ChatGPT

Designing Mega-Scale AI GPU Clusters (100K+ GPUs) – Architecture, Networking, and Best Practices

Introduction

Building an "AI factory" with tens of thousands of GPUs is an immense engineering feat. Recent projects in Europe underscore this scale: for example, Fluidstack and Eclairion are deploying a **40 MW, 18,000-GPU** supercomputer in France for Mistral AI, with plans to expand beyond **100 MW** (eventually over 100,000 GPUs). In fact, Fluidstack signed a €10 billion MoU to build a **1 GW AI supercomputer in France** housing ~500,000 next-gen AI chips by 2026. These sovereign European AI clusters leverage France's abundant nuclear power (100% carbon-free) – a strategic advantage highlighted by President Macron: *"our nuclear energy is controllable, safe, stable, and decarbonized – ideal for expanding our AI computing capabilities"*.

Designing an on-premises GPU cluster at this scale requires **cutting-edge hardware and thoughtful architecture** to ensure performance, efficiency, and resilience. This report dives into the latest **GPU server platforms** (e.g. NVIDIA **Grace Blackwell** GB200/GB300-based systems and AMD MI300), large-scale reference designs from NVIDIA, Dell, Supermicro, etc., and the **networking, storage, and cluster topology** considerations for 100K+ GPU deployments. We focus on *on-premises* "AI cloud" builds (e.g. in Europe under SecNumCloud-like security) rather than public cloud, and incorporate regional factors like power costs in France, Spain, and Germany. All information is backed by technical references (with PDF links) – no guesswork or hallucination.

Latest GPU Platforms for Extreme-Scale AI

NVIDIA Blackwell Architecture (B100/B200 GPUs) – NVIDIA's newest GPU generation (successor to "Hopper" H100) is built to handle trillion-parameter models. Blackwell GPUs (e.g. **B200** Tensor Core GPU) come paired with NVIDIA's **Grace** CPU in the Grace-Blackwell superchip. Each **GB200 Grace-Blackwell Superchip** integrates a 72-core Grace CPU with **two B200 GPUs** via a 5th-gen NVLink C2C interface delivering **900 GB/s** bandwidth [1]. This tight coupling provides enormous memory bandwidth and capacity in one package. Blackwell GPUs introduce faster Tensor Cores and **5th-Gen NVLink** connectivity: NVLink 5 reaches **1.8 TB/s bidirectional throughput per GPU** for peer-to-peer communication – double the prior generation. For multi-GPU servers, NVIDIA's HGX B200 platform includes 8× B200 GPUs connected by NVLink Switches, supporting up to **400 Gb/s** external networking (compatible with Quantum-2 InfiniBand and Spectrum-X Ethernet fabrics).

NVIDIA "Ultra" GPUs (B300) – Some Blackwell variants are geared for ultra-high-end performance. For instance, an NVIDIA HGX **B300** platform supports eight GPUs at **1100 W TDP each** (liquid-cooled), with up to 2.3 TB total HBM3e memory. These massive modules are intended for the most demanding workloads. An air-cooled 8U chassis can accommodate 8× 1100W GPUs plus oversized heatsinks. In practice, a B300 GPU likely offers more HBM3e memory per GPU (e.g. 288 GB vs ~180 GB on B200) – enabling a single 8-GPU

server to have **>2 TB of HBM** for training large models. Such extreme GPUs highlight the power/cooling demands: ~9 kW for GPUs alone in one server.

Grace Hopper vs. Grace Blackwell – NVIDIA's prior combo, the **GH200 (Grace-Hopper)**, paired Grace CPU with Hopper H100 GPU. It enabled multi-node NVLink clusters (e.g. the DGX GH200 design linked 256 H100 GPUs via NVLink Switch). The new **GB200 (Grace-Blackwell)** superchip similarly focuses on tightly coupling CPU and GPU for acceleration. NVIDIA has announced a multi-node GB200 system called **GB200 NVL72**: a **72-GPU, 36-Grace** cluster acting as one giant GPU memory pool [2]. This is essentially an "AI supercomputer in a rack" with GPUs and CPUs all linked by NVLink (more in Section 3). It's optimized for inference; NVIDIA reports that **GB200 NVL72 delivers up to 30× faster trillion-parameter LLM inference** than an H100-based setup, while reducing energy consumption **25×** [3]. In other words, one rack of GB200 NVL72 can achieve what dozens of H100 racks could do, thanks to tight coupling and massive NVLink bandwidth.

AMD Instinct MI300 – AMD's flagship GPU for AI, the MI300, is the closest competitor to Nvidia's offerings. The **MI300A** variant combines an EPYC CPU die with 96 CDNA3 GPU CUs (MI300A is being used in the El Capitan exascale supercomputer), whereas the **MI300X** is a GPU-only 192 GB HBM3e monster aimed at AI training. MI300X offers **5.3 TB/s memory bandwidth**, slightly higher than NVIDIA H100's ~4.8 TB/s. Early benchmarks indicate MI300X performs *absolutely on par* with or better than H100 in large-model inference, especially with large batch sizes. An 8×MI300X server (with 8× 192 GB = 1.5 TB HBM) can outperform an 8×H100 (8× 80 GB) in memory-heavy workloads. AMD's MI300 supports PCIe 5.0 and will support Infinity Fabric links between GPUs in the future (though not as extensive as NVLink Switch). For a sovereign cloud aiming to diversify, MI300-based systems (often coupled with AMD EPYC CPUs) provide an alternative to NVIDIA – indeed, Dell is also offering **PowerEdge servers with 8× Intel Gaudi 3** accelerators as part of its portfolio [4] (though Gaudi is less mature ecosystem-wise). While MI300X GPUs boast strong specs and **20–60% speedups over H100 in some tasks** (per AMD claims), NVIDIA's software stack (CUDA, libraries) remains a decisive factor. Thus, most mega-clusters today (Meta, OpenAI, etc.) still standardize on NVIDIA, but high-end AMD GPUs may appear in certain HPC/sovereign installations for diversification or if they prove cost-effective.

Power & Efficiency – Top-tier GPU nodes now regularly draw **>10 kW per server** (GPUs + CPUs + memory). Each NVIDIA B200 GPU is ~700W (and B300 ~1.1 kW), so 8 GPUs = 5.6–8.8 kW, plus ~500W for dual CPUs, plus NICs/DPUs (~300W, see below). Therefore, power delivery and cooling are first-order design constraints. Only direct liquid cooling can handle these densities efficiently – more on this in Section 6. Notably, NVIDIA and its partners emphasize efficiency gains of the new architectures: e.g., **a liquid-cooled GB200 NVL72 rack is ~25× more energy-efficient than an air-cooled H100 cluster** with equivalent performance [5]. Supermicro similarly claims the Blackwell platform delivers "**25× more performance at the same power**" compared to the previous generation [6]. In part, this is due to better performance per GPU and cooling effectiveness – but also because the NVLink-connected design avoids wasteful data movement over slower interconnects, keeping GPUs fed with data more efficiently.

High-Density GPU Servers and Reference Designs

To realize mega-scale clusters, vendors are delivering **integrated rack solutions** that maximize GPU density and simplify deployment. Rather than piecemeal servers, these come as pre-engineered racks with

power, cooling, and network configured for AI. We examine Dell, NVIDIA, and Supermicro's latest designs for Blackwell-generation clusters:

- **Dell Integrated Rack 5000 (IR5000)** – a traditional 19″-wide rack offering extreme density within EIA form factor. Dell packages their high-end PowerEdge servers (like XE9680L, XE9685L, XE7740 etc.) into IR5000 racks with either air or liquid cooling. An **IR5000 with direct liquid cooling supports up to 96 GPUs per rack** (e.g. 12 servers × 8 GPUs) while even air-cooled configs can do 72 GPUs/rack. This is *"the highest GPU density available for standard (19-inch) racks"*. For example, Dell's new PowerEdge **XE9685L** is a 4U liquid-cooled server with 2× 5th-Gen AMD EPYC CPUs and **8× NVIDIA H200 or B200 GPUs**, and Dell quotes **96 GPUs per rack** using this 4U node (i.e. 12 nodes × 8 GPUs). Such a rack would be ~48U of compute plus space for network gear. If using air-cooled 4U servers (like the XE7740 which can host 8 double-width accelerators with heavy-duty fans), the limit is ~72 GPUs/rack (because air cooling typically can't support full rack of 8kW servers).

