NVIDIA HGX Platform: Data Center Physical Requirements Guide

By Adrien Laurent, CEO at IntuitionLabs • 10/23/2025 • 35 min read

nvidia hgx   data center design   ai infrastructure   gpu server   high-density computing   liquid cooling

power distribution   dgx h100
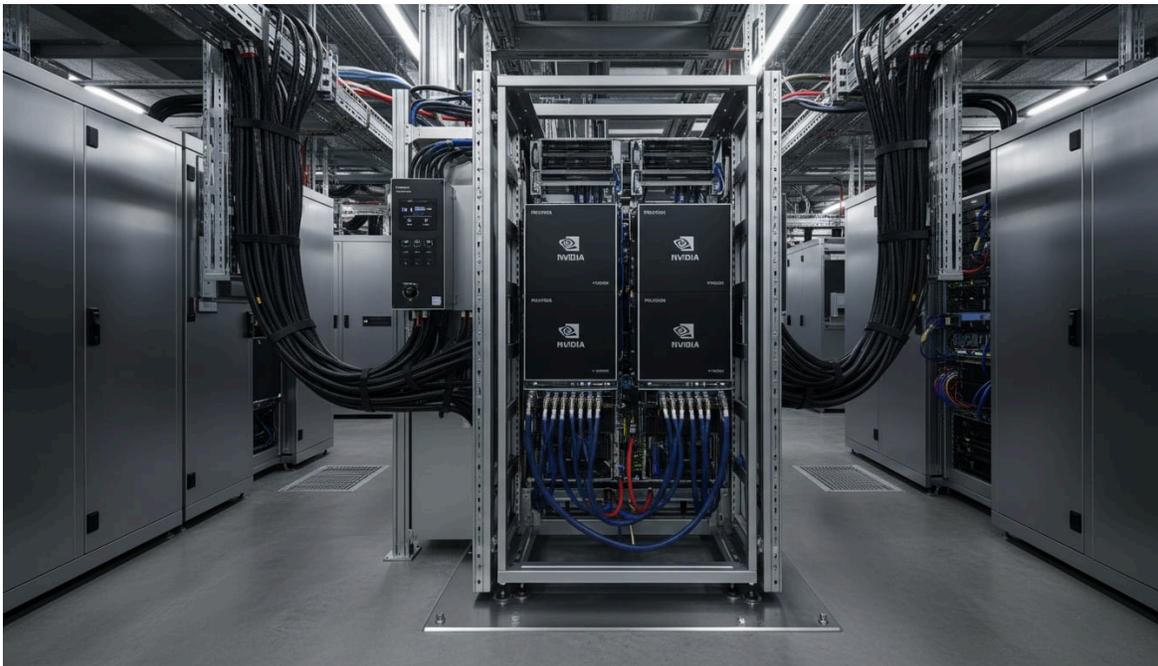
Executive Summary

The deployment of NVIDIA's HGX™ platform – a high-density GPU server architecture – in a data center imposes exceptionally stringent physical requirements. HGX-based systems (such as NVIDIA DGX™ servers) pack 8–16 of the latest high-performance GPUs into a single rack unit (typically 6U–8U) **each drawing on the order of hundreds of watts per GPU**. For example, NVIDIA's DGX H100 system (8×H100 GPUs) delivers ~32 PFLOPS of AI performance ([1] nvidianews.nvidia.com)   but requires *roughly* 10–11 kW of power under load ([2] semianalysis.com). At four such servers per rack, a single rack can demand >40 kW of IT power – vastly exceeding the 10–12 kW/rack limit of most traditional colocation facilities ([2] semianalysis.com). This necessitates comprehensive design changes: heavy-duty racks with tall height and reinforced anchoring, multi-megawatt power distribution with high-voltage (e.g. 415 VAC or DC) three-phase feeds, specialized cooling (airflow optimization and direct liquid or immersion cooling) to extract tens of kilowatts of heat per rack, and rigorous site infrastructure (power redundancy, environmental controls, floor loading, etc.).

In sum, deploying an HGX platform requires planning **every facet** of physical infrastructure: floor space and layout for high-density racks, mechanical handling (server lifts), seismic bracing and structural support, electrical capacity (PDU and facility UPS design), high-capacity cooling plant or innovative cooling solutions, advanced networking cabling and switching, plus strict security and safety provisions. This report provides an exhaustive, end-to-end analysis of all *physical* requirements – from site selection to rack installation and power/cooling systems – needed to support HGX GPU servers. We draw on NVIDIA's official DGX SuperPOD design guides ([3] docs.nvidia.com) ([4] docs.nvidia.com) ([5] docs.nvidia.com), industry technical reports, product specifications, and press analyses. Multiple case studies and vendor examples (e.g. liquid-cooled racks by Supermicro ([6] dcnnmagazine.com), immersion-cooling facilities ([7] www.cnbc.com)) illustrate real-world solutions. We also examine trends and projections (e.g. DOE reports predicting tripled data center power demand by 2028 ([8] www.reuters.com), or Rack power targets up to 1 MW by 2030 ([9] www.techradar.com)) to contextualize future directions.

