure 14. The 10 MW example: one building of two 5 MW halls, the generator yard to the south, the cooling yard to the north and the medium-voltage intake at the west end.

The 10 MW example	Specification	Note
IT load, halls, racks	10 MW, 2 halls, 76 racks	38 racks per hall in two rows of 19
GPUs, CPUs, compute trays	5,472, 2,736, 1,368	72 GPUs and 36 CPUs per rack
Rack power, weight, floor load	135 kW (155 peak), 1.5 t, 20.4 kPa	about 425 psf against a typical 250 psf slab
Heat to water, coolant	90%, 36 °C in and about 55 °C out	facility water 32 to 44 °C, dry coolers
Facility peak, design PUE	12.3 MW, 1.23	about 1.08 over a year, temperate US climate
Compute fabric	Ethernet, 2 planes of 400G, 3 tiers	694 switches, 32,832 links
All networks: transceivers, power	38,415; 1,306 kW, 12.7% of IT	not in any rack budget
Capital cost before IT	$167m, $16.7m per MW	US benchmark, about ±30%

Table 4. The 10 MW example, prepared by AGORA's engineering team for illustration, October 2026.

At this stage of a design our engineering team prepares a set of outline drawings covering the site, the building, the electrical and cooling systems, the network and the hall, and Figure 15 shows the ground floor plan from that set, whilst Figure 16 shows the same building in three dimensions.

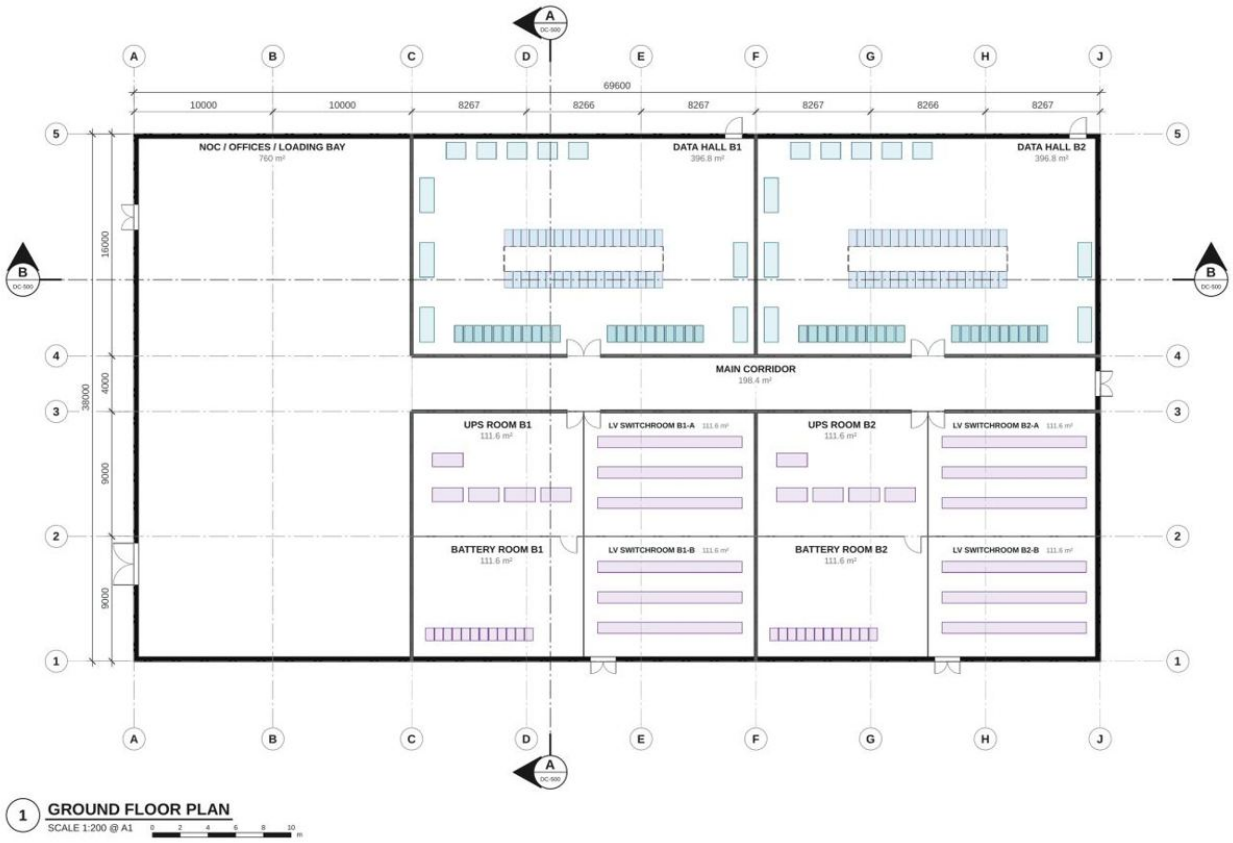

Figure 15. The ground floor plan from the general arrangement sheet of the concept drawing set for the 10 MW example, drawn at 1:200 on A1, with the structural grid, the dimensions in millimetres, the room areas and section lines A and B.