- **Dell Integrated Rack 7000 (IR7000)** – a next-generation **21″ OCP-standard rack** focused on maximum density and liquid cooling. Announced in 2024, the IR7000 is a 50OU (Open U) rack built **natively for liquid cooling**, capable of handling **up to 480 kW per rack** and capturing *"nearly 100% of heat"* via liquid [7]. The IR7000 uses **wider, taller sleds** (1OU height, 21″ width) to accommodate larger GPUs and cooling. Dell's new PowerEdge **XE9712** is a 1OU sled designed for IR7000: it's an **NVIDIA GB200 NVL72-based** system that effectively fills an IR7000 rack with **36 Grace CPUs + 72 Blackwell GPUs** all interconnected [3]. In other words, **one XE9712 rack = 72 GPUs** in a single NVLink domain. Dell confirms this 72-GPU "super-node" *"acts as a single GPU"* for LLM inference, yielding **up to 30× faster throughput** on trillion-parameter models, and **25× efficiency gain**, compared to conventional H100 setups [8]. The IR7000 is engineered for such high density: it uses **wider sleds** to fit more electronics and cooling, and has scalable power shelves to supply nearly half a megawatt to one rack [7] [3]. Dell quotes support for *"up to 144 GPUs per rack in a 50OU frame"* for IR7000 in some announcements – likely referring to a configuration with multiple 1OU sleds each holding fewer GPUs (e.g. perhaps 2–4 GPUs per sled but many sleds). The **Dell PowerEdge M7725** is another IR7000 sled example: a 1OU AMD EPYC server for dense CPU compute, allowing 64–72 dual-socket nodes per rack (24K–27K cores per rack) with a mix of direct liquid cooled CPUs and air-cooled components [9]. In short, IR7000 is *purpose-built for next-gen AI hardware*, providing a **"future-proof" OCP rack that can handle multigeneration heterogeneous nodes** under liquid cooling [10].

- **NVIDIA DGX SuperPOD with DGX B300 systems** – NVIDIA's own reference architecture for AI clusters. The latest design (2025) uses DGX **B300** nodes (presumably 8× Blackwell GPUs per node, possibly with Grace CPUs in each node) as the compute building block [11]. An NVIDIA SuperPOD is a **pod of DGX nodes plus all networking and storage** needed for turn-key AI. The new SuperPOD RA (reference architecture) provides two options for the high-speed compute fabric: either **InfiniBand XDR** (800 Gb/s) or **Spectrum-X Ethernet** (800 Gb/s). In either case the topology is a *"non-blocking twin-plane fat-tree"* for extreme scale. Each DGX B300 node has dual network interfaces (either two ConnectX-8 NICs or dual InfiniBand ports) connecting into two separate switch planes for fault tolerance. A single SuperPOD "Scalable Unit" might be 32 nodes (256 GPUs) or larger, and multiple such units can scale out. The SuperPOD RA emphasizes **modular deployment**: e.g. one could deploy a rack-level SuperPOD that includes a certain number of DGX nodes plus a pair of spine switches, then add more racks (each with leaf switches and nodes) scaling out like a Clos network [11]. Notably, NVIDIA also introduced a **DC** busbar power *option in the DGX B300 SuperPOD, meaning the racks are*

fed with high-voltage DC and use DC-to-DC converters, improving power distribution efficiency at scale [11]*. In terms of density, NVIDIA's DGX H100 (previous gen) achieved 8 GPUs per 6U and about 32 DGX per rack (~256 GPUs per rack). With B300 and liquid cooling, similar or higher per-rack counts are expected. But NVIDIA's own focus is less on cramming maximum GPUs in one rack, and more on balanced architecture* across many racks. We will cover their networking and storage guidelines in later sections.

- **Supermicro NVIDIA GB200 NVL72 SuperCluster** – Supermicro, known for OEMing many AI systems, has unveiled an **"Exascale Supercomputer in a Rack"** solution in Oct 2024 built around NVIDIA's 72-GPU Grace-Blackwell design [6]. This is essentially Supermicro's implementation of the **GB200 NVL72** concept in a turnkey rack. The SuperCluster rack contains **18× 1U compute trays**, each with **2 Grace CPUs + 4 Blackwell GPUs**, totaling 72 GPUs and 36 CPUs [12] [13]. These are interconnected by **9× NVLink Switches** (each switch board connecting 8 GPUs across trays) forming NVIDIA's *"largest NVLink network to date,"* with an all-to-all bandwidth of **130 TB/s** (!) across the 72 GPUs [14]. The NVLink Switch system provides **1.8 TB/s GPU-to-GPU interconnect** for every GPU pair in the 72-GPU domain [15]. In effect, the entire rack's GPUs operate as one giant pool with **13.5 TB of HBM3e memory** accessible (72 × ~187 GB each) [2]. Supermicro's design includes **8× 33 kW power shelves** (N+N redundant, 132 kW usable) and an **in-rack liquid cooling distribution unit (CDU)** (250 kW capacity) to manage the heat [16]. They also offer a *liquid-to-air heat exchanger* option (180–240 kW) if facility water is unavailable [17]. Impressively, Supermicro touts an **end-to-end solution** with manufacturing scale (5000+ liquid-cooled racks per month capacity) and global deployment services [18] [19] – highlighting how critical fast deployment is due to the surging AI demand. The SuperCluster supports **NVIDIA BlueField-3 DPUs ("SuperNICs") and Spectrum-X/Quantum-2** networking, meaning it's ready for either 800G Ethernet or InfiniBand out of the box [20]. This Supermicro system, like Dell's, claims **25× performance per watt vs. previous gen** and up to **40% reduction in datacenter electricity costs via liquid cooling** [6] [21].

Comparison: Both Dell and Supermicro are delivering *rack-level integrated "clusters"* for Blackwell: Dell's IR7000+XE9712 and Supermicro's NVL72 rack have similar core specs (72 GPUs + NVLink + liquid cooling). Dell's design is part of its broader **AI Factory** portfolio, which also includes enterprise storage (e.g. PowerScale NAS with NVIDIA certification) [22] and services, whereas Supermicro's emphasis is on raw performance and quick scaling of infrastructure. Another difference: Dell's IR7000 is OCP-standard, potentially accommodating third-party gear or future larger form-factors, while Supermicro uses its own rack design (likely also OCP-like). In terms of **density**, the NVL72 designs pack *fewer total GPUs per rack (72)* than a more conventional approach of many small servers per rack (e.g. one could fit 96–144 GPUs in a rack of 2U/4U servers). However, the **72 NVLinked GPUs act as one unit**, enabling training or inference jobs to utilize all 72 with *much faster interconnect* than if those 72 were split across multiple nodes with only InfiniBand/Ethernet links. This dramatically accelerates large model training/inference on that single "super-node" and reduces network bottlenecks. The trade-off is that a NVL72 rack is a *big failure domain* (if one NVLink switch fails, some portion of that giant GPU might be affected). Thus, in practice, mega-clusters might use *multiple* 72-GPU super-nodes as building blocks, connected via an Ethernet or IB fabric. For example, a 100,000-GPU deployment could be architected as ~1400 such 72-GPU supernodes, each internally NVLinked, and then networked to each other via a fat-tree fabric. That offers a hierarchical approach: NVLink for *intra-node* (fast, low-latency) and InfiniBand/Ethernet for *inter-node* (slower, but scalable to thousands of nodes). We discuss the networking next.

Networking Fabrics: InfiniBand vs. Spectrum-X Ethernet (RoCE v2)

High-performance interconnect is the lifeblood of large AI clusters – GPUs must communicate efficiently for distributed training (e.g. all-reduce operations in data-parallel training) and to access remote data. Two primary network technologies lead in this space: **NVIDIA InfiniBand** and **RDMA-enabled Ethernet**, specifically NVIDIA's **Spectrum-X** architecture for Ethernet at scale.

InfiniBand (IB) – The long-standing choice for HPC, InfiniBand is a purpose-built RDMA network with hardware-managed congestion control and very low latency. The latest generation **NDR InfiniBand (NVIDIA Quantum-2)** provides **400 Gb/s per port**, and the upcoming **XDR InfiniBand (Quantum-3)** doubles that to **800 Gb/s**. InfiniBand bandwidth is typically quoted bidirectionally, so 800G implies 400G each direction on a port. IB switches and NICs (e.g. ConnectX-7/8) also implement in-network computing features like NVIDIA's SHARP™ for collective reduction operations. For example, **SHARP v4** in Quantum-X800 can aggregate all-reduce operations in-switch with support for FP8 precision, providing a **9× increase (14.4 TFLOPS) in in-network compute capability** vs the previous gen. In practice, this speeds up large AI training by offloading some all-reduce work to the network. InfiniBand's advantages include extremely low latency (a few microseconds MPI latency) and very tight time-synchronized messaging – crucial for scaling to thousands of GPUs. It also has proven adaptive routing and congestion management suited for all-to-all communication patterns of AI.