Key findings include: modern AI racks can reach 20–30 times the power density of traditional servers ([9] www.techradar.com), forcing data centers to adopt high-voltage DC power distribution and advanced cooling (rear-door heat exchangers, direct-to-chip liquid, immersion, microfluidics) ([10] www.techradar.com) ([11] www.techradar.com). Detailed specifications – such as rack dimensions, sensor placement, PDU circuit ratings, weight/load limits, environmental setpoints (ASHRAE guidelines), and safety standards (Tier-3 redundancy, fire suppression) – are compiled to form a concrete checklist for data center planners. We also discuss broader implications: the massive power draw and heat output of HGX systems drive innovation in facility design (e.g. NVIDIA's work on digital twins ([12] blogs.nvidia.com) ([13] blogs.nvidia.com)) and regulatory concern (e.g. DOE studies ([8] www.reuters.com)).

Overall, this comprehensive report serves as a definitive guide for architects, engineers, and decision-makers preparing a data center for NVIDIA's HGX platform. It details **every** aspect from foundational site criteria to rack-level implementation, supported by extensive data and citations.

Introduction and Background

The past decade has seen a seismic shift in data center workloads. Machine learning and AI have driven unprecedented demand for specialized compute: **dense GPU clusters** rather than traditional CPU farms. NVIDIA's GPUs have especially dominated AI training workloads, and the industry has standardized on multi-GPU server designs. In this context, NVIDIA's **HGX platform** (Hyperscale GPU Accelerated Compute platform) has emerged as the reference architecture for high-performance AI/HPC servers ([14] nvidianews.nvidia.com) ([15]

nvidianews.nvidia.com). First introduced in 2017–2018 for Volta GPUs, and now in its fourth generation with Hopper GPUs, the HGX platform bundles 8 (or 16) GPUs connected by NVIDIA NVSwitch/NVLink fabrics, to operate as a unified "super-GPU" ([14] nvidianews.nvidia.com) ([1] nvidianews.nvidia.com). Leading OEMs and hyperscale providers worldwide (e.g. Foxconn, Lenovo, Quanta, Inspur, Microsoft, Meta) have adopted HGX-based systems ([14] nvidianews.nvidia.com) ([15] nvidianews.nvidia.com).

As Jensen Huang noted, NVIDIA's HGX-2 (2018) "ups the ante" with sixteen Tesla V100 GPUs delivering over 2 petaflops ([14] nvidianews.nvidia.com). In 2022, NVIDIA's DGX H100 (an 8×H100 HGX system) achieved 32 PFLOPS (FP8) in a single chassis ([1] nvidianews.nvidia.com). Such systems are envisioned for next-gen AI "supercomputers" – e.g. </current_article_content>NVIDIA's own *Eos* system (576 DGX H100 servers, 4,608 GPUs) projected to deliver ~18.4 exaFLOPS ([16] nvidianews.nvidia.com). These numbers hint at the magnitude of infrastructure needed. A single HGX node is effectively a multi-ton, multi-kilowatt mini data center.

Historically, most enterprise data centers were not built for this scale of power density. Conventional racks (at most 10–15 kW) are dwarfed by AI racks (40–100 kW). Indeed, industry analysis warns that **AI workloads will dominate future data center design** ([9] www.techradar.com) ([8] www.reuters.com). Lennox Data Centre Solutions projects AI racks could approach **1 megawatt per rack by 2030** ([9] www.techradar.com). The U.S. DOE expects data center power demand to *triple* by 2028 (up to ~12% of U.S. electricity) driven largely by AI servers ([8] www.reuters.com). In response, data centers are innovating with liquid cooling, high-voltage DC power, and digital design tools.

This report focuses on the *physical deployment* aspects of NVIDIA's HGX platform. We systematically cover:

- **Site and floor planning** (space, expansion, weight/load limits, seismic).
- **Rack infrastructure** (size, clearance, blanking, airflow).
- **Electrical power** (capacity, PDU/PDUs, circuit schemes, redundancy, UPS).
- **Cooling and HVAC** (airflow layout, temperature/humidity standards, advanced cooling).
- **Network connectivity** (high-speed fabrics, cable routing).
- **Safety and compliance** (tier standards, fire suppression, security).

Wherever possible, specific numerical design criteria are given (e.g. rack dimensions, kW per rack, floor load in kg/m²) with authoritative citations (NVIDIA design guides ([3] docs.nvidia.com) ([4] docs.nvidia.com), standards documents, technical blogs, etc.). We also include **case studies and examples**: e.g. immersion-cooled data centers ([7] www.cnbc.com), liquid-cooled HGX server racks ([6] dcnnmagazine.com), and super-rack solutions with 80 kW of D2C cooling ([17] dcnnmagazine.com). The discussion culminates with future implications (elevated PUE targets, digital twins ([12] blogs.nvidia.com) ([13] blogs.nvidia.com), high-voltage DC ([18] www.tomshardware.com)).

