ChatGPT

Designing a Next-Generation GPU Supercluster for Multi-Tenant AI Training

Overview and Goals

Building a state-of-the-art GPU supercluster for large language model (LLM) training requires integrating bleeding-edge compute, networking, and storage technologies with robust multi-tenant security and manageability. The goal is to architect an **AI "factory"** – similar to clusters used by OpenAI, Anthropic, etc. – that can train trillion-parameter models across dozens or hundreds of GPUs as a single unit, while securely serving multiple tenants side by side. Key design objectives include extreme GPU density and performance, high-bandwidth low-latency interconnects (without relying on InfiniBand), scalable storage for massive datasets, efficient cooling/power for racks drawing 100+ kW each, and a software stack enabling multi-tenant scheduling, isolation, and compliance (e.g. French SecNumCloud security standards).

In summary, the cluster will center on NVIDIA's latest **Grace-Blackwell** GPU superchips (GB200 and GB300 series) and their OEM implementations, augmented by 400–800 Gb/s Ethernet networking (RDMA over Converged Ethernet, RoCEv2), and tiered high-performance storage (WekaIO, VAST Data, DDN, Pure Storage, etc., plus a cost-efficient Ceph tier). We will detail **hardware specifications** (GPU nodes, reference rack designs, power/cooling, network fabrics, storage arrays) as well as **best practices** for network tuning (congestion control, BlueField DPU usage), **software/OS choices** (Ubuntu or DGX OS over RHEL, container orchestration, NVIDIA AI Enterprise, etc.), multi-tenancy and security (MIG partitioning, encryption in transit/at rest, confidential computing options), and vendor offerings from NVIDIA, Dell, Supermicro, HPE, Lenovo and others. All information is drawn from up-to-date 2024–2025 documentation and real-world reference designs, with links to original specs and whitepapers for further reference.

GPU Compute Infrastructure: NVIDIA Grace-Blackwell Superchips (GB200 & GB300)

At the heart of the cluster are NVIDIA's latest GPU superchips, which combine **Grace** data-center CPUs with **Blackwell** architecture GPUs in a tightly integrated package. The **GB200 NVL72** platform (launched late 2024) and newer **GB300 NVL72** (2025) are fully **rack-scale** units that connect 72 GPUs via 5th-gen NVLink within a single system domain. Each GB200 superchip consists of one Grace CPU paired with two Blackwell **B200** GPUs, whereas the GB300 uses enhanced **Blackwell Ultra (B300)** GPUs for even higher performance and memory capacity. Key specifications include:

- **GPU Count & Topology:** 72 Blackwell GPUs per rack unit, connected by NVLink switch fabric so that all GPUs communicate at full bandwidth. The NVLink-5 network offers an aggregate **130 TB/s** interconnect bandwidth in the 72-GPU domain, effectively treating 72 GPUs as one giant logical GPU for model parallelism. (By comparison, previous HGX systems had 8 GPUs per node.) Notably, NVLink5 can scale **beyond one rack** – up to 576 GPUs in one NVLink domain – by bridging multiple

chassis with NVLink Fusion links, enabling future expansion without falling back to slower Ethernet for GPU-to-GPU traffic.

- **GPU Performance:** Blackwell GPUs introduce a new Transformer Engine and lower precision modes (FP8, FP4) to vastly accelerate LLM training and inference. A 72-GPU GB200 NVL72 system delivers ~20 PFLOPS in FP4 tensor operations (with sparsity) and up to **30× faster** trillion-parameter LLM inference throughput than previous NVIDIA H100 clusters. The GB300 (Blackwell Ultra) goes further – its B300 GPUs have 1.5× the memory of B200 (288 GB vs 192 GB HBM3e each) and additional specialized throughput. NVIDIA reports the GB300 NVL72 achieves **10× the user-level responsiveness and 5× the throughput** of a Hopper H100 pod (overall ~50× more **AI inference output** per rack). This is pivotal for "AI reasoning" workloads where many inference queries run in parallel. For training, Blackwell also improves speed; e.g. GB200 showed ~4× faster training on some LLMs vs. H100 (thanks to FP8 precision and other improvements).

- **Memory Capacity:** Each Blackwell GPU comes with massive high-bandwidth memory. GB200 B200 GPUs have 192 GB HBM3e each (for ~13.4 TB total GPU memory across 72 GPUs). GB300's B300 GPUs have 288 GB each (≈20.7 TB total). In addition, the Grace CPUs contribute up to 512 GB each of LPDDR5x memory; with 36 Grace CPUs in the NVL72, that's ~18 TB of CPU memory. In total, a **single rack can have ~40 TB of fast memory** accessible (GPU + CPU) – extremely beneficial for huge models and datasets. The NVLink fabric and NVLink-C2C links between Grace and GPU allow unified memory access. (Grace's memory bandwidth is ~14 TB/s aggregate across the CPUs.)

- **Grace CPU:** The Grace processors (72 in GB200, 36 in GB300 where two GPUs share one CPU) are 72-core ARMv9 data center CPUs designed by NVIDIA. They provide high memory bandwidth to feed the GPUs (Grace has 512-bit LPDDR5X memory controllers) and are **2× more energy-efficient than leading x86 CPUs**. Grace can also accelerate certain data processing tasks: e.g. dedicated HW engines for compression, etc., yielding up to 18× faster data decompression or database scans vs. CPU-only systems. In a multi-tenant cluster, Grace CPUs handle pre-processing and can run secure OS services (with potential support for ARM Confidential Compute features, though details are emerging).

- **Power Envelope:** These are extremely power-hungry chips – each Blackwell GPU draws up to **1000–1200 W** (configurable to 1.2 kW), nearly double an H100. A GB200 NVL72 rack (72 GPUs + support hardware) consumes on the order of **120–132 kW** of power. The next-gen GB300 NVL72 is projected around **150 kW per rack** in power draw – a *single rack* as hungry as an entire small data center. We will discuss the power and cooling implications in the next section.

OEM Node Designs: NVIDIA provides the MGX reference architecture for modular servers, and OEMs have built solutions around the GB200/GB300. For example, each node/tray often contains **2× Grace CPUs + 4× Blackwell GPUs** (two GB superchips per tray). A full 72-GPU system might be composed of 18 such trays (18×4=72 GPUs) mounted in a rack chassis. These 18 nodes are then fully connected via NVLink Switch chips (the 5th-gen NVSwitch). NVIDIA's NVSwitch5 has 4+ TB/s port bandwidth, and 9 switch units can create the all-to-all topology for 72 GPUs. In practice, vendors like Supermicro, Dell, and Lenovo offer pre-integrated rack units: for instance, Supermicro's **SRS-GB200-NVL72** is a 48U rack housing 18 1U liquid-cooled server sleds with the full 72 GPU NVLink domain. Dell's equivalent is the **PowerEdge XE9712** server integrated into an IR7000 rack, also connecting 72 Blackwell GPUs with 36 Grace CPUs in one system.

In summary, the compute layer of our supercluster will use **NVIDIA GB200/GB300 NVL72 racks** as the basic building blocks. Each rack delivers on the order of **1+ exaFLOP** of AI compute (FP8/FP4) and tens of terabytes of HBM memory, acting as a single training unit for large models [1]. Multiple such racks can be interconnected via an Ethernet fabric for scale-out beyond 72 GPUs (though with some loss of efficiency compared to intra-rack NVLink). Next, we'll address how to power and cool these monsters, and how to network them without InfiniBand.

Rack-Scale Design: Power, Cooling, and Form Factor

Designing around 100–150 kW per rack means traditional air cooling and power delivery are pushed to their limits. **Liquid cooling is mandatory** to handle the thermal density of modern AI clusters. Key considerations and best practices include:

- **Rack Form Factor (21″ OCP vs 19″ Standard):** Several vendors have adopted the Open Compute Project (OCP) Open Rack Standard (ORv3) 21-inch width racks for AI clusters, as the extra width and depth allow larger, more power-dense server sleds and integrated liquid cooling manifolds. Dell's **Integrated Rack 7000 (IR7000)** is an example – a 48U, 21″ rack engineered for extreme GPU density and native liquid cooling. The IR7000 uses wider/taller sleds to fit the Grace+Blackwell trays and NVLink switches. It also features back-end busbars for power (no individual PSUs per node) and a liquid distribution manifold; this "cable-free" design improves serviceability and standardization. By contrast, some solutions stick to **19″ racks** for compatibility – e.g. Dell's IR5000 series racks support up to 72 GPUs (air-cooled) or 96 (liquid-cooled) in a standard EIA 19″ frame, and Lenovo's Neptune SC777 uses 19″ cabinets as well. The trade-off is slightly less density per rack with 19″. Our design can mix rack types as needed, but for maximal density we might use the 21″ format where facility allows.

- **Power Delivery:** A single rack drawing ~130 kW requires robust power distribution. Many designs use **busbar power shelves and high-voltage feeds**. For example, Supermicro's 72-GPU rack has **8× 33 kW power shelves** (each with six 5.5 kW PSUs) feeding a busbar – with N+N redundancy, the usable power is 132 kW. Lenovo's Neptune chassis includes **15 kW Power Conversion Stations (PCS)** that take AC input and produce 48V DC on a busbar; four PCS (N+1) per 13U chassis support ~54 kW, and up to 3 chassis per rack give **~162 kW per rack** capacity. We will need to provision adequate electrical capacity (likely 3-phase 400 V AC or 48 V DC bus) and upstream PDUs for each rack. Weight is also non-trivial – a fully populated liquid-cooled rack can weigh **1.5–1.8 tons (3000–3500 lbs)**, so floor loading and seismic bracing must be planned.