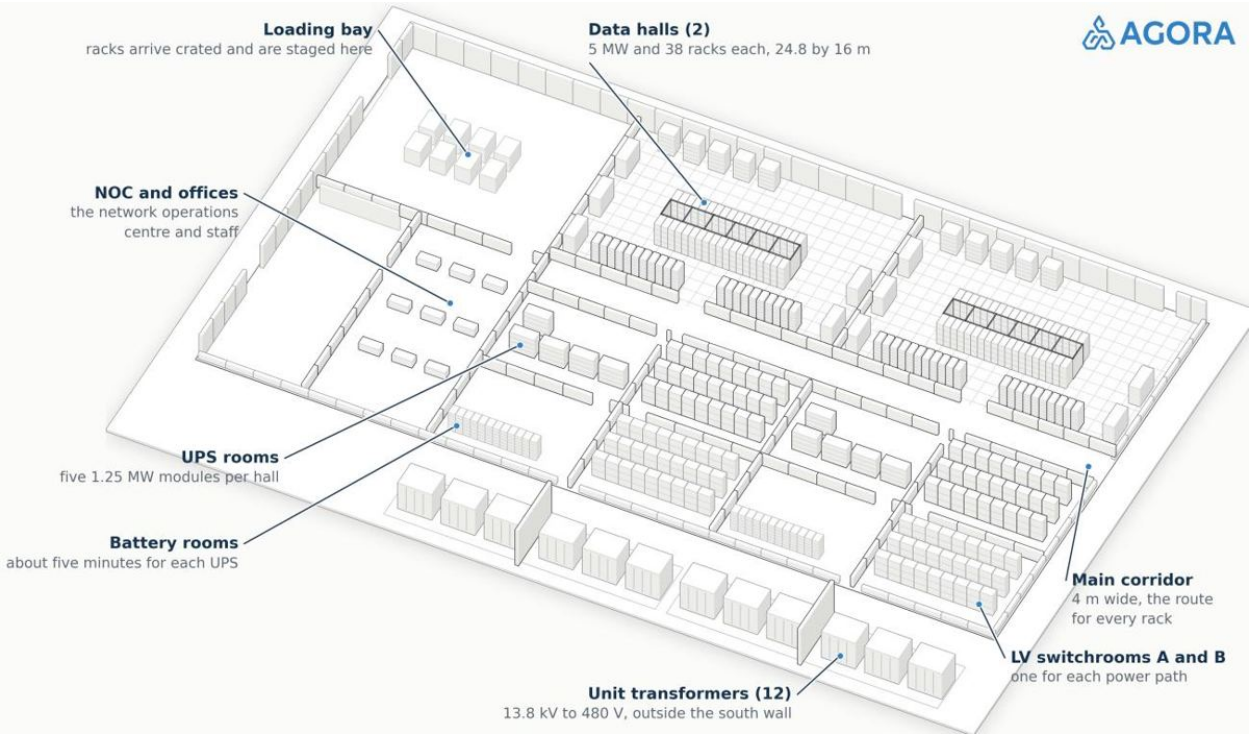

Figure 16. The building of the 10 MW example, 70 by 38 metres (228 by 125 feet), with its walls cut low: two halls along the north side, the main corridor, and for each hall a UPS room, a battery room and A and B switchrooms to the south, with the loading bay, NOC and offices at the west end and the unit transformers outside the south wall.

10. From one rack to a campus

The 10 MW example is a useful configuration for walking through each component, however clusters are built at many other sizes, so to show what changes as they grow we have built the same design at five sizes, from a single rack to a 100 MW campus, with the same racks, the same Ethernet fabric and the same Tier III rules throughout (Table 5)[18]. A single 72-GPU rack needs no scale-out fabric at all, as its GPUs are already joined by the scale-up network, so it can sit in a colocation cage with a feed on each side and an in-row CDU or sidecar, provided the floor can take its weight. One scalable unit of eight racks and 576 GPUs, which is the building block of the published reference designs, brings in the scale-out fabric on two tiers of switches, and at about 1.3 MW of facility load it still fits within a colocation hall or a small data hall.