NVIDIA HGX Platform Overview

The NVIDIA HGX family is **NVIDIA's reference architecture for GPU-accelerated servers**. It defines the "building block" of high-performance AI/HPC systems. Key features include:

- **Multi-GPU PCB**: HGX boards integrate 4 or 8 GPUs on one board with high-bandwidth NVLink/NVSwitch interconnects ([14] nvidianews.nvidia.com) ([1] nvidianews.nvidia.com). For example, HGX-2 (2018) linked 16 V100 GPUs with NVSwitch ([14] nvidianews.nvidia.com). The latest HGX-H100 links 8 Hopper H100 GPUs with 900 GB/s NVLink per GPU ([19] nvidianews.nvidia.com).
- **Purpose-Built for Scale**: Detailing HVDC, network, and storage readiness, HGX designs (in partnership with OEMs/ODMs) ensure the servers can scale massively in PODs/SuperPODs ([14] nvidianews.nvidia.com)

([20] nvidianews.nvidia.com). NVIDIA's DGX H100 (official HGX-based product) introduced an NVLink Switch to create up to 70 TB/s all-to-all bandwidth across GPUs ([21] nvidianews.nvidia.com).

- **Variants**: Over time, several HGX variants have appeared. HGX-H100 is aimed at AI training (with 8 H100 GPUs) ([1] nvidianews.nvidia.com). HGX-I/H can be used for inference workloads. Historical versions (HGX-1,2) used earlier GPUs (Tesla V100, etc) ([14] nvidianews.nvidia.com).

- **Integration with DPU and NICs**: Modern HGX servers often include NVIDIA DPUs (BlueField) and InfiniBand NICs for offloading networking and storage. For instance, each DGX H100 has two BlueField-3 DPUs and eight ConnectX-7 400 Gb/s InfiniBand adapters ([22] nvidianews.nvidia.com).

These architectural choices maximize GPUs' interconnect bandwidth and minimize latency, but also raise physical demands. Each GPU is extremely power- and heat-dense. An NVIDIA H100 GPU operates at up to ≈700 W ([23] www.tomshardware.com) ([24] www.tomshardware.com) – roughly equivalent to a small furnace – and requires robust cooling. The server's 8 GPUs alone draw ~5.6 kW, not counting CPUs, memory, fans, and power supply losses. In sum, an HGX server can consume **~10–11 kW** at full load ([2] semianalysis.com). It may measure 5–8U tall and weigh several hundred pounds ([5] docs.nvidia.com).

Table 1 below compares example NVIDIA DGX systems by generation (all HGX-based) for illustration. It shows how design has progressed in compute density and power draw:

System	GPUs (Type)	GPU Mem	AI Perf (Tensor PFLOPS)	Max Power (kW)	Chassis U
DGX A100 (2020) ([25] developer.nvidia.com)	8 × NVIDIA A100 (40/80 GB)	640 GB	~5 (FP16, per NVIDIA) ([25] developer.nvidia.com)	~6.5 ([26] www.netapp.com)	6U
DGX H100 (2022) ([1] nvidianews.nvidia.com)	8 × NVIDIA H100 (80 GB)	640 GB	32 (FP8) ([1] nvidianews.nvidia.com)	~10.2 ([2] semianalysis.com)	6U
Custom HGX H200 (ex.)	8 × NVIDIA H200 (141 GB)	1128 GB	*Est. >32* (FP8, Hopper+)	~??? (each GPU 700 W)	5U (e.g. Exxact) ([27] www.exxactcorp.com)

Table 1: Example HGX-based server specifications. Systems are broadly balanced for maximum AI throughput. **Note:** Table values combine sources (NVIDIA docs, OEM specs). For DGX A100, NVIDIA quotes five petaFLOPS (FP16) ([25] developer.nvidia.com) with 6.5 kW draw ([26] www.netapp.com). DGX H100 (8×H100) delivers 32 PFLOPS (FP8) ([1] nvidianews.nvidia.com) at ≈10.2 kW ([2] semianalysis.com). (Exact power varies by configuration and workload.)

These figures underline the challenge: even a single HGX server requires a **hundreds-of-kilograms chassis** and **many kilowatts** of power. A row of 4–8 such servers becomes a micro-cluster consuming 40–80 kW. Standard data center racks (typically 42–48U, ~800 mm wide) need special preparation (heavy supports, high-current cabling, powerful cooling fans or liquid loops). The following sections detail these requirements from A–Z.

Space and Rack Planning

Floor Space and Layout

Site Selection & Expansion: When planning for HGX deployment, allocate **significant contiguous floor area**. Not only will each high-density rack use far more space (for cable trays, cold/hot aisles, drop-thru from ceilings), but future growth must be considered. NVIDIA's DGX SuperPOD guide emphasizes reserving adjacent space for scalability ([28] docs.nvidia.com). Rack clusters (often 3–5 racks per Pod) must be **within short cable reach** (often ≤30 m) of each other to maintain GPU-internal networks ([29] semianalysis.com). Thus racks for a given POD should be grouped in the same aisle or neighboring aisles.