- **Liquid Cooling: Direct-to-chip liquid cooling** (water or coolant flowing through cold-plate loops on CPUs, GPUs, NICs, etc.) is the only practical way to remove ~100 kW+ from a single rack. Air cooling a 30 kW rack is challenging; at >100 kW it's practically impossible to stay in ASHRAE limits. Our cluster racks will use DLC for all high-power components. In-rack Coolant Distribution Units (CDUs) are often employed: e.g. Supermicro integrates a **250 kW CDU** in each rack, with redundant pumps, to circulate coolant through the 18 server nodes. This allows **nearly 100% heat capture** in liquid, meaning minimal hot air exhaust. The warm water can be cooled via heat exchangers connected to facility water loops or dry coolers. Dell IR7000 claims it can handle **up to 480 kW per rack** with its liquid cooling design (future-proofing for higher GPU counts or next-gen HPC ASICs). Lenovo's Neptune N1380 chassis similarly removes all heat via water so that **100 kW+ racks can run with no traditional CRAC cooling**. We should partner with cooling vendors (Schneider, Vertiv, Motivair, etc.,

with whom OEMs already collaborate) to deploy a suitable coolant distribution and monitoring system. Key metrics: coolant supply temperature (many DLC systems can use “warm water” ~35°C supply, reducing chiller needs), flow rates, water quality controls, and fail-safes for leak detection.

Example of a Dell Integrated Rack 7000 (IR7000) with direct liquid cooling. Each 21″ IR7000 rack can support up to **480 kW** of IT load by using liquid-cooled server sleds and rear-door liquid heat exchangers. These racks integrate features like busbar power distribution and centralized coolant manifolds to eliminate most cabling and piping clutter. Efficient cooling allows capturing ~100% of the heat in liquid, enabling deployment of **100 kW+ per rack** without additional air conditioning. (By comparison, a typical air-cooled rack might handle 10–30 kW.) Such high-density designs are crucial for modern AI superclusters in order to minimize floor space and maximize performance per square foot.

- **Thermal Monitoring and Redundancy:** With chips running at 1.2 kW each, thermal runaway is a concern. The cluster design should include granular monitoring of temperatures (GPU and CPU junction temps, coolant inlet/outlet, flow rates). Most OEM solutions come with management software to monitor and alert on cooling metrics. Our design should also consider redundancy in cooling – e.g. dual pump circuits, perhaps two separate CDUs (or an in-row CDU shared by multiple racks for redundancy). In case of a cooling failure, an emergency power-down of GPUs might be required within seconds to prevent overheating. This is part of cluster reliability engineering. (NVIDIA has noted some early reliability issues with NVLink backplanes on GB200, so we will incorporate improved diagnostics and burn-in testing as recommended.)

- **Noise and Environment:** These liquid-cooled systems generally have fewer fans (only for low-power components), making them quieter and more suitable for standard data centers (some can even be deployed in office environments if needed – e.g. NVIDIA’s DGX Station with similar tech is “whisper quiet,” though our racks will still have pump noise). The absence of high-volume airflow also means **less air particulate filtration** needed and no hot aisle/cold aisle containment is required – simplifying facility design. However, we must manage the facility water cooling plant capacity and ensure proper heat rejection (cooling towers/dry coolers). Liquid-cooled racks often allow warmer coolant temps (50°C+ return), which can improve cooling PUE.

In summary, the physical infrastructure for the GPU supercluster will adopt **liquid-cooled high-density racks**, either OCP 21″ or standard 19″ form factor depending on vendor. Each rack delivers on the order of **120–150 kW of GPUs** and will weigh ~1.5 tons, so facility prep is critical (power feed, floor strength, cooling distribution). The designs from Dell, Supermicro, Lenovo, etc. are all fairly aligned on using centralized cooling and busbar power to streamline deployment – Dell even offers **turnkey rack integration (IRSS)** so that an entire rack arrives assembled and tested, requiring only external hookups. This can greatly speed up Day-1 deployment (“plug-and-play” AI pods). Our plan is to leverage such reference designs to minimize custom integration on site.

High-Speed Networking: Ethernet (RoCEv2) with NVIDIA Spectrum-X and BlueField-3

To interconnect the cluster nodes (both within and across racks) for distributed training, we will use a cutting-edge **200–800 Gb/s Ethernet fabric** with RDMA capability, rather than InfiniBand. The choice of Ethernet (specifically **RoCE v2** – RDMA over Converged Ethernet) is intentional to align with standard data

center networking, multi-tenant flexibility, and avoidance of vendor lock-in to InfiniBand. However, modern Ethernet can achieve InfiniBand-like performance by using advanced congestion control, adaptive routing, and network offloads – epitomized by NVIDIA's **Spectrum-X** platform. Key aspects of the network design:

- **Network Topology & Bandwidth:** Each GPU server tray in the rack will be equipped with **ConnectX-7 or ConnectX-8** NICs (or integrated into BlueField-3 DPUs) providing at least *2× 200 Gb/s* ports (ConnectX-7) or *2× 400 Gb/s* (ConnectX-8). In fact, the GB300 NVL72 design dedicates an **800 Gb/s "SuperNIC"** module per 4-GPU tray – implemented as dual ConnectX-8 controllers. This equates to ~**11 Gb/s of network per GPU** if 72 GPUs share 800 Gb/s (sufficient for scaling to multi-rack training with minimal communication bottleneck, given NVLink handles intra-rack traffic). These NICs support PCIe Gen5/6 and have advanced offloads for RDMA. We will likely use a **leaf-spine CLOS topology** for scalability: e.g. each rack's nodes connect to redundant 400 GbE top-of-rack switches (Spectrum-4 based), and multiple racks' TOR switches uplink to spine switches at 400 or 800 Gbps. The exact oversubscription will be planned based on expected multi-rack job sizes – we might target a non-blocking fabric for at least up to 2–4 racks together if we want to train jobs across 288 GPUs, for example.

- **NVIDIA Spectrum-X Ethernet:** Spectrum-X is NVIDIA's latest end-to-end solution for AI networking over Ethernet, comprising the **Spectrum-4** switches (which can run up to 51.2 Tb/s and support 64×800G or 128×400G ports), and the **BlueField-3 DPU** SmartNICs working in concert. We will use Spectrum-4 based switches for 200/400G connectivity, and likely enable Spectrum-X software features: **adaptive routing**, congestion control, in-network telemetry, etc. Unlike traditional static ECMP, Spectrum-X can do **packet-by-packet dynamic load balancing** to avoid hot spots, and uses precise congestion signals to steer traffic away from congested paths. Tests have shown this can significantly improve performance: for instance, on the Israel-1 AI supercomputer, enabling Spectrum-X features improved storage throughput by **+20–48% for reads and 9–41% for writes** compared to standard RoCE v2 network settings. The adaptive routing is crucial when dealing with the "elephant flows" of AI (e.g. multi-terabyte model checkpoint writes) to prevent link buffer overflows. Our cluster network will thus leverage these capabilities to maximize effective bandwidth and minimize tail latency.

- **BlueField-3 DPUs:** Each server tray will either have a BlueField-3 DPU or a combination of ConnectX NIC + separate BlueField card. BlueField-3 effectively pairs a ConnectX-7 NIC (dual 200 GbE) with an array of Arm CPU cores and accelerators on one PCIe card. Running NVIDIA's DOCA software, the DPU can offload and isolate network functions from the host. **For multi-tenancy, BlueField is a game-changer**: we can use it to implement tenant-specific network virtualization, firewalls, encryption, and QoS at line rate, without burdening the host CPUs. Specifically, BlueField-3 supports features like: end-to-end packet telemetry, RDMA queue pair isolation, and even hardware-managed **performance isolation** for different traffic classes. One Spectrum-X advantage is the ability to handle congestion and ordering at the network edge – the BlueField DPUs can reassemble out-of-order RDMA packets and perform packet pacing to avoid congestion spreading. This reduces reliance on strict lossless fabric settings (like per-packet pause).

For our design, we will configure **RoCE v2 with ECN (Explicit Congestion Notification)** rather than relying purely on PFC (priority flow control). PFC (lossless Ethernet) can cause head-of-line blocking issues if misconfigured and is tricky in multi-tenant environments. Instead, Spectrum-X and BlueField offer **"adaptive routing + selective retransmit"** approaches that can tolerate an occasional dropped packet

without performance loss, by quickly rerouting and reordering packets. We will still ensure a mostly lossless network for RDMA (likely enabling PFC on an isolated priority for RDMA traffic), but the advanced congestion control (DCQCN algorithms tuned by NVIDIA) will handle incasts gracefully. Tuning guidelines (e.g. properly setting ECN marking thresholds on switches, ensuring end-to-end MTU 9000, and using DSCP for RDMA traffic class) will be followed according to NVIDIA's best practices for RoCE in HPC clusters.