For a mega-cluster, **InfiniBand is often deployed in a "fat-tree" (Clos) topology**, usually **two-layer (leaf-spine)** up to a certain size, and three-layer for the very largest. To avoid any oversubscription (full non-blocking), each leaf must have as much uplink bandwidth as downlink. This can get expensive at >10k nodes, so sometimes a slight oversubscription or a **Dragonfly+** topology is used (Cray's Slingshot network is an adaptive routing Dragonfly, for example). NVIDIA's reference suggests a **"twin-plane fat-tree topology"** for extreme-scale AI. **Twin-plane** means each node has dual network connections into two parallel fabrics (plane A and B). This can serve either for redundancy or aggregate bandwidth (often both). A *non-blocking twin-plane fat-tree* effectively doubles the bisection bandwidth and allows tolerating an entire plane's failure. HPC systems sometimes call this "dual-rail InfiniBand." Indeed, some leadership clusters use dual-rail NDR (each node 2× 200 Gb/s for 400 total) to get more throughput and failover capabilities. The DGX SuperPOD design explicitly lists a *"rail-optimized, non-blocking, twin-plane fat tree topology"* to support next-gen AI factories. This implies each DGX node has two NDR links going into two separate Clos networks that interconnect at the spine. In event of one rail's outage, the other still provides network connectivity (albeit at half bandwidth). Software like NCCL can be made topology-aware to use both rails or handle failures.

Ethernet (RoCE) – Ethernet is ubiquitous and cost-effective, but plain TCP/IP Ethernet is too slow (high latency, CPU overhead) for AI training. The solution is **RDMA over Converged Ethernet v2 (RoCEv2)**, which encapsulates IB-like RDMA packets in UDP/IP. RoCEv2 has become the standard for high-speed Ethernet in data centers (routable, works over L3 networks). When configured on a *lossless Ethernet fabric* with PFC (priority flow control) or ECN (explicit congestion notification), RoCEv2 can approach InfiniBand performance. However, RDMA Ethernet historically faced challenges at large scale – congestion and packet loss (due to imperfect PFC) can cripple throughput, and debugging is more complex than IB's end-to-end credit-based flow control. **NVIDIA Spectrum-X** is essentially a holistic solution to make Ethernet as robust and high-perf as InfiniBand for AI clusters. It combines the latest **Spectrum-4 switches** (e.g. SN5600) with

NVIDIA BlueField-3 DPUs (ConnectX-8 NICs) and a custom software stack for telemetry and control. Some highlights of Spectrum-X and modern RoCE networks:

- **800 Gb/s end-to-end** – The **Spectrum-X800** platform (announced Mar 2024) uses the **SN5600 switch (Spectrum-4 ASIC)** which has 64 ports of 800 GbE each (51.2 Tb/s total). Paired with **BlueField-3 "SuperNIC" DPUs** in servers, it achieves full 800 G throughput per link. In practice, current BlueField-3 cards provide 2× 200 Gb/s = 400 Gb/s; achieving 800 Gb/s per node is done by using two BF3 DPUs per node (one per plane, effectively). Spectrum-4 switches can be deployed in a Clos network similar to IB (leaf-spine). The **latency and bandwidth of Spectrum-X are comparable to IB XDR** according to NVIDIA, but with potentially lower cost (Ethernet optics and switches benefit from mass-market volumes).

- **Adaptive Routing & Congestion Control** – Like IB, Spectrum switches support dynamic routing of flows to avoid congested paths. They also implement advanced congestion control algorithms beyond simple PFC. For instance, NVIDIA's RoCE stack can use ECN marking and the **"HPCC"** algorithm or others to modulate send rates. A recent feature is **Swift (Switch-accelerated flow control)** which offloads some congestion management to the switch hardware. Overall, these reduce tail latency and prevent incast problems that plagued early RoCE deployments. Blocks & Files notes Spectrum's adaptive routing can significantly boost storage fabric performance under load – relevant because training tends to create bursty traffic patterns.

- **Multi-Tenant Isolation** – A key advantage touted for Spectrum-X in "AI cloud" scenarios is the ability to enforce tenant isolation at the network level. The **BlueField-3 DPUs** act as intelligent NICs that can perform traffic shaping, firewalling, and virtualization. According to NVIDIA, Spectrum-X is *"designed specifically for multi-tenant, hyperscale AI clouds"*, ensuring each tenant's workloads have **performance isolation** and quality-of-service controls. This prevents one user's noisy traffic from starving others – something native IB can struggle with unless carefully partitioned. In practical terms, the DPU can implement per-VM or per-container rate limiting and handle RDMA connection setup/teardown in hardware, offloading the host CPU.

- **Twin-Plane Ethernet Fabric** – Similar to IB dual-rail, NVIDIA's Spectrum-X reference uses a **two-plane topology**: *"The Spectrum-X based compute fabric features a twin-planar design (denoted in blue and green). Each GPU has 2× 400 GbE connectivity through two different planes"*. This means each server has two 400G ports plugged into two separate sets of switches (Plane A and B). This dual-plane provides both redundancy and an effective aggregate of 800G per node. In normal operation, traffic can be load-balanced across both planes (doubling bandwidth). If a switch/link fails on one plane, the other plane still carries traffic. This is analogous to dual-rail IB – a best practice for large clusters to avoid single points of failure.

- **Management & Telemetry** – NVIDIA provides tools like **UFM (Unified Fabric Manager)** for InfiniBand and **NetQ** for Ethernet to monitor the fabric health. NetQ can give real-time telemetry of latency, congestion events, etc., allowing operators to quickly pinpoint issues (e.g. a misbehaving node generating pause floods). This is critical at 1000+ node scale. Spectrum switches also support out-of-band network monitoring via telemetry agents. In the DGX SuperPOD RA, NVIDIA suggests deploying a **dedicated out-of-band management network** (e.g. separate 1 GbE switches for BMCs/OOB, rolled up to spine) [23] – of course, that applies regardless of IB or Eth for the data plane.

So, *Is RoCE v2 the "best and latest" for mega-clusters?* – In 2025, **RoCEv2 over Spectrum-X Ethernet is indeed a top-tier choice** for building large AI clusters, on par with InfiniBand. With Spectrum-4 switches and BlueField DPUs, Ethernet can achieve **800 Gb/s, low latency RDMA** similar to IB. NVIDIA explicitly positions Spectrum-X as an alternative to InfiniBand for AI clouds, boasting comparable performance at potentially lower cost. Early adopters like Microsoft Azure and Oracle Cloud have incorporated Spectrum-X in their AI infrastructure, indicating confidence in RoCE at cloud scale. That said, InfiniBand **Quantum-2/3** remains a gold standard for the absolutely highest performance (e.g. for tightly-coupled HPC simulations and large AI in supercomputer contexts). Many current mega-clusters (Meta's RSC, Microsoft/OpenAI's builds) actually use InfiniBand (e.g. RSC uses 200 Gb IB). But we are seeing a trend where next-gen "AI factories" might favor Ethernet for flexibility and cost: for example, **NVIDIA reports Spectrum-X can reduce AI job runtimes and do so with *"more affordable switches and optics"* compared to IB**. The choice often comes down to ecosystem and scale: if you need > tens of thousands of nodes, Ethernet's routability and broad vendor support can help, whereas IB is single-vendor and typically maxes out around a few thousand nodes per fabric (though that's increasing with Quantum-2).

In summary, **RoCE v2 is the state-of-the-art for Ethernet-based AI networking**, and under Spectrum-X or similar frameworks it is recommended for building multi-tenant clusters of 10k–100k GPUs. The best practice is to run it on a lossless, well-managed fabric (using features above) – essentially treating your Ethernet like an InfiniBand network in terms of configuration. Many large sites also segregate traffic types: for instance, **dedicated storage network vs. compute network**, even if both are Ethernet/RoCE. NVIDIA's SuperPOD RA provides two options for storage fabric: an InfiniBand storage network or an Ethernet (RoCE) storage network. Either way, **RDMA is essential** for storage IO to avoid CPU overhead, as noted: *"Storage is provided over InfiniBand or RoCE to provide maximum performance and minimize CPU usage"*.