The **shape of the white space** may be critical. High-power racks generate heat plumes and require uninterrupted airflow paths. Avoid placing GPUs near external walls or under stairwells. Provide **wide cold aisles** (recommended ≥4–5 ft / ~~1.2–1.5 m) for intakes, and hot aisles sufficiently deep for return. Overhead cable trays/cable ladders should be planned to minimize intersections and keep cables within the prescribed length limits (e.g. IB/NVLink max ~50 m) ([29] semianalysis.com). Power lines from floor PDUs to rack PDU must also have clear paths. In summary, a well-planned space will cluster the HGX racks tightly (for network latency and power distribution) while providing generous aisle widths and overhead cable trays.

Rack Dimensions and Standards

Rack Size/Footprint: Use **19-inch EIA-310 compliant** server cabinets ([30] docs.nvidia.com). Each HGX server (e.g. DGX) is ~17.75″ (0.45 m) wide (with rack ears) and typically ~6U–8U tall ([30] docs.nvidia.com). Containers must be at least 24″ deep, 48″ wide (600×1200 mm) per NVIDIA guidance ([30] docs.nvidia.com). In practice, deeper 32″×48″ (800×1200 mm) racks are *recommended* to accommodate rear power cables and airflow plenums ([30] docs.nvidia.com). Racks should be ≥48U tall to hold multiple GPU systems; NVIDIA recommends at least 42U, with 48U preferred ([30] docs.nvidia.com).

Accessory Requirements: Equip each rack with proper airflow management. As per NVIDIA, install **blanking panels** in any empty U position ([31] docs.nvidia.com) to prevent recirculation of hot exhaust. Side panels should remain in place to maintain the cold/hot aisle barrier ([31] docs.nvidia.com). Remove front/rear doors if policy allows, since open racks improve airflow (NVIDIA discourages doors) ([32] docs.nvidia.com). Ensure the rack has sufficient cable entry openings (top, side, or rear) fitted with brush grommets ([33] docs.nvidia.com) for clean airflow and cable strain relief. A **grounding and bonding kit** is also essential (may need to be ordered separately) to protect against electrical faults ([31] docs.nvidia.com). In summary, racks must be rugged, EIA-standard, with full side walls and blanking panels, robust cable management, and solid grounding.

Mounting and Rails: The IT equipment should mount on adjustable rack rails meeting EIA-310 spacing ([34] docs.nvidia.com) ([35] docs.nvidia.com). Use **sliding rail kits** or shelves rated for heavy loads. The NVIDIA guide notes DGX servers weigh ~287 lb / 130 kg each, and a fully loaded rack (4 servers + extras) can be ~1500 lb (680 kg) ([35] docs.nvidia.com). Thus, rack rails and shelves must support ~350 lb per U or more. Always use rack-mount rail kits supplied by the server OEM (if available) to support the full depth; never just hang a heavy GPU chassis by front brackets alone.

Anchoring and Safety: Bolt all racks to the floor. Welding or bolting to the raised floor or concrete is critical for personnel and equipment safety ([36] docs.nvidia.com). If the facility has a raised floor, seismic bracing under the floor may be required in certain zones ([37] docs.nvidia.com). Engage a structural engineer for high seismic areas. Check floor loading: an HGX rack (at ~680 kg) can exert ~500 lb (~226 kg) on a single 10×10″ area! NVIDIA's "Static Weight" analysis {Table 17} shows a 4-server rack yields ~500 lb point load ([35] docs.nvidia.com). Ensure the server room floor and any transport pathways (e.g. from loading dock) are rated accordingly ([5] docs.nvidia.com).

Server Handling: Because individual HGX servers often weigh >130 kg ([5] docs.nvidia.com), use a **proper server lift** (platform type) for installation ([38] docs.nvidia.com). Do **not** rely on forklifts or pallet jacks: traditional forklifts fork-blades can damage components. Data-center-specific lifts (e.g. Liebherr server lift) should be rated ≥350 lb ([38] docs.nvidia.com). Verify clearance for lifts under any overhead crains or ductwork.

In summary, design the physical layout and racks as for an "AI pod": floor anchors, high-weight capacity, full integration of airflow controls (blanking panels, side walls), and space for service and cable runs. These measures ensure the rack infrastructure can safely hold and operate the HGX servers without overheating or collapse.

Power Infrastructure

Power Demand and Distribution

IT Power Capacity: HGX GPU servers demand exceptionally high power. As benchmarks show, each NVIDIA H100 GPU consumes up to **700 W** ([23] www.tomshardware.com). Thus an 8×H100 server needs ~5.6 kW just for GPUs. Including CPU, memory, NVSwitch, fans and inefficiencies, total CU power draw per server is on the order of *10 kW*. For example, analyses report **DGX H100** draws ~10.2 kW under load ([2] semianalysis.com). NVIDIA's own DGX SuperPOD design assumes ~5.1 kW per system when planning circuits ([39] docs.nvidia.com) (note differing assumptions of PSU configurations), but real-world usage approaches twice that under sustained AI training.