The next step up has a greater effect on the building than on the network, as a full 5 MW hall of 38 racks still runs on a two-tier fabric, however it needs its own medium-voltage supply, generators and facility-water plant, and it brings the cost of building down from about $21.5 million per megawatt for a scalable unit built on its own to about $17.8 million. The network changes at around 4,600 GPUs, which is the most that two tiers of the example's Ethernet switches can join and which the example's 5,472 GPUs have just passed, and beyond that point the third tier raises the number of switches per thousand GPUs by about half and the network's share of the IT load from under 9 percent to nearly 13 percent, whilst on InfiniBand the same step comes at about 10,400 GPUs and on the co-packaged switches at about 18,400, which is why the choice of switch matters most to clusters close to those sizes. At 100 MW the design reaches a campus of three buildings fed from a 138 kV substation, with 66 generators, almost three million fibre cores and links between buildings in the two-kilometre optic class, whilst the cost of building falls to about $14.3 million per megawatt (Figure 17).

	One rack	One scalable unit	One hall	10 MW site	100 MW campus
GPUs (racks)	72 (1)	576 (8)	2,736 (38)	5,472 (76)	54,720 (760)
IT load	135 kW	1.1 MW	5.1 MW	10.3 MW	about 100 MW
Where it usually goes	colocation cage	colocation hall	its own hall	one building	three buildings
Scale-out tiers, switches	scale-up only	2 tiers, 44	2 tiers, 224	3 tiers, 694	3 tiers, 6,936
Network power, % of IT	frontend only	8.6%	8.7%	12.7%	12.9%
Longest scale-out link	in the rack	52 m, DR	70 m, DR	182 m, DR	517 m, FR
Utility supply	colocation	13.8 kV	34.5 kV	34.5 kV	138 kV
Generators, 2 MW	colocation	2	5	8	66
Facility peak	about 0.17 MW	1.3 MW	6.3 MW	12.3 MW	123 MW
Build cost per MW	colocation rent	$21.5m	$17.8m	$16.7m	$14.3m

Table 5. The same design at five sizes: 72-GPU racks on Ethernet in two 400G planes, Tier III, temperate US site. DR is the 500-metre single-mode class and FR the two-kilometre class.

[18] AGORA worked examples at five sizes on the same basis as footnote 1: 72-GPU racks with 128 × 400G Ethernet switches in two 400G planes, non-blocking, Tier III, the same climate and cost benchmarks. The single rack is not modelled as a building; the scalable unit is costed as a stand-alone building, which is why its cost per megawatt is high and why colocation usually takes its place; the campus is three buildings of six or seven 5 MW halls on a 138 kV supply, with 2 MW generators kept for comparison where a campus would normally use larger sets. From the same fat-tree sizing, two tiers reach 8 scalable units (4,608 GPUs) on 128 × 400G Ethernet switches in 400G planes, 18 (10,368 GPUs) on 144 × 800G InfiniBand switches and 32 (18,432 GPUs) on a 256-port co-packaged switch.