- **Throughput and Latency:** Each GPU-to-GPU communication across racks will incur some latency (on the order of a few microseconds for NIC + switch hops). While NVLink inside a rack is ~1.3 µs latency and >1 TB/s, the Ethernet fabric might have ~5–10 µs latency and 200–400 Gbps per link. For multi-rack training, this means we might not scale *as* efficiently beyond 72 GPUs, but it's mitigated by using strong parallelization strategies (e.g. pipeline parallelism across racks, data parallel within a rack). Also, the **GPUDirect RDMA** capability of ConnectX/BlueField ensures that GPU memory can be directly accessed by RDMA without CPU involvement, keeping latency low and avoiding extra copies. This is essential for multi-node NCCL (NVIDIA's collective communications library) – NCCL will use GPUDirect RDMA over RoCE so that all-reduce operations between racks happen efficiently. We will test that the network can sustain full bandwidth on all-reduce patterns common in training (Spectrum-X's adaptive routing should help achieve near ideal bisection utilization even under all-to-all communication loads).

- **Scaling and Uplinks:** For a moderate-sized cluster (say up to 4 racks of 72 GPUs = 288 GPUs) we might get away with a single-tier leaf-spine. If the cluster scales to e.g. 8 racks or more, we would add another spine layer or use a larger chassis switch. Spectrum-4 switches can be used in a two-tier nonblocking config for a few thousand GPUs if needed (just as InfiniBand usually uses fat-tree topology). The nice aspect of Ethernet is we can also integrate the storage and management traffic onto the same physical network (with VLAN or priority segregation), reducing complexity – more on storage network next.

- **In-Band vs Out-of-Band Management:** Each node will have both a 1/10 GbE out-of-band management port (for BMC/IPMI or Redfish control, on a separate management LAN), *and* possibly use the BlueField DPU for an **in-band management network**. The Supermicro design, for instance, mentions up to 200 Gb/s in-band management via BlueField-3, in addition to the main compute fabric. This could be used for things like a **storage or cluster control plane network** isolated from the GPU RDMA traffic. In our design, we might dedicate one of the DPU's ports to a storage network VLAN, ensuring that heavy storage I/O doesn't interfere with GPU all-reduce traffic (or we rely on QoS if converged). BlueField allows partitioning NIC functions, so tenant VMs/containers could get virtual NICs limited to certain bandwidth.

- **Security & Multi-Tenant Features:** From a security standpoint, running a multi-tenant AI cluster means untrusted workloads may share the physical network. We will leverage **isolation at the DPU/Switch level** – e.g., using VLAN or VXLAN segments per tenant, implemented in the BlueField so that the host OS cannot bypass it. BlueField can enforce L3/L4 firewall rules in hardware, separating tenants' traffic. It also can encrypt traffic: for instance, IPsec or TLS offload on BlueField-3 can provide wire-speed encryption for data-in-transit between nodes. If required (e.g. by SecNumCloud or similar standards for sensitive data), we can configure end-to-end encryption on the overlay network between tenant partitions, with negligible CPU overhead (the DPUs handle the crypto). This might be an extra precaution if multiple orgs (Anthropic, OpenAI, etc.) share a cluster – even if they are logically separated, one might insist on encryption so that even if traffic were sniffed, it's

gibberish. The Spectrum switches support IEEE 802.1AE MACsec as well, but at 400 Gbps scale that becomes complex, so host-level IPsec is more flexible. We will evaluate performance impacts; early data suggests BlueField can do line-rate IPsec with minor latency addition (since crypto is hardware-accelerated).

- **Time Synchronization:** Training on many GPUs doesn't usually need strict time sync, but for logging and distributed tracing we will ensure all nodes have NTP or PTP time sync. If using PTP (Precision Time Protocol), many NICs (including ConnectX) support hardware timestamping, which could be useful if we ever run protocols that require tight sync or measure latency.

Overall, our **network fabric will be a state-of-the-art 400 GbE RDMA network** using NVIDIA's **Spectrum-X Ethernet platform**, tuned for zero packet loss and minimal congestion. It forgoes InfiniBand but retains high performance: we get RDMA, collective offloads (NCCL/SHARP-like all-reduce can be done via software since BlueField can act as a rendezvous), and adaptive routing. This approach also better supports multi-tenant integration with standard DC networking gear, and allows mixing of Ethernet-attached storage, firewalls, etc.

Before moving on, it's worth noting real-world usage: NVIDIA's own DGX SuperPODs historically used Mellanox InfiniBand, but with Spectrum-4, even **large AI supercomputers like Israel-1 are built on Ethernet** and achieving excellent performance. This validates our approach. We will work closely with NVIDIA engineers (perhaps through their Network Design Advisory service) to implement best practices from day one, avoiding pitfalls like untreated congestion that could cripple multi-tenant workload performance.

Scalable Storage Architecture: Tiered High-Performance Storage for AI

AI superclusters not only need massive compute – they must feed those GPUs with massive amounts of data at high throughput. For multi-tenant environments, the storage subsystem must be **flexible (support many datasets, users)** and **securely isolated**, while offering extreme performance for training (hundreds of GB/s of throughput, millions of IOPS) and scalable capacity (petabytes, possibly into exabytes for certain research). We propose a **tiered storage solution** combining the strengths of specialized parallel file systems and object storage:

Tier-1: Ultra-High-Performance Parallel Storage (Flash)

The top tier will consist of an **all-flash distributed file system** optimized for AI/workload throughput. The likely candidates are **WekaFS (WekaIO)**, **VAST Data**, **DDN's AI appliances**, or **Pure Storage FlashBlade**, as mentioned – all of which are proven in AI/HPC environments:

- **WekaFS:** Weka's software-defined parallel file system pools NVMe SSDs across a cluster of storage nodes, presenting a single POSIX namespace with very high throughput and low latency (small file and metadata performance are also excellent, which is useful for AI datasets with millions of files). Weka is already used in many GPU clusters and is **certified for NVIDIA GPUDirect Storage (GDS)**, meaning GPUs can directly DMA data from Weka client buffers via RDMA, bypassing CPU copying. Weka supports **tiering to object storage** natively: it can automatically move colder data to an S3-

compatible object store while keeping hot data on flash. This is ideal for multi-tenant use – we can allocate each tenant a Weka namespace or directory, with quotas, and their inactive data can flow to a cheaper tier transparently. Weka also offers encryption at rest and multi-tenant auth (it can integrate with LDAP/AD for user separation). Performance-wise, a modest Weka cluster (for example 10 nodes with NVMe) can deliver 10s of GB/s; large installations (Facebook/Meta reportedly used Weka for some AI training clusters) can hit **>300 GB/s** aggregated throughput. We will need to size it based on how many GPUs and how big the data shards per GPU; as a rule of thumb, feeding 576 GPUs (8 racks) might require on the order of 200–400 GB/s to keep them saturated (assuming ~0.5–1 GB/s per GPU for training throughput). This is achievable with ~a couple dozen NVMe-based storage servers. Weka's strong point is also latency – for interactive workloads or random reads (like inference), it's very fast due to distributed metadata and no single controller bottleneck.

- **VAST Data:** VAST's architecture pairs NVMe **QLC flash for capacity + Optane persistent memory for caching** (though newer versions might use DDR or other NVRAM since Optane is discontinued). It provides a global namespace accessible via NFS, SMB or S3, and uses aggressive data reduction (compression, dedup) to lower cost. VAST is known for good performance on read-heavy or mixed workloads and ultra-large capacity on flash (their clusters can scale to tens of petabytes of effective storage). VAST does not tier to slower storage itself (it advocates **all-flash** for everything with data reduction), so in our design we might use VAST for certain workloads that demand consistent low latency. Like Weka, VAST supports **RDMA over Ethernet** clients for performance. A potential edge for VAST is simplicity – one big system for both NAS and Object interface. However, multi-tenancy on VAST would rely on creating separate storage pools or share permissions.

- **DDN (DataDirect Networks):** DDN is an established HPC storage vendor, offering solutions like **EXAScaler** (Lustre-based parallel file system) and **AI400X** appliances. DDN's appliances were used in NVIDIA's DGX SuperPOD reference designs historically, and they are tuned for GPU workloads (with GDS support). A single DDN AI400X array, for example, can provide ~48 GB/s read throughput and scale linearly with more appliances. Lustre or DDN's new solutions can be a bit complex to manage, but they offer robust **fine-grained control** and have options for encryption and snapshots. In multi-tenant settings, Lustre is less commonly used (since it's more a single-project high-throughput scratch space), but DDN also offers NFS/SMB interfaces if needed. Alternatively, DDN's newer product (if any for Blackwell generation) could integrate easier – we should check if DDN has a "reference architecture for AI" whitepaper; DDN is indeed partnering with NVIDIA Spectrum-X as well, meaning their systems will integrate with our Ethernet network with ease.

- **Pure Storage (FlashBlade):** Pure's FlashBlade//S is a scale-out NAS + Object platform built on high-performance NVMe flash and an ultra-parallel architecture. It's known for very easy management and fast metadata operations. FlashBlade systems have been popular in AI for hosting miscellaneous files and as an NFS store, though for pure training data, Weka or Lustre often outperform it at scale. However, Pure recently announced FlashBlade can deliver up to ~>100 GB/s in a multi-chassis configuration and they position it for AI/ML pipelines (notably, Pure was the first Ethernet-based storage certified for NVIDIA SuperPODs). For multi-tenancy, FlashBlade is attractive because it provides **multi-protocol** support (a user can access data via NFS, SMB, or S3 with the same namespace, which can be useful for different tenant workflows). It also has built-in encryption at rest and transparent compression.