Network Topologies: Fat-Tree, Dragonfly, Bcube – The vast majority of large GPU clusters use a **Clos fat-tree** topology because it's straightforward and guarantees uniform bandwidth. However, alternative topologies exist and may become relevant as clusters approach exascale sizes:

- **Dragonfly** – A high-radix low-diameter topology used in some supercomputers (e.g. Cray Aries, HPE Slingshot interconnect). Dragonfly groups nodes into local groups fully connected, and connects groups via select global links. It reduces the number of switch hops (often only 2 or 3 hops across the whole network vs 5+ in a large fat-tree). This can reduce latency and cost (fewer total switches), but requires very high-radix switches and sophisticated adaptive routing to avoid global link congestion. Modern Ethernet/IB switches (radix 64–128 at 200–400G) are high-radix enough to consider dragonfly or similar. If one had, say, 100k GPUs, a dragonfly might partition them into ~200 groups of 500 GPUs each, with a certain number of inter-group links. It could be beneficial if the communication is mostly localized (which in AI training, often many GPUs are participating in collective ops across all nodes, so traffic isn't that local). Thus, Dragonfly is more common in HPC simulation where communication patterns can be mixed local/global. For AI, fat-tree is still preferred to get full bisection bandwidth for all-reduce scaling.

- **BCube and Multi-Torus** – BCube is a multi-layer network topology proposed for data centers (originating from Microsoft research) where servers have multiple NICs connecting to multiple layers of mini-switches. It provides a lot of parallel paths and is highly fault-tolerant. A BCube(n,k) structure can connect large numbers of nodes with only small switches. While elegant, BCube hasn't seen mainstream adoption for GPU clusters – primarily because typical GPU servers have at most 2 NICs (for dual-plane), not the multiple ports needed to build a BCube. However, as servers start having

more network adapters (for example, some HGX systems might have 4 NICs – two per GPU pair, etc.), these alternative topologies could emerge. **Fat-tree topologies can be augmented with extra links** to improve fault tolerance in similar fashion. We mention BCube mainly as a conceptual alternative; modern AI clusters are not using BCube explicitly, but they do use **multi-homed connections** (each server to multiple switches) which is essentially a simplified BCube (depth 1).

Resiliency – At 100K GPU scale, network failures are inevitable (links, transceivers, switch ports *will* fail). The architecture must localize failures and prevent them from cascading. This is where **"pod" or cell-based design** comes in: the cluster can be divided into pods (say 1024 GPUs each) with well-defined intra-pod bandwidth and limited inter-pod bandwidth. If a spine switch fails, it might isolate one pod's external connectivity but not kill the whole cluster. Additionally, **dual-plane fabrics** ensure that no single switch failure disconnects nodes entirely. Routing protocols (for Ethernet, usually CLOS/ECMP or even BGP in large DCs) should converge quickly on failures. InfiniBand uses a subnet manager that can reroute around failed links as well. *Blast radius* is a key concept – for example, not scheduling a single training job across the entire 100k GPUs blindly; instead, perhaps limit a job to one pod or one supernode, so that a failure beyond that scope doesn't crash the job. This is more of a software scheduler concern, but it interplays with network design. Generally, **highly scalable clusters are built with failure containment zones** – whether by pods, or by using chassis-based switches that isolate linecard failures, etc. We will see in Section 6 how multi-pod layouts and scheduling strategy can mitigate failures.

Before moving on, a quick note on **BlueField-3 DPUs** and power: as mentioned, these DPUs are integral to Ethernet-based designs for offload and isolation. Each BlueField-3 DPU can consume up to about **150 W** (they have an 8-core ARM SoC and high-speed NIC on board). Servers with **two BF-3 DPUs may draw ~300 W just for NICs**. This is significant – roughly 5–10% of a server's power. It's a trade-off: one pays that power cost to save CPU overhead and gain security features. In cluster power calculations, planners must include DPU power. For instance, a 1000-node cluster with 2 DPUs each is consuming 300 kW just in DPU power. However, not using DPUs would mean using CPU cores for networking (less efficient) and potentially worse GPU utilization (costlier in lost productivity). In multi-tenant sovereign clouds, DPUs are often considered essential despite the power hit, because they enable features like **zero-trust security, encryption, and bare-metal provisioning** without involving host CPUs. In the **GB200 NVL72 architecture, BlueField-3 DPUs are included** to handle cloud networking, storage offload, and security, underscoring that even these supernodes anticipate being part of larger multi-tenant environments.

High-Performance Storage: Feeding 100K GPUs

Massive GPU clusters place *unprecedented demand on storage systems*. Training jobs can read **petabytes of data and write checkpoints at terabytes per second** aggregate throughput. A rule of thumb: each GPU might require on the order of **GB/s of I/O** (especially during data ingest for training). For 100,000 GPUs, storage must scale to **hundreds of TB/s** to keep them busy – an extreme challenge. Thus, a **tiered storage strategy** is used:

Tier-1: Ultra-High-Performance Distributed Storage – This typically comprises an all-flash parallel file system or object store that can deliver very high throughput and IOPS to the compute nodes with low latency. Key players and technologies here include **DDN (Lustre & EXAScaler/APEX)**, **VAST Data**, **WEKA Data**, **IBM Spectrum Scale (GPFS)**, and **Azure's FS for AI (for cloud)**, among others. We'll focus on DDN

and VAST as requested, since they are commonly proposed for large AI clusters, and also discuss how they compare, along with the option of open-source Ceph for a capacity tier.

- **DDN (DataDirect Networks)** – A veteran in HPC storage, DDN's solutions (often branded A³I when bundled for AI) traditionally use Lustre or their own parallel filesystem on arrays of SSDs or NVMe drives. DDN has achieved top positions in the IO500 benchmarks. In an AI context, DDN recently introduced **Infinia**, a high-performance S3-compatible object storage purpose-built for AI data pipelines. One success story: DDN described a deployment for a "global AI leader" with *>10,000 GPUs scaling to 100,000+ GPUs*, where other storage failed to keep GPUs fed. The DDN Infinia solution delivered **>1.1 TB/s sustained read throughput with 150,000 simultaneous connections**, achieving near-100% GPU utilization. It also provided extremely fast metadata ops – listing 30 million objects per second (75× faster than prior solutions) – eliminating delays when an AI job starts reading a huge dataset. These numbers are astounding: **1.1 TB/s to 100k GPUs** means each GPU effectively gets ~11 MB/s if all are active (but in practice, not all GPUs read at max simultaneously; the key is the storage can handle spikes and heavy concurrency without slowdowns). The result for the client was *"near 100% GPU productivity, unlocking $8M–$35M annual savings"* in GPU time by eliminating idle waits. DDN also notes that when *"every experiment costs $50k–$500k, storage isn't infrastructure – it's survival"*. This underlines how high-performance storage, though expensive, **pays for itself by preventing idle GPUs** (which are far costlier). Technically, DDN achieves this via a scale-out architecture: Infinia can scale to **150 PB+** capacity across many nodes, and handle multi-petabyte daily ingest. The storage exposes a unified namespace (object or POSIX via gateways) to simplify data access. DDN's solution in that case used an S3 interface with huge throughput and an ability to absorb **5–6 PB of new data per day** continuously for model training. Infinia likely uses NVMe drives and advanced indexing/caching to reach those numbers. For more conventional setups, DDN's **AI400X** appliances (parallel filesystem with NVMe) are often deployed – the NVIDIA SuperPOD reference BOM actually lists **DDN AI400X** systems as the certified storage, connected via 100 GbE or InfiniBand [24]. Each DDN AI400X can provide on the order of 40–60 GB/s, so dozens of them are clustered to get multiple TB/s.