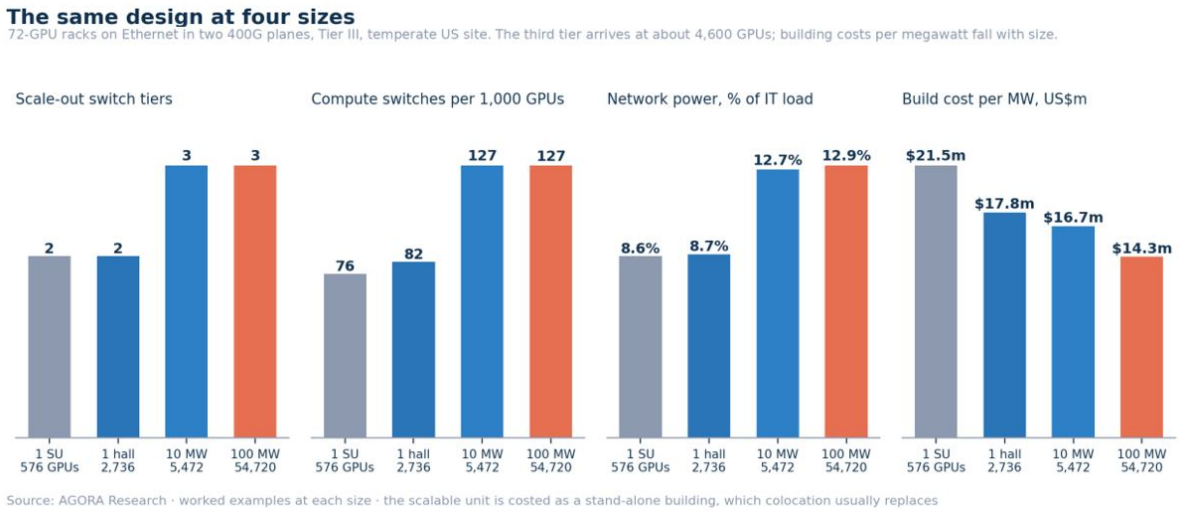

Figure 17. The same design at four sizes: switch tiers, compute switches per thousand GPUs, network power and the cost of building per megawatt.

11. What it means

11.1 For anyone buying a cluster

For anyone buying a cluster, the parts of the specification that matter most are the network, the cooling and the hall rather than the GPU, as choosing 5,472 GPUs only settles the easy half of the design, whilst the questions that decide the other half, which are the fabric, the oversubscription, the optics, the water temperature, the slab and the busway, are the ones to settle before the purchase order rather than after it, and are the questions our engineering team works through with clients at the design stage. Three numbers from this paper are worth keeping in mind, as the network is about an eighth of the power and a third of the racks on the floor, co-packaged optics could take almost half of that power and a tier of switching away, and the long lead times are in the plant rather than the IT, as a substation transformer is a three-year item on our lead-time series against a few months for the racks and, the GPU vendors' own modules aside, for the optics that connect them.

11.2 For anyone financing one

Our own node bill of materials puts the non-GPU content of a node at about 42 percent, and the non-GPU content of a cluster is larger again once the fabric, the optics, the distribution units and the building are counted[19]. That matters for residual value, which our lending decisions and our view of what the hardware will be worth at the end of a term both depend on, as the parts of a cluster hold their value differently, as a GPU has a used market, a switch has a smaller one, a transceiver has almost none, and a cold-plate manifold has a value only in the hall it was built for. A lender who takes security over a cluster is taking security over four or five different kinds of asset with four or five different curves, and the GPU curve is the most favourable of them, whilst the move to co-packaged optics and 800-volt DC means that some of today's network and power hardware will be a generation behind sooner than the GPUs it serves.

12. Limitations

The numbers in this paper come from a worked example at concept stage and from vendor material, and neither is a built cluster or a design for a real site. The example simplifies in several places, as it

[19]AGORA analysis, October 2026.

uses one cable length per switch tier, treats every scale-out link as optical, keeps the spine and core switches in the halls rather than in a central network room, and sizes the electrical and cooling plant on the 10 MW of compute with the network's 1.3 MW identified separately, whereas a detailed design would carry it through every system. The smaller sizes in section 10 are costed as stand-alone buildings where most would go into colocation, the co-packaged variant rests on products that were only entering production in 2026 and on our own estimate of their power, and the climate, voltages and costs are typical US values rather than those of any site, utility or market. The costs are benchmarks good to perhaps 30 percent and are intended for comparing options rather than budgeting, and nothing here includes a rental price or says what any of this hardware is worth over time.

13. Conclusion

Most of a working cluster is made up of parts other than the GPU, as behind every compute tray of the example run four networks, one in copper that never leaves the rack and three in fibre that fill 42 racks with 739 switches and over 38,000 transceivers, and draw 1.3 MW that is not in the 10 MW specification. The choice of fabric changes the switch count nearly fivefold before InfiniBand or Ethernet is even considered, the choice of optics determines where the switches stand and, with co-packaged optics, whether a third tier is required at all, the choice of cooling determines how many racks a hall holds and whether its heat is removed by compressors, and the power reaches each rack through a chain of conversions that the next generation of hardware is about to shorten. Underneath all of this is a rack that weighs 1.5 tonnes, loads the floor at about 1.7 times what a typical hall was built for and sends 90 percent of its heat into water, therefore the data centre needs to be designed with the rack in mind from the start, and it is these parts, rather than the GPU, that determine whether a cluster performs as intended, what it costs to run and whether it was worth buying.

About us

AGORA designs, deploys and finances GPU clusters across the US. Our engineering team works on the networks, cooling and power behind these clusters and commissions them on site, which is where the material in this paper comes from, and as we also finance the hardware we take a view on what each part of a cluster is worth over time. Our sister company, Arden Infrastructure Group, secures the power and builds the sites that house them. The worked example in this paper was put together by our engineering team for illustration and is not a design for any site or client, and the paper is published at agoraexchange.ai.