Given the above, **our design can incorporate one or more of these Tier-1 storage solutions**. In fact, a common approach is to use **multiple storage tiers even at the top level**: e.g., Weka or Lustre for the actual training scratch (where maximum I/O throughput is needed), and something like a FlashBlade or VAST system as a "landing zone" or secondary storage for datasets that don't require the last ounce of performance. Tenants could stage their data from an object store into WekaFS just for the training run, then data could be evicted back out after training. This keeps the expensive flash tier relatively small and focused.

Crucially, all these storage systems integrate with **NVIDIA GPUDirect Storage (GDS)** and our RoCE network. This means GPUs can perform DMA reads/writes directly to storage buffers via RDMA, reducing latency and CPU load. For example, using Spectrum-X network, tests on a supercomputer showed that the storage fabric (with DDN, VAST, Weka as participants) improved significantly under Spectrum's adaptive routing. We will ensure to enable those features – e.g., run our storage on the same lossless network with perhaps a dedicated VLAN, and let Spectrum-X handle adaptive routing so that one tenant's heavy reads don't interfere with another's writes.

Security for Tier-1 storage: We will enable **encryption at rest** on these systems (all listed vendors support it, usually with self-encrypting drives or software encryption). We'll also consider using separate filesystems or at least separate directory trees per tenant with strong ACLs. Some systems (Weka, Lustre) can also support multi-tenant by separating at the client level (for instance, if we mount a Weka filesystem on different Kubernetes namespaces with only certain pods having keys to certain directories).

Tier-2: Capacity and Cost-Optimized Storage (Object Store / Ceph)

Below the ultra-fast tier, we plan a large **capacity tier** for data that is not currently in use but must be stored, or for workloads that are not IO-bound. This could also serve as a "cold" layer for backups, log storage, or older checkpoints. For this tier, cost per TB is a bigger concern than raw performance, so we likely use high-capacity HDDs or mixed HDD/SSD, and open-source solutions to avoid heavy licensing costs. The prime candidate here is **Ceph**, deployed with a focus on multi-tier performance:

- **Ceph (Open-source)**: Ceph is a distributed storage platform that can provide object storage (via the S3-compatible RGW interface), block storage (RBD), and even a filesystem (CephFS). We prefer a Ceph distribution not tied to Red Hat (since we avoid RH due to subscription costs), so likely **Ceph on Ubuntu/Canonical** or a community version. Canonical provides **Charmed Ceph** or Ceph pre-integrated in their OpenStack and Kubernetes offerings, which might align well. We could use Ceph's object store as the main interface for Tier-2 (tenants get S3 buckets for their raw data, archives, etc.), and possibly CephFS if a POSIX interface is needed for certain applications (though CephFS performance is moderate).

Ceph can be configured with multiple performance tiers: e.g., an **all-HDD pool** for archival data and an **SSD pool** for more active data or metadata. It also has a concept of **cache tiering** (although in practice that's been tricky, many opt for manual tiering). Perhaps a simpler approach: use Ceph mainly for object storage (S3) – this pairs well with Weka, because Weka can tier to any S3 target. We could have Weka automatically offload cold data to a **Ceph S3 cluster**. This way, when a tenant's dataset is not actively used, it lives on the cheaper Ceph storage (which might be large HDD-based), and when needed, Weka seamlessly pulls it back to NVMe tier.

We will ensure Ceph is tuned for large objects (AI data tends to be big files or batches), and replicate data adequately (SecNumCloud likely requires at least 2–3 copies or erasure coding across data centers for durability). We might use erasure coding on Ceph to save space for the really cold stuff (with a small performance penalty). The Ceph cluster can also be scaled independently of the training cluster – it's more about capacity scaling (we could have e.g. 10 PB of Ceph storage backing a 1 PB Weka active tier).

- **Other Options:** Alternatives in this tier include **Apache Ozone** (an HDFS derivative for object storage), **MinIO** (lightweight S3 server – could be great for per-tenant isolated object stores, but at large scale Ceph is more battle-tested), or even cloud storage (if not fully sovereign). But given sovereignty and economics, a self-hosted Ceph seems best. Another possibility: use a **tape archive** or **glacier** for deep cold storage if some tenants have regulatory retention needs – but that's probably outside scope; we mention it for completeness.

Performance vs Cost: By combining these tiers, we aim to achieve a balance: The expensive flash tier (Weka/VAST/Pure) is relatively small (maybe 5–20% of total data volume) but delivers huge performance for active training. The Ceph tier can hold the bulk of data cheaply (using high-density 20+ TB HDDs, potentially in a dense storage enclosure like 4U90 disk systems). Tenants could also directly use the Ceph/S3 tier for workloads that are less IO intensive (e.g. analyzing results, or training smaller models). Ceph's performance per TB is lower (throughput per node might be in the 1–5 GB/s range depending on how many OSDs, etc.), but it's fine for many use cases. Ceph also allows multi-site replication which could be useful for DR or for a multi-data-center cluster spanning France and Germany as hinted.

Networking for Storage: The storage systems will connect via the same Ethernet network (Spectrum-X). Ideally, we isolate storage traffic on its own VLAN or priority. Spectrum-X's adaptive routing and congestion control will treat the storage flows separately to avoid interference with GPU all-reduce flows. The technical blog confirms integrating storage on Spectrum-X yields faster job completion – effectively, by preventing congestion, we ensure that, say, a big checkpoint write doesn't clog the network and slow down GPU-to-GPU communications.

Multi-Tenancy & Security in Storage: Each tenant's data must be isolated from others. On the object store (Ceph RGW), we'll create separate buckets/tenants with unique credentials. Ceph supports bucket policies and user quotas, which we will use to prevent one tenant from accidentally reading another's data. For the file system tier, we can either stand up separate filesystems per tenant (e.g., a separate Weka filesystem per tenant organization) – WekaFS can carve out separate file systems with independent quotas and even encryption keys. Or simpler, use a single filesystem but separate directories with ACLs and perhaps encryption. If extremely high security is needed, we could even run separate storage clusters per tenant, but that loses efficiency; instead, enabling encryption at rest (so if one tenant's data somehow was accessed by another, it's gibberish without the key) and strict permission controls should suffice.

We also consider **data ingress/egress**: Multi-tenant clusters often need to ingest large datasets from external sources. We should provide secure methods for tenants to upload data to the storage (e.g. an S3 upload endpoint to the Ceph cluster, or a high-speed data transfer node). Possibly integrate with cloud storage (if tenant has data in AWS S3, use a fast network link or Snowball device to transfer). The design should include a plan for moving 100s of TBs into the cluster efficiently.

Finally, backup: The cluster's valuable trained model checkpoints should be backed up (maybe to the Ceph tier or even offsite). We might allocate part of Ceph to hold periodic snapshots from Weka or copies of final

models. If required by compliance, we could even encrypt each tenant's data with tenant-specific keys (so that even admins of the cluster can't see raw data). This could be done via application-level encryption or using storage that supports multi-tenant encryption keys. Some storage solutions allow "bring your own key" for encryption.

Software Stack and Cluster Management

With the hardware in place, a critical aspect is the software layer that orchestrates resources and provides a usable environment for multiple teams. We will use an **Ubuntu-based OS stack** with containerization and an AI cluster management toolkit, avoiding proprietary OS lock-in where possible. Key components:

- **Base Operating System:** We opt for **Ubuntu 22.04 LTS (or 24.04 LTS if available)** as the host OS on all Grace CPU nodes. NVIDIA's DGX OS is essentially an Ubuntu LTS with NVIDIA drivers and CUDA stack pre-installed, so using Ubuntu aligns well with NVIDIA's supported environment without the need for Red Hat. (Red Hat's licensing and support model is less appealing, and many AI frameworks are developed/tested on Ubuntu/Debian environments). Grace CPUs run Linux just like x86; NVIDIA has been using Ubuntu 20.04 on DGX systems and likely will support Ubuntu 22.04+ on Grace systems. We will harden the OS (per SecNumCloud guidelines) – enabling secure boot, minimal package installation, regular kernel updates, etc. If certain components (like GPFS or professional schedulers) required RHEL, we would consider Rocky or AlmaLinux, but at this time Ubuntu should cover all needs. **Real-time kernel** is not necessary for training, but we might tune some kernel parameters: for example, increase max IPC limits (for many parallel processes), enable CPU isolation for housekeeping threads (Grace has many cores, we can dedicate some to OS background tasks and others to ML processes), and ensure BIOS/firmware settings are optimized (C-states, NUMA, etc.).

- **GPU Drivers and Libraries:** We will install the latest **NVIDIA HPC SDK / CUDA toolkit** and drivers that support Blackwell GPUs. This includes CUDA 12 or newer, cuDNN, NCCL (which supports NVLink and RoCE). Since our cluster uses MIG? (We might allow MIG – Multi-Instance GPU – for inference jobs, but for training large models MIG will likely be disabled to use full GPUs). The drivers will be configured accordingly. We'll also deploy **NVIDIA Fabric Manager** (for NVSwitch management) on each node as needed, and DCGM (Datacenter GPU Manager) for monitoring GPU health and telemetry.