- **VAST Data** – A newer entrant that has disrupted storage for AI with an **all-flash, scale-out NAS** design. VAST's architecture, **DASE (Disaggregated, Shared-Everything)**, puts all NVMe flash in shared enclosures and uses stateless server pods to present a global namespace. It leverages compression and deduplication to make flash cost-effective, claiming that **all-flash can actually be cheaper than HDD tiering** for many workloads (due to high data reduction and better utilization). VAST emphasizes simplicity: one big system for both hot and warm data, instead of separate tiers. For performance, VAST's latest systems use NVMeoF (NVMe-over-Fabrics) with RoCE and can saturate multiple 100 GbE links per server. VAST claims its storage can **"feed 100k+ GPU clusters at TB/s rates"**, eliminating data bottlenecks. In a marketing highlight, they mention >**50% lower TCO** for AI workloads by using their architecture. One reason is that VAST uses cheaper **QLC flash** (higher density, lower cost per TB) and compensates for QLC's speed limitations with heavy caching (e.g. each data is written once to persistent *Optane* storage class memory in older models, or now maybe CXL.memory pools, then destaged to QLC, etc.). This yields flash performance at near-disk pricing. VAST also integrates well with GPU-direct RDMA and Kubernetes, etc. The **real-world evidence**: VAST has been adopted by clients like Cloud AI providers (CoreWeave, Lambda Labs, etc.), and they partner with NVIDIA (they demonstrated using NVIDIA BlueField DPUs for isolated storage networks in a reference design). In terms of throughput, a single VAST cluster with dozens of storage servers can easily provide **multi-terabytes/s** – e.g. one published number is 2.4 TB/s from a VAST cluster

with 20 servers. Scaling further just means adding stateless servers (each with 2×100 Gb or now 200 Gb links). VAST's strength is **mixed workload performance** – it can handle both streaming reads and small random reads/writes efficiently, due to its global flash level cache and log-structured layout. For AI training, this means it can supply training data (often large sequential reads) while simultaneously absorbing checkpoints and random writes from many jobs. One potential drawback might be latency (anything involving heavy inline data reduction can add a bit of latency), but for large batch training, a few milliseconds is not an issue. VAST is often compared to **WEKA** and **IBM** in this space: WEKA is a high-performance parallel filesystem that also scales well (WekaFS holds some metadata performance records, and it uses clients on GPU nodes to accelerate IO, but requires more memory on clients). IBM Spectrum Scale (GPFS) is very mature and used in many HPC sites, but can be complex to manage at scale without the right expertise (and IBM is working on an all-flash Spectrum Fusion for AI).

DDN vs VAST – technical and commercial comparison: Both can deliver the needed performance, but via different philosophies. DDN (with Lustre/EXAScaler or Infinia) might be more **bare-metal optimized**, getting every ounce of throughput with relatively straightforward architecture (just many storage servers each with SSDs, and a parallel filesystem stripe). It may, however, require more manual tuning (ensuring Lustre stripe settings, etc., and it may not have advanced global compression). VAST is more **feature-rich** (inline data reduction, snapshots, encryption, multi-protocol NAS/Object access) and focuses on ease of scaling (adding nodes seamlessly). Commercially, costs depend on scale and discounts – DDN is usually a premium turnkey solution (often sold as appliances), whereas VAST sells a software+hardware package and claims better $/TB when their data reduction is effective (e.g. vision datasets with many duplicate images could see big savings). In a **mega-cluster scenario (100 PB+ storage)**, a DDN solution might involve a lot of OSS/MDS servers and can scale, but the management overhead grows, whereas VAST would treat it as one giant cluster. **Reliability**: both are enterprise-grade; DDN has years of mission-critical HPC use (which often lacks fancy GUIs but is robust), VAST is newer but designed with no single points of failure (their stateless controllers can all take over). For an AI cloud that wants *multi-tenancy* at the storage level, VAST's support for multi-tenant NFS buckets with secure isolation might be attractive, whereas Lustre is usually single-tenant (all users see all files, unless complex ACLs).

Ceph and Tier-2 storage: Many sovereign cloud architects consider using **Ceph** for bulk storage, because it's open-source, flexible (offers object, block, and now high-performance CephFS for file). Historically, Ceph was viewed as **too slow for Tier-1 HPC storage** – its latency overhead and write amplification were problematic. But it has improved, and for **Tier-2 (capacity/cold storage)** it shines. As a distributed object store on commodity servers, Ceph can scale to exabytes relatively cheaply by using dense HDDs with erasure coding. A whitepaper by SoftIron notes: *"years ago Ceph would not be the best choice for Tier 1 storage… however, it is well-suited for use as Tier 2 storage"* in HPC. Ceph's performance can scale by adding more OSD nodes; it won't hit the per-node speeds of Lustre or DDN, but aggregated it can reach tens of GB/s with enough nodes. For example, an all-NVMe Ceph cluster built by OpenMetal achieved **71 GB/s** large-file read throughput (with 4.4M IOPS) using 24 NVMe drives across 8 servers. On the HDD side, CERN has famously used Ceph for years to store physics data: they run **tens of PB on Ceph** (one instance was 65 PB using 10,800 drives), showing it can manage *huge* datasets reliably. In AI, one could envision Ceph (in object mode) storing raw training data (images, text dumps, etc.) and then when a training job starts, the needed subset is staged to the high-speed tier (DDN/VAST) either manually or via automated tiering. Ceph even has a **tiering feature** where a fast pool (NVMe) can serve as a cache and spill to a capacity pool (HDD), though tuning that for AI workloads might be tricky. Some HPC centers also use **CephFS** (Ceph's POSIX file system) for shared home directories or project spaces – not super fast, but convenient for general use.

Multi-Tier Architecture – Therefore, an optimal mega-cluster storage might combine **a fast tier and a capacity tier**:

- **Fast Tier** (Tier-0/1): All-flash, RDMA-enabled, co-located with GPUs. Provide ~1–5 TB/s throughput and high IOPS. Example: 10 racks of VAST or DDN, each delivering ~100 GB/s, totaling 1 TB/s. Or a handful of DDN Infinia systems achieving >1 TB/s as described. These typically use **RoCE or InfiniBand** (GPUDirect Storage) to let GPUs read data with minimal CPU involvement. Indeed, in the DGX SuperPOD design, the **storage fabric is separate** and can be either 100 Gb or 200 Gb Ethernet with RoCE or NDR InfiniBand, connecting to storage arrays. The goal is to sustain high throughput for active training data and checkpointing. The storage in this tier often sits on higher-cost media (NVMe, Optane, etc.), so capacity might be smaller (a few PB).

- **Capacity Tier** (Tier-2): Mostly HDD (possibly with SSD cache) object storage. Provide massive capacity (hundreds of PB) at lower cost, albeit with lower bandwidth (~10s of GB/s). Ceph, Hadoop HDFS, or even tape archives fall here. In sovereign contexts, Ceph is popular because it's self-hosted. This tier holds datasets at rest, older checkpoints, etc. When a new training needs data, either the training framework streams from Ceph (if it can tolerate slower throughput) or data is pre-copied to Tier-1. Some frameworks (like PyTorch's dataloader) can actually read from an object store directly, possibly saturating a few 100 Gb links, but for 100k GPUs it likely isn't enough – thus, staging to flash is common.

- **MetaData and Caching**: There might also be an intermediate caching layer – e.g. using node-local NVMe or RAM as cache (some sites use **Lustre HSM** or **Burst Buffers**). However, newer storage like VAST can reduce need for explicit cache by handling it internally (they essentially provide tiered storage within their system using NVRAM/optane as write cache and SSD as storage, etc.). NVIDIA's SuperPOD certified storage systems (DDN, VAST, IBM, etc.) all include some form of client-side cache too (e.g. using GPU memory or system RAM to cache recently used data).

Networking for Storage – In large designs, **separating the storage network from the main compute network** is advisable. For instance, in the DGX SuperPOD RA, they describe an **InfiniBand storage fabric** (option 1) using dedicated NDR switches to connect DGX nodes to storage servers. The IB storage fabric had a 4:3 oversub ratio in one example (since not all GPUs will saturate storage concurrently). Alternatively, an **Ethernet storage fabric** using 200 Gb or 400 Gb switches (Spectrum SN5600, etc.) is supported. In either case, the storage network is separate from the MPI/compute network to isolate heavy I/O traffic (which can be bursty during checkpoints) so it doesn't interfere with inter-GPU communications. This aligns with best practice: *"user storage differs from high-performance storage… user NFS share on in-band (management) network"* for certain data [25], while the high-perf storage is on a dedicated fabric. Essentially, large sites often have **three networks**: 1) compute fabric (IB or RoCE), 2) storage fabric (IB or RoCE, possibly lower speed), 3) management (1 GbE or similar for IPMI, etc.). Some modern DPUs can even combine these logically while isolating traffic via QoS or separate VLANs if using one physical network for both compute and storage (Spectrum-X could conceivably partition bandwidth per traffic class). But to keep things simple and scalable, physically separate networks are common at the high end.