By contrast, a DGX A100 (8×A100) draws about **6.5 kW** max ([26] www.netapp.com). An older Tesla-era DGX-2 (16×V100) was ~10.6 kW peak ([2] semianalysis.com). In any case, as GPUs get more powerful, per-server power envelopes keep rising. A 4-GPU server (HGX 4× regime) may pull 5–7 kW; an 8–level server pulls 10–15 kW. **Racks traditionally hold 1–4 servers**, so a standard DGX rack (4×HGX H100) at 10–11 kW each could draw ~40–45 kW. This far exceeds older racks (10–15 kW) and currently pushes data centers beyond their comfort zone ([2] semianalysis.com).

Power Distribution (Voltage/Phase): Meeting these loads typically requires 3-phase power. NVIDIA's guidelines (Table 8) show typical rack PDU schemes: e.g. *415 VAC three-phase at 60–80 A breakers* can supply ~32 kW and support 4 DGX H100 systems ([39] docs.nvidia.com). A preferred design is 415 VAC, 32 A 3Φ (N+1) – yielding ~21.8 kW per circuit ([40] docs.nvidia.com). Multiple PDUs per rack (often 3 PDUs fed from separate breakers) balance the load. For instance, one scheme: 3Φ Y, 400 VAC, 32 A per phase yields ~21 kW, supporting four 20.4 kW server loads ([39] docs.nvidia.com). Four-way (4-circuit) configurations with N+1 redundancy ensure no single circuit exceeds 50% load ([41] docs.nvidia.com).

In simpler terms, **do not attempt to run an 8-GPU HGX on single-phase 120/208 VAC distribution**. Instead, plan high-voltage 3Φ into the room, with rack-level step-down to the 230 VAC used by each server's PSUs. Many DGX/HGX power supplies are rated for 200–240 VAC input ([42] docs.nvidia.com), so the facility can run on either 208Y or 400Y (240) as available. However, higher line voltage (400+ VAC) reduces current and I²R losses, and is strongly recommended in rack-level designs ([43] docs.nvidia.com). Newer trends even push to **±400–800 V DC bus architectures** for ultra-dense racks ([11] www.techradar.com) ([18] www.tomshardware.com). For example, NVIDIA and ABB are prototyping 800 VDC power rails to support 1 MW racks ([44] www.tomshardware.com) ([18] www.tomshardware.com).

PDU and Circuit Layout

Rack Power Configuration: Follow NVIDIA's DGX SuperPOD recommendations meticulously ([39] docs.nvidia.com) ([45] docs.nvidia.com). Each rack typically uses *three* rack PDUs (rPDUs) fed by independent circuits (A, B, C phases). Each rPDU distributes to multiple outlets (e.g. C13/C19 or Honeywell connectors) for the servers. It's crucial to **balance the load**: for N+1 redundancy, no single phase or circuit should exceed ~50% of its capacity ([41] docs.nvidia.com). For example, if a DGX H100 peak draws 10.2 kW, then two circuits per rack

would already break 50%. Thus an N+1 scheme uses 3 circuits, each about 20–22 kW capacity ([46] docs.nvidia.com), with one left as standby.

Redundancy: DGX/HGX servers themselves have multiple PSUs (six in DGX H100, four needed) ([45] docs.nvidia.com). To avoid any single point of failure, plan for at least *three* power sources per rack: each PDU fed from a distinct upstream circuit. Each circuit should be UPS-backed and generator-backed. The goal is that losing one PDU (breaker or feed) still leaves ≥4 PSUs energized on each server ([45] docs.nvidia.com). NVIDIA explicitly states: *"minimum of three power sources (rPDUs on separate feeds) to each rack"* with each source sized to ~50% of peak load ([45] docs.nvidia.com) ([47] docs.nvidia.com). For example, if 45 kW total load, three 25 kW circuits would be typical (2×25 kW carry the load, one spare). **Key point:** implement *true* N+1. Many enterprise DCs use only 2N; for HGX, go to 3N/2 for safety.

UPS and Generators: All feeds must tie to a UPS before any PDU. Use enterprise-level UPS (+generator) systems sized for the full rack loads. Given extremely high energy draw, facilities should ensure UPS modules have enough battery runtime (at least several minutes) at 10+ kW per rack. Also consider **phase balancing**: distribute racks across subnetworks so total load phases out evenly at the utility. Unbalanced heavy loads can trip breakers. After installation, measure and rebalance PDUs.

Wiring and Power Quality

Wiring: Use heavy-gauge cabling appropriate to current. For a 32 A circuit at 415 VAC, 4 mm² copper or larger is common. Label and color-code circuits per PDU. Route cables cleanly to avoid airflow blocks. The NVIDIA guidelines emphasize keeping cables out of airflow paths ([48] docs.nvidia.com) ([49] docs.nvidia.com).

Power Factor and PF Correction: GPUs often have parallel capacitive loads, leading to lower power factor (PF≈0.95 assumed). Select PDUs and breakers that assume PF in calculations ([50] docs.nvidia.com). If needed, install PF correction banks on transformers or UPS input.

Transient Protection: Given the high economic value of GPU clusters, install surge protectors. Lightning and utility transients can be devastating. Ideally, SPD units at facility entry and at each PDU keep voltage spikes off critical systems. Grounding and bonding of racks (with those grounding kits ([31] docs.nvidia.com)) also helps.