- **NVIDIA AI Enterprise (NVAIE):** This is NVIDIA's licensed software suite for enterprise AI, which provides a curated set of frameworks (TensorFlow, PyTorch containers, RAPIDS, etc.), tools like NVIDIA Base Command (for job scheduling/management), and support. We will strongly consider using NVAIE to accelerate software provisioning. Notably, NVAIE provides **NVIDIA Base Command Manager**, which can manage multi-node training jobs, handle user authentication, and resource scheduling across the GPUs [2]. It basically is the software that powers NVIDIA's own DGX SuperPODs, packaged for others. However, NVAIE is a licensed product charged per GPU. Based on public info, an **NVAIE subscription costs on the order of \$1,800–\$3,600 per GPU per year** (exact pricing varies by term and support level). For example, Dell quotes \$18k for 5-year, 1-GPU support (which is \$3.6k/yr), and 3-year licenses were around \$7.2k/yr in some listings. At our scale (potentially hundreds of GPUs), this is a significant cost, but it does come with enterprise support and validated software. We will evaluate the value: if we need guaranteed support and the convenience of ready-to-run AI frameworks, we might invest in NVAIE for at least the critical nodes.

Alternatively, since we have strong in-house expertise, we could use the open NVIDIA NGC containers and our own orchestration without paying per-GPU software fees.

- **Cluster Orchestration & Scheduling:** To manage multi-tenant workloads, we have a few choices:

- **Kubernetes-based approach:** Use Kubernetes as the underlying scheduler, with NVIDIA's device plugin for GPUs and possibly a custom operator for multi-node training (MPI Operator or Kubeflow for ML). SpectroCloud's **Palette** is explicitly of interest – *Palette AI* is a solution that can deploy and manage K8s clusters optimized for AI (with built-in support for NVIDIA drivers, GPU isolation, and even multi-cloud bursting). Palette would allow us to treat each tenant as a namespace or separate virtual cluster, enforce quotas, and schedule jobs in containers. It supports Ubuntu nodes and can integrate with Slurm or other job backends if needed. If we choose this path, each training job might be a Kubernetes Batch Job or Argo Workflow, requesting a certain number of GPUs. Advanced features like GPU-sharing (MIG) can be enabled via the device plugin if needed for small inference tasks. One benefit is that Kubernetes + Palette can also manage the **lifecycle** (Day-2 ops) of the cluster – upgrades, monitoring, etc., which is valuable for a lean ops team.

- **Traditional HPC Scheduler (Slurm):** Slurm is widely used in multi-tenant HPC centers to schedule GPU jobs. It's proven and can handle advanced scheduling policies (fair share, partitions per tenant, exclusive vs. shared node allocations). We could run Slurm across the cluster nodes, possibly with adaptations for Grace (should be fine) and use plugins for NVIDIA GPUs (Slurm has native support for GRES – generic resources like GPUs). Slurm would allow us to create **accounts/partitions** for each tenant, enforcing quotas on GPU hours, etc. It can also schedule multi-node jobs easily. Slurm's downside is it's more static and not as user-friendly for ML engineers used to cloud. But we can build user-friendly submission wrappers or use a GUI like Open OnDemand on top. Slurm is free (open-source) which is nice. It doesn't directly handle containerization, but one can use it in conjunction with Singularity/Apptainer or simply have users run Docker inside their job allocation.

- **Hybrid:** We might even run Kubernetes on top of Slurm or vice versa – e.g., use Slurm as a low-level scheduler but expose a Kubernetes interface to users (there are tools to integrate the two, or one can dedicate some nodes to a K8s cluster for certain interactive services and others for batch jobs).

Given the multi-tenant requirement and desire for cloud-like flexibility, a **Kubernetes-centric solution with SpectroCloud Palette** seems promising. Palette can manage bare-metal Kubernetes across many nodes and includes GPU support. It also can manage **Air-gapped or sovereign deployments** (which is relevant for SecNumCloud – ensuring no telemetry leaks to outside). With K8s, each tenant could be given their own namespace and perhaps even their own **Virtual/Kube cluster** (Palette supports hierarchical multi-tenancy). We can implement network policies so tenants' pods cannot talk to each other except through controlled channels.

Additionally, **NVIDIA offers Base Command Platform** (BCP) which is a higher-level job scheduler with a nice GUI for ML workflows – it typically runs on top of Kubernetes. For example, NVIDIA Base Command (part of NVAIE) could run in our cluster and provide a SaaS-like experience: users log in to a web portal, upload datasets (or connect to our storage), and launch training jobs that run on the cluster's GPUs, with the platform handling scheduling and reporting. This is similar to what CoreWeave (mentioned in Dell's

news) provides as a service. Since we may want to allow external customers to run workloads, having such a user-friendly interface might be valuable. The cost is the licensing and some onboarding effort.

- **Resource Partitioning:** Multi-tenancy means we might sometimes partition the cluster – e.g., reserve certain racks or nodes for a specific tenant during a contract. We can implement that either with **Slurm partitions or K8s taints/labels** to dedicate resources. However, to maximize utilization, we'd ideally keep it flexible (one tenant can use any free GPU when another tenant isn't using them). We will rely on the scheduler's fairness mechanisms for that. Also, **job preemption** or QoS levels can be set – e.g., internal research jobs might run at lower priority and get preempted if a paying tenant job comes in.

- **Data Management Software:** We will likely deploy supporting software for data management: for instance, **NVIDIA Morpheus** (for cybersecurity logs AI) is not relevant here, but **NeMo** (for building LLMs) and other frameworks from NVIDIA can be containerized. We may also consider **Horovod** or native PyTorch DDP for distributed training – those come as part of the frameworks. Monitoring-wise, we'll deploy **Prometheus/Grafana** stack to collect metrics from GPUs (DCGM exporter), DPUs, network (switch telemetry via SNMP or streaming), storage (Ceph and Weka have their own metrics). This is important for both performance tuning and demonstrating compliance (SecNumCloud might require certain monitoring and incident response capabilities).

- **Automation and Deployment:** The initial provisioning of this complex system will itself require automation. We will use tools like **Ansible, Terraform** and perhaps vendor-provided automation (Dell and others have deployment services). For example, Dell's IRSS includes factory integration and on-site deployment support – we should take advantage of that for setting up racks. For software, SpectroCloud Palette if used will automate installing Kubernetes and desired software on each node. If we roll our own, we might use **Canonical MAAS + Juju** to bare-metal provision Ubuntu and then charmed operators for services (like a charmed Ceph cluster, charmed K8s). This ties into our desire not to be locked to one vendor's proprietary automation – using open tools will give us long-term flexibility.

- **Confidential Compute:** The question of confidential computing arises – if tenants are very sensitive (e.g., one is a government or medical organization), they might want assurances that even if they share hardware, their data is never exposed in plain form. There are a few angles:

- **MIG (Multi-Instance GPU):** This allows partitioning an NVIDIA GPU into smaller slices (each with its own memory isolation). MIG can be used to securely run different clients on one GPU, with hardware isolation. We could use MIG on some GPUs for inference services (e.g., one tenant runs a GPT inference on 20% of a GPU while another uses the rest). MIG isolation is quite strong (each MIG has separate context and QoS enforcement). However, MIG is mostly for A100/H100 – Blackwell's MIG details are not yet fully known but expected to continue (the spec says B200 has 7 MIG partitions possible). We can enable MIG via software if needed. For training though, MIG is usually off (need full GPUs).

- **Confidential VMs:** If we provide tenants virtual machines (or containers) on shared hosts, technologies like **AMD SEV** or **Intel TDX** encrypt VM memory such that even the host admin or a rogue peer VM can't read it. Our cluster is Grace CPU based, which currently doesn't have a known encrypted memory feature (AMD EPYC does). If confidential compute becomes a must for

customers, one approach is to include some **AMD EPYC CPU nodes with MI300 GPUs** for those specific tenants, since AMD is pushing confidential GPU computing (they announced MI300 supports encrypted compute when paired with EPYC). Alternatively, NVIDIA has been working on **Confidential Computing for GPUs** too – H100 introduced support for confidential containers where the GPU memory is isolated and encrypted keys are needed to access it. It's possible Blackwell/Grace platform will enhance this. We'll keep an eye on NVIDIA's **Confidential Computing** roadmap. In practice, enabling full confidential computing can complicate software (you need attestation, etc.), so we might offer it as an optional mode for extremely sensitive workloads, rather than default.

- **Tenant Separation:** On the network and storage we covered isolation. On the compute side, the primary isolation is via the Linux kernel (namespaces, cgroups) or via hypervisors if we go that route (we likely won't use VMs for training, as it adds overhead, except possibly lightweight hypervisor like Kata containers or gVisor if needed for an extra security boundary). We will ensure that no two tenants share the same OS instance unless confident in container isolation. Using K8s, each tenant's workload can run in separate Linux user IDs and with AppArmor/SELinux profiles to mitigate any container escape risk.

- **Compliance Considerations (SecNumCloud, etc.):** SecNumCloud (by ANSSI, France) demands strict security for cloud services: data encryption at rest and in transit, admin personnel background checks, intrusion detection, etc. Our cluster will implement:

- Full disk encryption on OS drives (Grace nodes may netboot, but local NVMe scratch should be LUKS encrypted).
- Encryption at rest on storage as mentioned.
- In transit encryption on external connections (e.g., when a user accesses the cluster from outside or data is imported/exported). Inside the cluster, we might not encrypt all internode traffic for performance, but for tenant-to-tenant isolation we will treat the internal network as untrusted beyond each tenant context.
- Strong identity management: integrate with an Identity Provider for user auth, use MFA for console access, separate admin roles for hardware vs tenant ops.
- Logging and monitoring of all admin actions, perhaps using something like a bastion with session recording for any maintenance.