Blast Radius & Data Management – On the storage side, blast radius refers to ensuring a single storage node failure doesn't stall training. Distributed storage with erasure coding or replication ensures that if a disk or even an entire storage server fails, data is still available (with some performance hit until rebuilt). Ceph, Lustre, VAST all have mechanisms for this (Ceph replicates 3x or erasure codes across servers; Lustre

can use dual-OSS mirroring; VAST writes two or three copies of data across failure domains). For multi-tenant clusters, one might even **partition storage by user or project** (e.g. each tenant gets a separate Lustre filesystem or a namespace in Ceph). But more often, all GPUs share a common high-performance file system for simplicity, with proper POSIX permissions or tenant IDs for security. Given sovereign cloud needs, encryption at rest is often required – VAST and Ceph offer encryption options, as do DDN (especially on S3 object data).

Finally, **cost**: Storage can be a large fraction of cluster cost. At multi-petabyte scale, flash storage costs many millions of dollars. For example, 10 PB of flash at ~$1000/TB is $10M. If data reduction halves the effective cost (VAST's 50% TCO claim), it's still ~$5M. HDD storage is cheaper (10 PB HDD maybe $1–2M), but as explained, the hidden cost is slower turn-around and potential GPU idle time which can cost far more. In Europe, there are efforts to use more open-source storage to avoid vendor lock-in (e.g. some EU HPC centers use **BeeGFS** or **Lustre** on commodity servers). Those can reduce upfront cost but require in-house expertise to tune for AI. Vendor solutions like DDN/VAST come with that expertise baked in – which is valuable when time-to-results is crucial.

In summary, a mega GPU cluster today often uses **DDN or VAST all-flash storage for the performance tier**, and possibly something like **Ceph for a capacity tier**. Both DDN and VAST can meet technical requirements (multiple TB/s throughput, millions of IOPS). DDN brings HPC-proven stability and raw speed (especially for large sequential IO), while VAST brings innovative efficiency (data reduction, single-tier simplicity) and strong random IO performance. Commercially, each has wins: DDN perhaps in more conservative HPC/enterprise, VAST in agile AI startups and some enterprise. It's not uncommon for bidders of an AI system to consider **both**: for example, one might get quotes for "DDN ExaScaler 30PB + Lustre" vs "VAST 30PB", comparing overall cost including power (VAST's claim of 40% lower power due to fewer total drives and better efficiency is something to evaluate). Actually, VAST specifically advertises *"40% reduction in electricity cost for data center"* when using their liquid-cooled NVMe systems [21] [26] – which dovetails with the sustainability theme.

Best Practices in Large-Scale AI Cluster Design

Designing a 10000–200000 GPU cluster is not just about hardware capacities; it requires **careful planning of topology, power, cooling, and operational strategy** to ensure scalability and resiliency. Here we outline some best practices and considerations, incorporating ideas mentioned (blast radius, multi-tenant, power, cooling, etc.):

Pod Architecture and Blast Radius

Instead of treating 100k GPUs as one monolithic cluster, it's wise to partition into **pods** or **cells**. A pod might be, for example, **1024 GPUs (128 nodes with 8 GPUs each)** or a certain number of racks that form a logical grouping (like one DGX SuperPOD unit). Within a pod, full-bandwidth networking is provided, but inter-pod bandwidth might be limited or routed through a higher-level core switch. This hierarchy has multiple benefits:

- **Failure Isolation**: If a network issue (misconfigured switch, heavy traffic spike, etc.) occurs in one pod, it can be contained without impacting the entire system. Similarly, a software bug that crashes nodes might be isolated if each pod runs separate job schedulers or domains.

- **Upgradability**: Pods can be taken down for maintenance or upgraded one at a time (rolling upgrade) while others continue running, minimizing whole-system downtime. This is important for applying firmware updates, replacing end-of-life hardware, etc., in a phased way.

- **Scheduling Efficiency**: Many training jobs do not need 100k GPUs; they might run on 512, 1024, 2048 GPUs. By aligning pod size with typical job sizes, you can schedule jobs to reside fully within pods, thus containing their high-traffic communications inside a pod's network (which is usually a single local switch domain). This reduces inter-pod traffic, avoiding core network bottlenecks. It also simplifies performance predictability.

For example, Meta's **RSC (AI Research SuperCluster)** (which had ~6080 GPUs initially) was logically divided (they didn't publish details, but one can infer multiple racks were grouped with internal InfiniBand and then connected via higher-level switches). Similarly, Nvidia's DGX SuperPOD reference of ~20 DGX nodes (160 GPUs) can be seen as a "pod" that can be replicated and connected via a larger network for scale-out [11].

Blast radius in context: The idea is to limit the scope of worst-case failures. In a multi-tenant environment, you also want to avoid one tenant's actions affecting others. Network storm caused by tenant A should ideally be isolated to A's partition. Using DPUs helps enforce that (rate-limiting) at the network level. At the job scheduling level, one might reserve certain pods for high-priority or sensitive workloads so that if another pod is overwhelmed, it doesn't delay critical jobs.

Resilient Networking (Fabric A/B)

As extensively discussed in the networking section, deploying **redundant network planes** is a must. A practical guideline: **each server needs at least two high-speed NIC/IB ports** connecting to independent switches. Moreover, those switches should ideally go to separate power feeds (so that a PDU failure doesn't take out the entire network plane). The **twin-plane fat tree** is the de facto solution in both IB and Ethernet world for large systems.

In addition to physical redundancy, networks should use features like **adaptive routing** to handle congested links and **link-level retry** (both IB and modern Ethernet have link-layer FEC and retry that correct most errors). Multpathing (ECMP or IB multipath) ensures traffic can route around failures quickly.

When using Ethernet, one should configure Data Center Bridging (DCB) properly: enable PFC on only the RDMA priority traffic class to avoid head-of-line blocking on other traffic; or use ECN marking so that senders slow down before drops occur. With InfiniBand, similar care is needed to configure virtual lanes or use QoS if mixing traffic types.

One more layer of resilience: **management network** separation. The cluster's out-of-band control (IPMI/iDRAC for servers, switch management ports, etc.) should be on a separate network (usually a low-speed Ethernet network connecting to a bastion). This way, if the main fabric is saturated or down, admins can still reach all components to diagnose and fix. NVIDIA's RA recommends an OOB management network rolled up to spine as a VXLAN or separate VLAN [23], and Dell IR solutions no doubt integrate management for the whole rack that can be accessed externally even if the data plane is off.

Power and Cooling Infrastructure

Power Delivery: At mega scale, power availability is a gating factor. A 100k-GPU data center can easily draw **100+ megawatts**. For instance, the 18k-GPU Fluidstack/Mistral cluster is starting at 40 MW and planning to expand beyond 100 MW. Ensure early collaboration with utilities and grid operators. Often, **new substations** or HV lines are needed. Large projects may secure **dedicated power contracts** – e.g. the French government partnering on the 1 GW AI center suggests allocation of nuclear generation capacity at favorable rates. In France, thanks to nuclear, bulk power can be relatively affordable (~€50–70/MWh for industrial tariffs), whereas in Germany prices might be higher and more volatile due to renewables and gas (unless long-term PPAs are signed). Such differences impact OPEX significantly – e.g. running 100 MW year-round at €50/MWh costs €43.8M/year in electricity, whereas at €100/MWh it's €87.6M/year. These costs dwarf many capital expenses, so countries like France offering stable low-carbon power is a major draw for locating AI clusters (hence the recent flurry of announcements in France).

From a design perspective, power must be delivered to racks possibly at medium or high voltage then stepped down. **Busbar DC distribution** inside racks (as NVIDIA supports) can improve efficiency by reducing AC/DC conversion losses [11]. Having **dual-power feeds (A+B)** to each rack and each critical component ensures that a single power source failure (utility feed, transformer, UPS, PDU) doesn't cut off the cluster. Each IR rack solution (Dell, Supermicro) includes redundant PSUs and PDUs (Supermicro's NVL72 rack has 8 power shelves, likely 4 on feed A and 4 on feed B, each capable of full load so that the total 132 kW can be supplied from either side if needed) [27]. Also, backup power (generators, etc.) should be planned if cluster uptime is mission-critical, but the power draw is so high that running generators for 100MW is nontrivial (would require many large gensets). Some sites might instead rely on grid redundancy and battery backup for short outages.