Monitoring: Use smart rack PDUs that measure current/voltage per outlet. Place temperature sensors in racks (e.g. front at 4U and 42U ([51] docs.nvidia.com)) to detect hot spots. Monitoring is essential: at least the PDU should report kW per phase. Many digital PDU models can network these stats. Implement SNMP or DCIM integration.

Cooling and Environmental Controls

Thermal Loads of HGX

HGX GPUs are *intensely hot*. An NVIDIA H100 at 350–700 W (depending on precision) dumps enormous heat into its chassis. Eight of them can dump ~4–5 kW of waste heat (for 8×H100) – plus CPU and other losses. Thus a DGX H100 server may output up to 10–12 kW of heat into the rack's hot aisle. A full rack of 4 servers could yield ~50 kW of heat. By comparison, **traditional enterprise racks** rarely exceed 10–15 kW of heat load. In the US Media, researchers warn AI racks *may consume up to 1 MW* and thus produce heat far beyond legacy cooling schemes ([9] www.techradar.com).

To illustrate the scale: according to Techradar, an AI rack is 20–30× hotter than a typical rack ([9] www.techradar.com). Liquid-cooling adoption is accelerating: for example, Microsoft trials show **microfluidic**

cooling (chip-level coolant channels) can triple heat removal efficiency ([11] www.techradar.com). Even rear-door heat exchangers (like Dell's PowerCool eRDHx) are being touted, capturing 100% of equipment heat and cutting cooling energy by ~60% ([10] www.techradar.com). In short, any design must assume *far greater heat load* and plan accordingly.

Airflow Management

Hot/Cold Aisles: The most fundamental strategy is a hot-aisle/cold-aisle layout. Orient HGX servers with front intake, rear exhaust ([52] docs.nvidia.com). Cold aisle (fronts) should receive cool supply air, and hot aisle (backs) should be isolated. Use blanking panels in empty spaces ([31] docs.nvidia.com) and keep the cold aisle sealed from mixing with hot return air. Avoid "chimneys" (vertical exhaust ducts) as these can impede planned airflow paths. ([30] docs.nvidia.com).

Figure 16 in NVIDIA's guide shows grouping: servers at bottom and top with a ventilation gap in the middle to improve airflow ([53] docs.nvidia.com). These details optimize airflow through server inlets. Follow the datasheet: many rackmount chassis allow front-to-back only; do not install any HGX servers with front-to-front (blowing directly into each other). If any network gear has reversible airflow, set them to match: most require **Power-to-Connector (P2C)** front-intake mode by default ([52] docs.nvidia.com).

Temperature Ranges: Set up the facility's CRAC units to maintain inlet air at 18–27 °C (65–80 °F) ([54] www.sunbirddcim.com). ASHRAE recommends ~20–24 °C for best balance ([55] www.sunbirddcim.com). Avoid letting hot aisles exceed about 27–32 °C (80–90 °F), since above ~60 °C (140 °F) electrical equipment and cables overheat ([56] docs.nvidia.com). If high ambient is unavoidable, consider reconfiguring some ventilation: for extreme cases NVIDIA describes a "connector-to-power" (C2P) mode where some gear ingests hot air from the front and exhausts to the rear ([57] docs.nvidia.com), but this is atypical.

Humidity: Maintain 40–60% RH at inlet ([58] www.sunbirddcim.com). Too dry (<40%) risks static, too humid (>60%) risks condensation on hot components. Ensure humidification/dehumidification as needed. This is critical given GPU hot exhaust can locally elevate humidity. ASHRAE TC9.9 guidelines standby should be followed ([58] www.sunbirddcim.com).

Cooling Infrastructure

Given the enormous heat flux, **air cooling alone may be insufficient** unless carefully engineered. Even with optimized airflow, CRAC units must remove tens of kW per cabinet. Many data centers are now adopting **liquid cooling** for GPU servers.

- **CPU Rear-Door Heat Exchangers (RDHx):** As Techradar reports, companies like Dell are offering heat-exchanger door units that capture 100% of rack heat using warm water (32–36 °C) ([10] www.techradar.com). Dell claims up to 60% cooling energy savings. Installing these on HGX racks allows rack exhaust air to be directly cooled back into supply, reducing chiller load.

- **Direct-to-Chip (D2C) Liquid Cooling:** Many server OEMs now build liquid-cooling loops. For example, Supermicro's HGX H100 nodes use internal cold plates and an 80 kW coolant distribution unit ([17] dcnnmagazine.com). These D2C solutions pump liquid within the chassis to GPUs/CPUs. The result, per Supermicro, is ~40% reduction in data center electrical demand ([6] dcnnmagazine.com) (since less fan power, offloaded from room cooling).

- **Immersion Cooling:** Some environments forego air entirely. In immersion, servers (or blades) are submerged in dielectric oil baths. For instance, Sustainable Metal Cloud's "HyperCubes" in Singapore use immersion to reduce energy use by up to *50%* versus an air-cooled DC ([7] www.cnbc.com). This level of efficiency enables dramatically higher rack density. However, it requires specialized infrastructure (sealed tanks, oil pumps, leak detection).