- Possibly an on-premises Hardware Security Module (HSM) or key management service to manage encryption keys (so that keys are not stored on the same machine as data). We could use a French certified HSM (per SecNumCloud recommendations) for master keys that encrypt storage keys, etc.

- **Lifecycle and Updates:** The cluster software stack should be regularly updated (esp. CUDA libraries, framework versions). Using containerized workloads helps – users can bring their container with a specific version of PyTorch, etc. We will maintain a set of NVIDIA NGC containers locally (maybe in a registry) to ensure no internet pull is needed (for sovereignty). For the host, we will schedule rolling upgrades (draining one node at a time) to apply OS security patches. Since Grace is ARM, we must ensure all our software (K8s, monitoring agents, etc.) supports ARM – which nowadays most do, thanks to Graviton adoption.

- **Vendor Automation Stacks:** The user request asked whether vendors offer "full automation stacks" for rapid deployment. NVIDIA, through Base Command and their partnership with VMware, offers

things like **NVIDIA AI Enterprise on VMware** (we likely avoid since we don't want the overhead of VMware hypervisor for bare-metal training). Dell might have some automation via their OpenManage suite or DevOps tools for IR7000 (Dell could pre-rack and maybe even pre-install a software image). However, since we are aiming for sovereignty and flexibility, we will likely do our own automation or use an open platform like SpectroCloud. Another note: some cloud providers use **OpenStack** to manage multi-tenant GPU clusters (creating virtual GPUs to tenants, etc.). We could consider an OpenStack layer to give tenants self-service VMs with GPUs. But given the complexity and the fact that ML engineers prefer container or bare-metal, we lean against adding OpenStack's complexity. Kubernetes with proper plugins can achieve similar "cloud-like" self-service (especially with web portals atop it).

To ensure **"day-1 deployment, scaling and lifecycle management"** is smooth, we will follow reference architectures – e.g., Dell's **AI Factory** solutions come with services, and NVIDIA's **Mission Control** software is aimed to monitor and manage AI datacenters. NVIDIA Mission Control (just announced for Blackwell) will provide a software suite for operating Grace-Blackwell clusters, handling tasks like workload scheduling, telemetry, and "full-stack intelligence for infrastructure resilience". This sounds like a complement to our chosen scheduler, possibly giving us AI ops insights (maybe an AI-driven monitoring that predicts failures or performance issues – NVIDIA hints at their expertise delivered as software). We will evaluate Mission Control once more details are available, but it could be a valuable tool to run the cluster at peak efficiency and to assist our ops team.

Security and Multi-Tenancy Considerations

(We have interwoven many security aspects above, but we'll summarize and add any remaining points here.)

Multi-tenancy is a core design point: the cluster must robustly segregate tenants' workloads and data while still allowing shared use of the massive compute resources. The main security domains are **compute isolation**, **network segmentation**, and **storage/data protection**:

- **Compute Isolation:** We will not run two different tenant training jobs in the same OS process or container – each job gets its own container or VM. Linux containers (with user namespace, cgroups, seccomp, AppArmor) are the baseline isolation; for stronger isolation, we could assign separate physical nodes to different tenants for critical jobs (affinity rules in scheduler). Also, when possible, a tenant gets exclusive use of a GPU (no time-slicing between tenants on one GPU) unless using MIG slices. This prevents side-channel or timing attacks via shared GPU. As mentioned, MIG can provide hardware isolation if we intentionally want to share one GPU for smaller tasks (with MIG, each partition's SMs and memory are separate and performance is partitioned, so one tenant's partition can't access another's data).

- **Network Segmentation:** The network design will enforce that tenants cannot sniff or interfere with each other's traffic. On Ethernet, without precautions, a user with root in a container could potentially craft packets. But because we have BlueField DPUs, we can place each tenant's container interface on a distinct **virtual network** (VXLAN or even a separate VLAN or VRF). BlueField can act as a vSwitch or SmartNIC, ensuring one tenant's packets can only go to allowed destinations. We will likely implement an overlay network (like Kubernetes' Calico or Multus with SR-IOV DPUs) where each tenant's pods get an IP in a tenant-specific subnet, and the DPU enforces L3 separation. For

additional security, enabling **end-to-end encryption** (IPsec VPN per tenant) would mean even if traffic crosses shared wires/switches, it's encrypted with tenant-specific keys. This might be overkill internally, but it's an option if mandated (the BlueField can offload IPsec so it's viable).

Another aspect is **DoS protection** – one tenant should not flood the interconnect and starve others. We will use network QoS on the switches and DPUs: e.g., allocate separate queue/TC for each tenant or at least for each traffic type. Spectrum-4 switches have advanced QoS capabilities (eight traffic classes, etc.), and BlueField can tag traffic. The goal is to prevent noisy-neighbor issues on the network.

- **Storage Protection:** Each tenant's data is logically separated as discussed (separate volumes, buckets, access credentials). Even on the high-speed file system, careful permission and perhaps encryption of certain sensitive datasets will be applied. Weka, for example, can encrypt data with a cluster key, but ideally we'd have tenant-specific keys. If that's not directly supported, we might rely on application-level encryption for extremely sensitive data (the tenant can encrypt their data before storing it). On Ceph S3, we can enable S3 bucket encryption with unique keys per bucket via a KMS. This means even Ceph admins cannot read the data without keys.

Backups, if taken, will also be encrypted and stored in tenant-specific spaces.

- **Admin Access:** Multi-tenant also means we must partition responsibilities between *infrastructure admins* and *tenant users*. Tenants should not have root on the host or any access to others' jobs. We will provide them a constrained interface (like submitting jobs via an API/portal or accessing a dedicated login node that is containerized per tenant). Administrative access to the physical nodes will be limited to our cluster operators, who will follow strict security practices. If needed, we can utilize **secure enclaves** for admins too – e.g., manage the cluster from a secure jump host and never directly from the open internet.

- **Compliance (SecNumCloud):** To specifically address SecNumCloud (which is aimed at cloud providers handling sensitive French govt data), we anticipate requirements such as: data must reside in France, operated by personnel with certain clearance; encryption keys under customer control; continuous monitoring for intrusions; strong identity management; and the cloud management systems themselves hardened. We will ensure all data is hosted in-country (and similarly, for Germany, if we host there, compliance with BSI C5 or similar frameworks). The cluster's management plane (K8s control, etc.) will be on isolated networks and possibly have two-factor auth for any access. We can implement an **IDS/IPS** (Intrusion Detection/Prevention System) on the management network, maybe leveraging the DPU's ability to sniff traffic or using a virtual appliance.

We should document every measure, as typically we'd need SecNumCloud certification via audit.

- **User Separation & Fairness:** Multi-tenancy is not just security – it's also about fairness and quality of service. We'll implement scheduling policies so that one tenant can't monopolize all GPUs unless others are idle. Likewise, if one tenant job crashes a node (e.g., due to a hardware fault or kernel panic), our system should contain the issue (with job isolation, only that job fails, not others). We'll use health checks on GPUs – if a GPU encounters ECC errors or NVLink errors, we can mark it offline and not let it degrade others' tasks.

- **Real-World Observations:** In existing multi-tenant GPU clouds (like AWS, Azure GPU instances, or CoreWeave's cloud), the providers use strategies we are adopting: e.g., AWS uses Nitro (DPUs) for network and storage isolation; Azure uses SR-IOV NIC partitioning; both rely on hypervisors (which we are trying to avoid for performance). Our approach with BlueField and bare-metal containers is cutting-edge but feasible, essentially similar to how some HPC centers share resources securely via container tech.

Finally, we also consider **auditability** – we might enable logs for every job (which user ran what, how much resources consumed, any security events). In case of a breach or misuse, we need forensic data. Tools like **NVIDIA's NeMo Auditing** are not known, but for our purposes standard Linux audit and network logging should do.

Vendor Reference Designs and Hardware Options

To concretely plan the cluster, it's useful to list the available **vendor systems** for each component (compute, network, storage) and their specs. Below is a comprehensive list of relevant hardware along with key specifications and references:

GPU Compute & Integrated Racks:

Vendor / System	Description and Specs	Reference / Data
NVIDIA GB200 NVL72 (Grace+Blackwell)	Official NVIDIA rack-scale design with 72 Blackwell **B200** GPUs (192 GB HBM3e each) and 36 Grace CPUs, fully liquid-cooled. NVLink-5 connects all 72 GPUs with **>1 PB/s** bandwidth and unified 240 TB fast memory across the system. Acts as one giant GPU for training/inference. Power ~120 kW per rack (air-cooled equivalent would be far less efficient – NVL72 is *25× more energy-efficient* than an H100 air-cooled setup).	NVIDIA product page
NVIDIA GB300 NVL72 (Grace+Blackwell Ultra)	Next-gen NVIDIA design unifying 72 **Blackwell Ultra B300** GPUs (288 GB each) + 36 Grace. Targeted for large-scale *inference* and "AI reasoning" with 50× the output of Hopper systems. NVLink-5 bandwidth ~130 TB/s in one rack [3]. Fast memory ~40 TB total. *Enhanced performance:* 1.5× more HBM per GPU and 2× Transformer Engine speed vs B200. Power ~150 kW/rack (estimated +25% over GB200). Requires liquid cooling.	NVIDIA product page [3]