Cooling: As we've repeated, **direct liquid cooling (DLC)** is now the norm for high-density AI clusters. Air cooling simply cannot handle ~30 kW per rack with reasonable efficiency, let alone 200+ kW/rack that these GPU racks require. Liquid cooling comes in a few flavors:

- **Cold Plate (Direct to Chip)** – Most vendors (Dell XE9685L, Supermicro, HPE) use cold-plate loops to take heat from GPUs, CPUs (and sometimes NICs/DPUs) and circulate a coolant (often water or glycol mix) to either a CDU (Coolant Distribution Unit) or directly to facility water. For example, Dell IR7000 has dual liquid manifolds per rack and can capture nearly all heat in liquid [7]. Supermicro's solution includes CDUs in-rack for up to 250 kW and can even tie into cooling towers [28] [14]. Cold plate cooling typically uses water at maybe 30°C–45°C inlet (warm water cooling) to maximize use of free cooling (no chillers) – which improves PUE.

- **Immersion Cooling** – Not mentioned explicitly in our sources, but immersion tanks (submerging servers in dielectric fluid) are an alternative. Some startups and European projects explore immersion for AI clusters, as it can allow even higher power density (multiple servers in one tank, up to 100 kW in a small footprint). However, retrofitting immersion is complex, and major vendors currently favor cold-plate because it's easier to integrate into standard rack form factors.

- **Liquid-to-Air** – As an interim solution where facility liquid loops are not available, companies like Supermicro offer **liquid-to-air heat exchangers** that fit in-row or in-rack [16]. These use a radiator with fans to dump heat from the liquid into the room air. For example, a 240 kW liquid-to-air unit can

allow deploying a liquid-cooled rack without hooking to an external chiller plant [16]. The efficiency is lower (since ultimately air cooling the room), but it eases deployment in existing data centers.

The goal is to achieve a low **PUE (Power Usage Effectiveness)** despite enormous power density. Many AI datacenters target PUE ~1.1 or better, even at scale. Using warm-water liquid cooling, waste heat can sometimes be reused – some European sites pipe warm water to district heating or absorption chillers. That ties into sustainability requirements (governments may favor projects that plan to reuse heat to avoid just dumping it). With IR7000 capturing ~100% heat in liquid, it's feasible to harvest that heat.

Facility and Layout: Traditional data center layouts (rows of 42U racks with hot/cold aisles) are evolving. For instance, OCP racks like IR7000 are taller (sometimes 48U or 50U) and often require overhead cabling and piping for liquid. Planning space for CDUs, coolant pumps, and possibly rear-door heat exchangers (if used) is needed. Weight is also a factor: a rack filled with liquid cooling gear and GPUs can weigh >1500 kg. Raised floors might not suffice; many modern designs go with slab floor and overhead power/cooling distribution.

Additionally, **fire suppression** and safety need to account for liquids and high power. Fluids used are typically water-based, which is non-flammable, but leaks are a hazard (thus many quick-disconnect couplings and leak sensors are installed).

From an operational standpoint, **spare capacity** for power and cooling must be considered – e.g. design for 20% headroom in cooling so that on a hot day or future higher-TDP upgrades, the system can cope. Dell explicitly says IR7000 is "future-ready" for larger CPUs/GPUs beyond current gen [7], implying the cooling and rack design can take even more wattage if needed (e.g. 480 kW now, maybe 600 kW in future with upgraded CDU).

Multi-Tenancy and Security

For sovereign cloud scenarios, multi-tenancy is a given – multiple organizations or teams share the cluster, possibly with sensitive data. Incorporating **security isolation** at various levels is vital:

- **Network segmentation** – Using DPUs (BlueField) to implement **zero-trust networking** inside the cluster. Each tenant's traffic can be isolated via VLAN or VRF, and only allowed to communicate where necessary. BlueField can act as a firewall for each server, offloading security policies. Also, features like **GPUDirect RDMA** can be controlled so that one tenant cannot inadvertently read another's GPU memory over RDMA. In GB200 NVL72, BlueField-3 DPUs are specifically noted to enable *"composable storage, zero-trust security and GPU compute elasticity in hyperscale AI clouds."*.

- **Storage security** – Multi-tenant cluster will have data from different parties. Solutions include encryption at rest (with keys per tenant) and strong authentication for access to the storage. Ceph can isolate pools per tenant; Lustre/GPFS can use filesystem permissions or even separate filesystems. VAST supports multi-tenant namespaces with secure delegation. Strict auditing (knowing who accessed what data when) should be in place, likely integrated with the AI cloud management.

- **Scheduling & Compute isolation** – Kubernetes or SLURM might be used to schedule jobs in containers or VMs. Ensuring one tenant's jobs do not interfere or see another's is crucial (for example, side-channel attacks on GPUs are an emerging consideration – so one might choose to not

co-locate different tenant containers on the same physical GPU, etc.). Some cloud-like deployments dedicate entire pods or sets of GPUs to a tenant for the duration of a job for stronger isolation.

- **Compliance** – Sovereign clouds in EU often need certifications (SecNumCloud in France, C5 in Germany, etc.). This includes physical security (secure locations, access control) and data sovereignty (data not leaving the country, which is given if on-prem in country). The cluster design might incorporate encrypted links between sites if multi-site, and remote attestation for firmware (ensuring no tampering). While beyond pure hardware scope, it's worth noting that employing DPUs and isolated management networks actually helps with compliance: e.g., one can run security monitoring software on DPUs (since they have ARM cores) to scan traffic for anomalies without burdening hosts.

Costs and TCO Considerations

We touched on power cost, but overall **TCO (Total Cost of Ownership)** includes capital expenditure on hardware, facilities, and operational costs (power, cooling, staffing, maintenance). For mega AI clusters, **economies of scale** can reduce per-unit costs (bulk procurement discounts, optimized cooling for high density, etc.), but there are also diminishing returns if not fully utilized.

Some best practices: - **Utilization is King**: The single biggest cost is the GPUs themselves (which can be ~$10k–$20k each for H100/Blackwell). If you have 100k GPUs, that's easily $1–2 billion of hardware. Any idle time is extremely costly. So design choices that improve utilization (like faster storage reducing wait times, or good scheduling that packs jobs efficiently) directly improve ROI. This is why the storage solutions emphasize eliminating GPU idle time, and why networking is built to minimize wait on gradients, etc.

- **Phased Deployment**: No one deploys 100k GPUs in one go without testing smaller scales. Typically, one might deploy a smaller cluster (say 1k GPUs), run pilot workloads, then scale out in phases (5k, 10k, etc.). Each pod added provides learning. Reference architectures help shorten this cycle – e.g., using Dell IRSS racks delivered pre-integrated in **"days rather than weeks" (3× faster deployment)**.

- **Energy Reuse & Incentives**: In Europe especially, clusters that can reuse waste heat might get incentives or at least goodwill (for example, a data center near a town that heats homes with waste heat could get tax breaks). It's wise to consider locating near facilities that can absorb the heat (district heating network, greenhouses, etc.). Alternatively, being near a power source (like next to a power plant) can reduce transmission costs and ease getting large power capacity.

- **Future-proofing**: Hardware evolves fast – the design should allow inserting next-gen components. Dell IR7000 explicitly states it is multigeneration capable [10] . This might mean oversizing power and cooling slightly now, and ensuring network backplanes can scale (e.g. lots of extra fiber capacity installed, or modular switch chassis that can go from 400G blades to 800G blades later). The modular OCP rack approach is beneficial here.

- **Commercial models**: Some sovereign cloud providers might not purchase all hardware up front; they might use leasing or as-a-service models (for instance, Dell and HPE both offer HPC-as-a-service where they retain ownership and charge monthly). This can shift CapEx to OpEx and provide flexibility to upgrade hardware in a few years. It's beyond technical design, but impacts how you plan expansions.

Finally, **software**: A cluster of this scale needs excellent software management – from cluster schedulers, provisioning (automated bare-metal provisioning for thousands of nodes, possibly using tools like OpenStack Ironic or Kubernetes Cluster API), monitoring (Prometheus/Grafana collecting millions of metrics), and automation for failover. **AI-specific software** like NVIDIA's Base Command or OpenAI's tuning orchestration might be in play to manage distributed training across so many GPUs. Though not hardware per se, the cluster architecture should accommodate this – e.g., ensuring that the control plane (master nodes, head nodes) are on resilient servers and possibly not overloaded. Some designs include dedicated "management nodes" or login nodes that are separate from the GPU worker nodes. This keeps user interactions and scheduling decisions off the performance-critical path of GPUs.