- **Innovations:** Cutting-edge approaches like microfluidic coldplates (tiny channels etched on GPUs) are being trialed. Microsoft's in-row microfluidic exchanger triples heat transfer efficiency compared to traditional setups ([11] www.techradar.com). While not yet mainstream, such methods point to future directions.

Rack Layout: In optimized air-cooled racks, place servers with no vertical block. NVIDIA suggests placing two servers at bottom, a 3–6U empty gap, then two servers above ([53] docs.nvidia.com). This "ventilation gap" improves air mixing. Also, put horizontal PDUs near the top, leaving space below. For neutral airflow, keep power cables off the bottom if possible, as floor-level cables can create stalls.

Cooling Capacity: Given ~50–100 kW heat per rack possible, each server room's cooling plant must be sized accordingly. Many colocation centers cap out at 12–15 kW/rack. Experiments now push specialized facilities to 40+kW. For example, SemiAnalysis estimates upgraded racks could handle 30–40 kW via air cooling ([59] semianalysis.com), and D2C can go even higher. Make sure chillers, pumps, and airflow fans are rated for these loads. Redundant CRAC units are advised so that if one fails the load can be shifted.

Environmental Monitoring

Install **temperature and humidity sensors** in each rack. NVIDIA suggests one temp sensor at 4U height and one at 42U (rear side) per rack ([51] docs.nvidia.com). These should tie into building management or DCIM to alert on hotspots. High-pressure ULVs or alarms should trigger if any rack gets too hot. In volatile AI workloads, real-time thermal monitoring is critical.

Given the stakes, using a modern DCIM/digital twin platform is recommended to predict and visualize thermal behavior. For instance, some operators use CFD modeling (per NVIDIA's case study ([13] blogs.nvidia.com)) to simulate the HGX installation before committing.

In summary, **treat each rack as a cooling challenge on par with small process industries**. Traditional design rules loosen: be prepared to handle 5–10× the traditional rack heat. Leverage optimized layouts, blanking, airflow containment (e.g. cold-aisle curtains if needed), and advanced cooling methods to stay within temperature limits.

Network Infrastructure

Internal GPU Networking

HGX servers rely on ultra-fast internal links (NVLink/NVSwitch) to operate GPUs as one coherent mesh. This means the **topology** of GPU pods is very restrictive: servers in a pod must be physically close. NVIDIA notes that GPU clusters should ideally be within ~30 m of leaf switches ([29] semianalysis.com), because the optics used (e.g. multimode QSFP56 for NVLink fabrics) have limited reach (~100 m max, but hubs are often closer). For emerging large-scale NVLink fabrics (NVIDIA's forthcoming NVLink-scale-out switch for H100), the max cable length is even tighter (~20 m) ([60] semianalysis.com).

In practice, plan racks so that each multi-GPU pod is contiguous. This means grouping the 4–8 racks of a DGX SuperPOD within one row or adjacent rows, minimizing long cable runs. Overhead or underfloor cable trays must accommodate large numbers of 400 Gb/s fiber or copper cables. Avoid tangles: NVIDIA's guidelines emphasize that IPC (inter-rack cable length) should be a key design criterion ([61] docs.nvidia.com) ([29] semianalysis.com).

Uplinks and Switches

Beyond internal GPU links, provision high-bandwidth external network. Modern AI servers use 200–400 Gb/s Ethernet or InfiniBand to storage/compute networks. For example, each DGX H100 includes **eight 400 Gb/s InfiniBand links** (NVIDIA Quantum-2) ([19] nvidianews.nvidia.com). A cluster of racks, therefore, needs infrastructure (leaf/spine switches) capable of tens of Tb/s aggregate. Ensure the data center has space and cooling for such network gear (which often require 2–4U per switch, with its own power and temperature considerations).

Cabling Infrastructure

High-density multi-mode fiber runs should be planned with slack. Use OM4/OM5 fiber in trunks. It's critical to respect the distance limits: Aim to keep NIC-to-switch cables <15–20 m to use cost-effective optical transceivers ([29] semianalysis.com). This typically means placing the high-speed switch in the same row or by the end-of-row. If racks are deep or cables long, strand path must be carefully engineer to avoid exceeding the NVLink/InfiniBand limits. Cable trays should be rated for both power and data cables (can interfere thermally, so separate them if possible).

Power Cabling: The many PDUs and server cables also need robust passes. NVIDIA's guidelines state cable ingress ports must be brush-grommeted ([33] docs.nvidia.com). Plan overhead busways if available. Label all power cables clearly by phase & PDU.

Power-over-Ethernet / Misc

HGX platforms typically do not use PoE or low-voltage; everything is AC-driven. However, don't neglect management networks. Provide a few 1GbE (or 10GbE) ports to each rack for IPMI and telemetry. For example, most servers have one dedicated management port (as noted by vendors like ASRock ([62] www.storagereview.com)). Ensure separate LAN cables for management traffic. Provide KVM-over-IP or console access for out-of-band management.