Vendor / System	Description and Specs	Reference / Data
Dell PowerEdge XE9712 (with IR7000 Rack)	Dell's implementation of NVIDIA NVL72 in a factory-integrated rack. **72 GPUs + 36 Grace** in a 21″ wide **IR7000** rack. IR7000 provides up to **480 kW cooling** capacity via direct liquid and supports multi-generation upgrades. Example configs: *144 GPUs (2× NVL72 domains) per rack with NVLink or 72 GPUs with full NVLink*. One XE9712 rack delivers >1 ExaFLOP AI performance and ~40 TB memory [1]. Power usage ~132 kW (Dell uses 8×PSU shelf similar to Supermicro) for 72-GPU config. Dell provides **"IRSS" turnkey deployment** service.	Dell Press Release & Blog [1]
Dell Integrated Rack 5000 (IR5000)	Dell's 19″ rack solution for AI. Standard EIA 19″ rack supporting heterogeneous servers (Intel, AMD, NVIDIA). Up to **72 GPUs per rack (air-cooled)** or **96 GPUs (liquid-cooled)**. Slightly lower density than IR7000 but easier to fit in existing datacenters. Can mix GPU and CPU-only nodes. IR5000 can be air or liquid – useful if we deploy some air-cooled systems (e.g., lower-power inference nodes). Fully validated with Dell's networking and cooling options.	Dell IRSS FAQ
Supermicro SRS-GB200-NVL72	Supermicro's full-rack **GB200 NVL72** solution. 19″ 48U rack, **18× 1U** GPU servers each with 4× B200 GPUs + 2× Grace (total 72 GPUs). Integrated with **9× NVSwitch** chips to link GPUs at 1.8 TB/s each. Contains **8× 33 kW PSU shelves** (132 kW usable) and a **250 kW liquid CDU** in-rack. Networking configurable with NVIDIA Quantum-2 InfiniBand *or* Spectrum-X Ethernet (ConnectX-7 or BlueField-3) up to 400 Gb/s. Rack dimensions: 600 mm W × 1068 mm D × 2236 mm H (about 2.2 m), weight ~1360 kg. An **air-cooled variant (NVL4)** is possible where GPUs are in groups of 4 with external fabric (less NVLink), but our focus is the NVL72. Supermicro provides end-to-end services (consultation to onsite install).	Supermicro Specs

Vendor / System	Description and Specs	Reference / Data
HPE "AI/ML Supercomputer" (Cray/ HPE)	HPE offers integrated systems via its Cray line. The HPE solution for Blackwell is marketed as **"HPE with NVIDIA GB300 NVL72"**, featuring a 72-GPU rack similar to others. HPE emphasizes a mix of air+liquid cooling, with each rack ~132 kW (and possibly ~150 kW for GB300). One HPE offering uses a **Cray EX chassis** with Grace+Blackwell blades, but more concrete is an upcoming HPE ProLiant or Apollo system for Blackwell. HPE's site mentions their design has a **72-GPU NVLink domain in a fully integrated rack**, delivering ~**1 PFLOP of FP64** or higher of AI performance. HPE also integrates with their storage solutions (e.g., Cray ClusterStor or partner storage). While details are sparse publicly, it's comparable to Dell/Supermicro. HPE's strength is in large supercomputers (they built e.g. Leonardi (France) with GPUs, etc.). We will keep contact with HPE for reference designs once available.	HPE Brief (launch video)
Lenovo ThinkSystem SC777 + N1380 Neptune	Lenovo's liquid-cooled Blackwell platform. It consists of the **N1380 Neptune chassis** (13U high) and **SC777 V4 server trays**. Each SC777 tray holds **2× Grace + 4× Blackwell (GB200)** in a vertical form factor. Eight trays in one 13U chassis = 32 GPUs per chassis. Up to **3 chassis per 42U rack** gives **96 GPUs/rack** with full liquid cooling. Total power **≈162 kW per rack** (54 kW ×3) with 96 GPUs. This design uses standard 19″ racks (Neptune chassis fits in 19″). Cooling: 100% heat removal via water; 4× 15 kW PCS (power converters) per chassis for 54 kW DC output (N+1). Lenovo highlights standard floor tile footprint (two chassis per 60×60 cm tile) and high efficiency (no fans). Networking: supports NVIDIA Quantum-2 (IB) or Spectrum-X Ethernet, and NVIDIA AI Enterprise stack. Lenovo also has smaller 8-GPU air-cooled servers (SR675 V3 etc.) which could complement for less intense needs.	Lenovo Press & DCD

Vendor / System	Description and Specs	Reference / Data
ASUS & Others (Inspur, QCT)	ASUS has previewed an **"AI Pod" with NVIDIA GB300 NVL72**, likely similar to others (72 GPUs, 36 Grace, liquid-cooled) boasting 1.5× more AI perf vs previous gen. Inspur and QCT, major server OEMs, will undoubtedly offer HGX-based 8-GPU servers and possibly integrated racks (Inspur had solutions for HGX A100; for Blackwell they likely will too). These could be considered especially for cost competition. For example, if Inspur offers a lower-cost 8-GPU server, we might use those for certain parts of the cluster (keeping in mind any supply chain or security concerns for gov clients). Since our focus is on best-of-breed and support, we lean towards Dell/Lenovo/HPE for primary racks, but it's good to note alternatives in case of supply constraints.	ASUS announcement (others TBD)

Networking:

Component	Description and Specs	Reference
NVIDIA Spectrum-4 Switch (Spectrum-X)	Ethernet switch ASIC family used in Spectrum-X platform (e.g., **Spectrum-4** MQM9700 series switches). Offers **51.2 Tb/s throughput**, supporting configurations like 64 ports of 800 GbE or 128 ports of 400 GbE. These switches have advanced features for AI: <u>**RoCE v2**</u> with sophisticated congestion management, **Dynamic Routing** (per-packet adaptive routing to avoid hot spots), and enhanced buffer architecture to deal with incast. They also support **collectives offload (SHARP)** in InfiniBand mode – not directly applicable in Ethernet mode, but similar can be achieved via NCCL + tree algorithms. Spectrum-4 latency is on par with IB (sub-1µs for port-to-port). For our cluster, a couple of 64-port 400G Spectrum-4 switches could serve as the spines, with 400G links to each rack's leaf. The Spectrum-X solution also includes software to orchestrate the adaptive routing and telemetry.	NVIDIA/ Naddod info
NVIDIA ConnectX-7 NIC	Dual-port NIC (often 200 Gbps per port, total 400 Gb/s) used in BlueField-3 and standalone. Supports **PCIe Gen4/5**, RoCEv2 and InfiniBand, GPUDirect RDMA, and in hardware: tag matching, reduction operations (for MPI/NCCL offload), and NVMe-oF offloads. We may use ConnectX-7 in systems where we don't deploy a full DPU. ConnectX-8, its successor, doubles speeds.	

Component	Description and Specs	Reference
NVIDIA ConnectX-8 / "SuperNIC"	Latest NIC, each device is 400 Gbps; in GB300 NVL72 design, two ConnectX-8 are combined into an 800 Gb/s module per node. It supports PCIe Gen5/6 and likely has enhanced offloads and security. We will incorporate these for maximum bandwidth – e.g., each 4-GPU tray gets 2×400G.	
NVIDIA BlueField-3 DPU	SmartNIC with **16 Arm A78 cores**, 2×100 GbE or 1×200 GbE ConnectX-7 NIC on board, plus accelerators (regex, crypto, etc.). BlueField-3 ("BF3") can handle up to 400 Gb/s with minimal host interaction. For us, BF3 will serve both as high-performance NIC *and* as an embedded network appliance on each node: running security policies, tenant vSwitch, storage offloads (NVMe-oF target initiator), etc. Each Grace+Blackwell tray could have one BF3 (maybe one per 2 GPUs or per 4 GPUs). BlueField-3 in **"embedded Trust"** mode can even be the only network interface, with the host OS's control plane separated. This might aid in compliance by locking down host networking. NVIDIA's literature emphasizes BF3's role in **RoCE packet reordering** and delivering guaranteed QoS per flow. We will use DOCA frameworks on BF3 to implement desired functionalities (like micro-segmentation).	
Cumulus NOS / SONiC	These are software options: Spectrum switches can run NVIDIA Cumulus Linux or SONiC. We would likely use Cumulus (which NVIDIA provides) as it's tailored for Spectrum hardware and has better support for Spectrum-X features out of the box. This gives us a Linux-based switch OS we can automate (NetDevOps with Ansible).	
Optics and Cabling	We'll use **400GBASE-DR4 or FR4** optical modules between racks (for up to 500m if needed) and possibly **Copper DACs** or **Active Optical Cables** within racks if distances are short (<=3m DACs for leaf-to-server). Ensure they are **Pan-FEU** (France and EU) certified sources if any restrictions (some government contracts require certain country-of-origin for network gear – we should verify if any such requirement exists). For neatness and speed, some integrated systems (Dell IR7000) come pre-cabled.	