In conclusion, building an on-premises mega GPU cluster requires an **end-to-end approach**: cutting-edge GPU servers, a robust high-speed fabric (InfiniBand or Spectrum-X Ethernet with RoCEv2), a tiered storage solution that can supply data at massive scale, and careful planning for power, cooling, and failure domains. The latest reference designs from NVIDIA, Dell, Supermicro, etc., give a blueprint for such "AI factories." By following best practices – such as dual-plane networks, liquid cooling, pod segmentation, and integrating DPUs for offload – one can achieve a scalable, efficient, and secure AI infrastructure. The result is the ability to train frontier AI models (for healthcare, defense, large language models, etc.) on European soil, in a sovereign cloud, at a scale and cost-efficiency on par with the world's leading AI labs – truly a strategic capability for the coming decade.

References (Technical Documents & Sources)

The table below lists all referenced documents and sources used in this research, with links (including PDFs where available):

Source / Document	Description	Link
Dell IR5000 & IR7000 Integrated Racks – *Dell Technologies brochure/brief*	Overview of Dell's Integrated Rack Scalable Systems (IR5000 standard 19″ and IR7000 OCP 21″ racks), including GPU densities and cooling.	
Dell "Accelerating AI Innovation" – *Dell Blog (Nov 2024)*	Dell announcement at SC24 introducing PowerEdge XE7740, XE9685L servers and IR5000/IR7000 racks; first support for NVIDIA GB200 NVL4 in IR7000. Discusses 96 GPUs/rack (liquid cooled) etc.	
Dell Press Release Oct 15, 2024 – *"AI Factory with Advanced Cooling"*	Dell Newsroom release detailing IR7000 features, PowerEdge XE9712 (72-GPU NVL72 rack-scale node) and M7725, plus PowerScale storage updates. Confirms 480kW liquid cooling, 72 GPU acts as one GPU, 30× inference boost, 25× efficiency vs H100.	3 9

Source / Document	Description	Link
NVIDIA Blackwell Architecture & Grace Superchip – *NVIDIA Press Release (PDF)*	NVIDIA Newsroom PDF announcing Blackwell GPU architecture and Grace-Blackwell (GB200) Superchip. Contains NVLink5 bandwidth (900 GB/s per direction) and description of GB200 NVL72 system with 72 GPUs + 36 CPUs + BlueField-3 DPUs (30× performance, 25× efficiency vs H100 for LLM inference).	1 29
NVIDIA DGX SuperPOD Reference Arch (DGX B300, 2025) – *NVIDIA RA-11337-001 PDF*	Technical reference architecture for a DGX SuperPOD using DGX B300 systems, Spectrum-4 Ethernet or Quantum InfiniBand, and DC busbar power. Includes network topology (twin-plane fat-tree), storage fabric options (RoCE vs IB), and sample rack layouts.	
NVIDIA Spectrum-X Networking Platform – *NVIDIA GTC 2024 Announcements*	Press release and related info on Spectrum-X and Quantum-X 800 Gb/s switches, BlueField-3 "SuperNIC" DPUs, and early adopters (Azure, Oracle). Emphasizes Spectrum-X for multi-tenant AI clouds with performance isolation.	
Supermicro NVIDIA GB200 NVL72 SuperCluster – *Supermicro October 2024 Brochure (PDF)*	Detailed brochure for Supermicro's "Exascale-in-a-Rack" 72-GPU Grace-Blackwell SuperCluster. Provides specs: 72× B200 GPUs + 36 Grace, 130 TB/s NVLink network (9× switches), 132 kW power, liquid cooling (250 kW CDU or liquid-air options), BlueField-3 and 800G network ready.	14 13
Supermicro Case Study – Lambda Labs (Aug 2025) – *PDF*	Success story of Lambda upgrading to Supermicro servers with NVIDIA Blackwell. Lists server models: A21 (5th Gen Xeon + HGX B200 8-GPU with 180 GB HBM each), etc., and mentions a Supermicro AI Supercluster with 72 Blackwell GPUs and 9 NVLink switches (1.8 TB/s interconnect, 13.5 TB HBM).	2 2
DDN Infinia for AI Success Story (May 2025) – *DDN Blog*	"Inside a $70B AI Giant: How DDN Infinia…" – Describes a large AI company's storage challenges and DDN's results: 1.1 TB/s reads across 150k connections, 75× faster metadata, near 100% GPU utilization, scaling to 100k GPUs, etc. Great insight into storage performance requirements and GPU cost savings.	
VAST Data "AI Operating System" Announcement – *VAST website (2024)*	VAST marketing content highlighting its platform's ability to feed 100k+ GPU clusters at TB/s and >50% TCO savings. Useful for understanding VAST's claims on scale, performance, reliability (DASE architecture, multi-tenant, etc.).	

Source / Document	Description	Link
Fluidstack 18K-GPU Supercomputer in France – *InsideHPC News*	News article on Fluidstack & Eclairion building Europe's largest GPU cluster (18,000 GPUs, 40 MW) for Mistral AI (startup), with expansion to 100 MW and an eye toward 1 GW. Emphasizes France's sovereign AI push and nuclear-powered, decarbonized energy.	
Fluidstack 1 GW MoU with France – *DataCenterDynamics (Feb 2025)*	Article covering Fluidstack's agreement with French government for a $10.4B, 1 GW AI datacenter by 2026, with ~500k next-gen AI chips. Includes Macron's quote on France's AI leadership and nuclear energy benefits.	
SoftIron Ceph for HPC White Paper – *HPCwire/SoftIron*	"Thought Ceph was a bad fit for HPC? Think again." – Discusses how Ceph has evolved for HPC use, suggesting it as good Tier-2 storage. Notes Ceph's scalability and that it was historically not used for Tier-1 but is now viable for large datasets (CERN's usage, etc.).	
Dell PowerEdge XE9685L Server – *Dell Product Info (via Dell blog)*	Details on Dell's 4U direct-liquid-cooled server with 8× H200 or B200 GPUs and dual AMD EPYC 9005, used in IR5000. Indicates 3× higher LLM training performance vs prior gen, and highlights 96 GPUs/rack capability.	
NVIDIA HGX B200/B300 Platform Info – *NVIDIA Compute*	Specifications on NVIDIA's HGX platforms for Blackwell (B200: 8-GPU baseboard, likely ~700W/GPU; B300: "Ultra" 8-GPU with 1100W GPUs). Confirms support for 400 Gb/s networking and NVLink Switch system connectivity between boards. (Derived from NVIDIA announcements, e.g., UncoverAlpha analysis).	30
Meta AI Research SuperCluster (RSC) – *Meta blog (Jan 2022)*	*(Contextual reference)* Described Meta's 6080-GPU cluster using Nvidia A100, InfiniBand networking, and plans to scale to 16,000 GPUs. While not directly cited above (due to older hardware), it provides background on big pod design in industry.	N/A (Meta AI Blog)
Elon Musk on AI Power & Transformers – *SemiAnalysis (Mar 2024)*	*(Contextual reference)* Analysis of AI datacenter energy demands, quoting Elon Musk on limits (transformers, grid). SemiAnalysis data on AI compute growth (50–60% QoQ). Illustrates concerns about infrastructure scaling which validate design precautions we discuss (power being a limit).	

1 29 nvidianews.nvidia.com
https://nvidianews.nvidia.com/_gallery/download_pdf/65f8a7843d633205563719fc/

2 Lambda Selects Supermicro GPU Optimized Servers
https://www.supermicro.com/CaseStudies/Success_Story_Lambda_NVIDIA_Blackwell.pdf

3 4 5 7 8 9 10 22 Dell AI Factory Transforms Data Centers with Advanced Cooling, High Density Compute and AI Storage Innovations | Dell USA
https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2024~10~dell-servers-storage-at-ocp.htm

6 12 13 14 15 16 17 18 19 20 21 26 27 28 30 Supermicro NVIDIA GB200 NVL72 SuperCluster
https://device.report/m/97f063254b010f7301978a5808a604255f60d249da50e16e8b875b8ef356878c

11 23 24 25 NVIDIA SuperPOD DGX B300 Systems, Spectrum-4 Ethernet and DC Busbar Power Reference Architecture
https://docs.nvidia.com/dgx-superpod/reference-architecture/scalable-infrastructure-b300/latest/_downloads/5fe4960ce43a21f33d6a5919d57bd583/RA11337001-DSPB300-ReferenceArch.pdf