High-Performance Storage:

Vendor / System	Description	Reference
Weka Data Platform (WekaFS)	Software-defined parallel file system. Typically runs on a cluster of x86 or ARM servers with NVMe SSDs. Delivers extreme performance (linear scaling, tested to millions of IOPS and 100s GB/s). Supports **POSIX, NFS, S3** interfaces. Key features: **Tiering to Object** (transparently offload cold data to S3), snapshots, encryption. Many AI orgs use Weka for training data (it's designed to handle the random read/write of AI training and checkpointing efficiently). In our cluster, we might deploy e.g. 8–16 Weka nodes (could even co-locate on some of the GPU nodes if needed, but better on dedicated storage nodes for isolation). Weka will utilize our 200/400G RoCE network (it has RDMA support). Its client can be installed on the GPU nodes, or accessed via standard protocols. Multi-tenancy: We can create separate file systems per tenant if desired. Also, Weka can ensure data protection (distributed erasure coding across nodes, etc.).	Weka docs
VAST Data Platform	Unified data store using all-flash. Presents an NFS and S3 interface (and SMB). Internally uses a scale-out architecture with storage "pods". Known for simplifying data tiering by using one tier of QLC flash + memory caching. VAST's performance is very high for large sequential reads and writes, slightly behind Weka in metadata perhaps. It's a good option if we want to reduce complexity (tenants could just use S3 or NFS on VAST for everything except the most IO-heavy training). VAST is also integrating with NVIDIA's Spectrum-X program, indicating we can get optimized performance on our Ethernet fabric. It does *not* automatically tier to another storage (because it assumes all data stays on flash with compression). Cost is mitigated by high compression ratios for AI data (often there's redundancy to exploit). We might deploy VAST for use cases like large-block reads (e.g., scanning through image datasets) and use Weka for heavy random I/O (small batch training). Both can coexist.	NVIDIA Tech Blog (mentions VAST integration)

Vendor / System	Description	Reference
DDN A^3I (Accelerated, Any-Scale AI)	DDN's solution stack for AI, typically comprising their **storage appliances (e.g., AI400X)** running Lustre or their own EXAScaler, plus the DDN Insight software for monitoring. A single AI400X appliance provides ~23 GB/s write, 90 GB/s read (depending on config) and they scale in a cluster. DDN storage was used in the MLPerf reference systems and is certified with NVIDIA (including with Spectrum networking). For our design, a DDN system could be used if we prefer a turnkey hardware storage. DDN also now has **NFS/GPFS** options (they acquired IBM Spectrum Scale rights for certain markets). Actually, IBM Spectrum Scale (GPFS) is another parallel FS known for reliability and multi-tenant use (GPFS supports multi-tenancy well, with fileset quotas and separation). However, GPFS licensing is expensive and performance for pure throughput is slightly lower than Weka for example. The Reddit discussion pointed out GPFS is "most reliable, albeit expensive". We likely stick to Weka/VAST unless a specific requirement pushes GPFS. DDN's advantage: **integrated solution with vendor support** – if we want one throat to choke for storage, DDN can do it. But mixing DDN with our open approach might conflict (DDN appliances can come with Red Hat, etc.).	Reddit HPC discussion (notes on GPFS, Lustre, DDN)
Pure Storage FlashBlade//S	A high-end NAS/Object appliance. Each chassis has multiple blades with NVMe. Delivers unified file and object, and easy scaling (add more blades). For AI, it's been used to store training checkpoints and as the backend for MLOps pipelines. Pure was notably the storage for the **NVIDIA DGX-1 SuperPOD** (first iteration) and is **Ethernet-based certified**. In our context, Pure FlashBlade could serve as an **auxiliary store** – for example, a place to dump model checkpoints or to serve datasets to many small jobs (its NFS performance for small files is top-notch). It might not reach the per-thread throughput of Weka for a single large job, but a lot of concurrent reads it can handle well. Also, multi-tenancy is straightforward – different shares/buckets per tenant. We could incorporate a FlashBlade to provide an **S3 API for tenants directly**, complementing Ceph. Pure also has a feature where it can replicate or tier to cheaper storage (FlashBlade//E or to S3 cloud) for older data, though their architecture is mostly all-flash.	

Vendor / System	Description	Reference
Ceph Storage Cluster	Our chosen **Tier-2** solution. We will build a Ceph cluster with, say, **several petabytes** of raw capacity using a mix of NVMe (for Ceph's internal metadata/index pool) and HDDs (for bulk data). Ceph will expose an **S3-compatible endpoint** that tenants or even our Weka system can use. We might run Ceph on commodity servers (maybe even utilize some of the Grace CPU nodes when not busy, though it's better to use separate storage servers with lots of disk slots). Given Grace is ARM and Ceph runs on ARM, we could have low-power high-core count Ceph OSD nodes (or consider using some AMD EPYC storage nodes if better suited for many HDDs). Ceph's performance: With enough tuning (thread pinning, BlueStore on NVMe WAL, etc.), we can get decent throughput but realistically HDDs will limit per-stream speeds. That's acceptable because this tier is for capacity. Ceph is open-source and free, though enterprise support is available from Canonical or others. Ceph integrates fine with Kubernetes (we can provide tenants with object storage and even block volumes via RBD if someone needed a VM with persistent disk). Ceph multi-tenancy: we will implement each tenant as a separate Ceph object user with quotas. Ceph can also enable encryption per pool. Possibly use CephFS for a shared space if needed, but CephFS is less performance and not necessary if we have other file systems.	– (Ceph docs for reference on tiering and multi-tenancy)
Backup/ Archive (optional)	If needed, an offline tier: e.g., **LTO tape library** or connection to a cloud archive for backups. Not in scope unless required by compliance (some government might want an offline copy of data). We note it but it's not core to design.	

This table (and discussion above) essentially forms a **library of up-to-date components** to choose from when configuring the supercluster.

Pricing and Procurement Notes: High-end AI hardware is expensive, but large purchases often come with significant discounts. Some indicative prices (from analyses and public info):

- NVIDIA **B100/B200 GPU list prices** are estimated around \$30k–\$35k for a B100 and \$60k–\$70k for a GB200 superchip (Grace + 2×GPU). That implies a 72-GPU NVL72 system might list at \$3M+ for GPUs alone. However, **hyper-scalers pay less**; SemiAnalysis noted a **GB200 NVL72 rack costs ~$3.1M to hyperscalers** hardware-only, and ~$3.9M including storage, networking etc.. This presumably reflects ~20–30% discounts off list. So per-GPU cost at scale maybe \$40k–\$50k. They also observed H100 8-GPU servers dropped to ~$190k each for hyperscalers (about \$23.5k per H100), so similar or slightly higher discount ranges may apply for Blackwell once supply stabilizes.

- **Network costs:** 400G switches are pricey (~\$1000 per 100G port equivalent typically), but again volume helps. NICs like ConnectX-7 might be few thousand each. BlueField-3 DPUs are maybe \$8k+

each at list (just an estimate from previous gen). Overall networking might be 10–15% of total cluster cost.

- **Storage costs:** All-flash systems like Weka/VAST will dominate storage cost if we go heavy there (think on the order of \$100/TB for high-end NVMe after factoring replication). Ceph with HDDs can be as low as \$10/TB raw. A balanced approach keeps storage cost maybe 15–20% of total.

When procuring at scale, we'd expect **20-40% off list** commonly. For example, large cloud deals often get 30%+ discounts (NVIDIA offered major deals to cloud providers recently). Our cluster being sovereign might limit some negotiation compared to a hyperscaler, but we can play Dell, HPE, Lenovo against each other for the best price on integration. Also note, certain governments have budget constraints so demonstrating TCO/performance is key; e.g. if GB200 is 1.6× the TCO of H100 per GPU, we need to justify that with the performance gain, which it does when utilized fully.

In terms of **availability (Q3 2025)**: GB200 systems are shipping (though in limited quantity early 2025), GB300 might be just ramping production. We may need to mix generations if supply is an issue (some initial racks of GB200, later racks of GB300 B300 GPUs when available). They are NVLink-compatible generationally or maybe not? (Usually NVLink domains can't mix architectures, so probably separate domains). But from a software view, they can still join the Ethernet cluster.

Deployment Plan: We will likely start with a **single rack NVL72** (or a couple) as a pilot, then scale out. Using vendor integrated racks (Dell IR7000 or Lenovo Neptune) should compress deployment time – e.g., Dell's white-glove service can deliver a rack ready to plug in within weeks of order. Automation using Palette or similar means we can have the software environment up in a day or two after hardware arrives, rather than months of manual setup.

Conclusion: By combining these best-of-breed components – NVIDIA's cutting-edge Grace-Blackwell GPUs in high-density liquid-cooled racks, a high-speed Ethernet fabric with advanced networking features, and a tiered storage architecture – we can design an AI supercluster that is **state-of-the-art for 2025**. It will be capable of training the largest AI models, while meeting multi-tenant isolation needs and sovereign cloud requirements in France/EU. All specifications and practices outlined are grounded in the latest documentation and real deployments, as evidenced by the cited reference designs and performance analyses. This positions us to build a **world-class GPU supercluster** that rivals those of leading AI labs, with the flexibility to support multiple organizations securely under one roof.

Sources: The information above was gathered from official NVIDIA product briefs, OEM press releases (Dell, Lenovo, Supermicro), technical analyses (SemiAnalysis, NVIDIA Tech Blogs), and other reputable sources, all cited inline. For further reading, please refer to the linked documents and whitepapers for details on specific components and best practices.

---

1 Dell Delivers Market's First NVIDIA GB300 NVL72 to CoreWeave | Dell
https://www.dell.com/en-us/blog/dell-delivers-market-s-first-nvidia-gb300-nvl72-to-coreweave/

2 3 Designed for AI Reasoning Performance & Efficiency | NVIDIA GB300 NVL72
https://www.nvidia.com/en-us/data-center/gb300-nvl72/