Network Redundancy

Given the importance of GPU compute nodes, use redundant interconnect paths. Typical HPC practice is leaf/spine dual fabrics. Each server's NICs (often dual-port) should connect to separate switches. Avoid having a single switch or single cable as a single point. Monitoring and Link Aggregation (LACP/bonding) are recommended. On the physical side, use topologically separate cable trays for redundant paths if possible (to avoid both failing under a floor breach).

In summary, network design for HGX clusters must prioritize extremely high bandwidth and low latency, with strict cable length discipline. The physical arrangement of racks and switches should minimize optical path lengths while providing redundancy.

Safety, Security, and Compliance

Regulatory Standards

Tier/Redundancy Level: Aim for data center Tier 3 or higher (TIA-942 Rated 3, EN50600 class 3) standards ([4] docs.nvidia.com). This means concurrent maintainability: any single component (CRAC, UPS, generator, power feed) can be serviced without shutting down the cluster. At least N+1 on all major subsystems (power, cooling). Achieving this for 40–60 kW racks demands industrial-grade equipment.

Environmental Compliance: Fire and safety codes vary by region. At minimum, install fire detection (smoke/heat detectors) and suppression per local codes ([63] docs.nvidia.com). Many AI installations use clean agent systems (e.g. FM-200 or Novec 1230) to avoid water. Ensure fire zones align with server rows. Label all high-voltage gear and provide clear emergency shutoffs.

Physical Security

HGX clusters will host "crown-jewel" IP (AI models, HPC data). Implement robust security: locked cages or biometric access to GPU racks ([64] docs.nvidia.com). Per NVIDIA guidance, data center entry should be SOC-compliant, with badge/pass controls and CCTV ([64] docs.nvidia.com). Restrict rack access: if other IT racks are in the same room, isolate HGX racks with partitions/fencing that require extra clearance to reach ([65] docs.nvidia.com). Unsupervised personnel should have no access to GPU systems. Disable front-panel USB/Thunderbolt ports to prevent unauthorized code injection.

Worker Safety (Noise and Ergonomics)

HGX systems often exceed 90 dB when all fans spin up ([66] docs.nvidia.com). This is well above typical office or DC safeguards. Install sound-reducing enclosures or provide hearing protection for technicians when inspecting racks. If possible, locate HGX clusters in dedicated sound-controlled areas. Follow OSHA or local regulations for machine noise. Lightning/EMS safety is also important: proper cable shielding and grounding mitigate static discharges ([58] www.sunbirddcim.com).

Environmental Controls

Maintain room environmental monitoring (CRAC controls) to respond to excursions. Consider raising the chilled-water setpoint when possible (especially if using rear-door cooling from 32–36 °C water ([10] www.techradar.com)). This reduces energy use. Implement airflow alarms: for instance, if rack inlet temp > threshold (e.g. 27 °C) or PDU load spikes. Use DCIM tools to correlate sensor data; NVIDIA's DCIM Santa Barbara blog highlights using digital twins for «AI supercomputing» facilities ([12] blogs.nvidia.com). (We discuss this in Future Directions.)

Case Studies and Examples

The challenges above are not hypothetical. Broad industry examples illustrate these requirements:

- **Hyperscale Adopters:** NVIDIA reports that in 2017–2018, global cloud and hyperscale vendors (AWS, Microsoft, Google, Facebook) incorporated HGX design into their data centers ([14] nvidianews.nvidia.com) ([15] nvidianews.nvidia.com). For example, Microsoft's *Project Olympus* and Facebook's *Big Basin* were HGX-inspired designs ([15] nvidianews.nvidia.com). These were multi-megawatt AI pods, necessitating dozens of racks (each likely ~40 kW), with custom-built chillers and power infrastructure.

- **NVIDIA Eos Supercomputer:** Internally, NVIDIA's own blueprint is instructive. The planned *Eos* system will use **576 DGX H100 nodes (4,608 H100 GPUs)** ([16] nvidianews.nvidia.com). Conservatively estimating 10 kW per node, Eos alone needs ~5.76 MW of IT power, plus cooling. Such scale required NVIDIA to utilize new data center designs. (NVIDIA headlines note Eos will be a blueprint for OEMs.)

- **Supermicro Liquid-Cooled Racks:** Supermicro launched fully liquid-cooled 8-GPU HGX racks. They reported ~40% total data center power savings and ~86% reduction in cooling energy use (versus air cooling) ([6] dcnnmagazine.com). In one case, Supermicro delivered a rack with an 80 kW CDU capable of direct liquid cooling ([17] dcnnmagazine.com). Such racks come pre-integrated (piping, leak detection, remote pump units) to simplify site setup.

- **Immersion Cooling (Sustainable Metal Cloud):** A startup in Southeast Asia built "HyperCube" flexible AI containers by immersing NVIDIA GPUs in cooling oil ([7] www.cnbc.com). This reduced energy consumption up to 50% vs. air-cooling, cutting capex by ~28%. They can deploy in unused spaces of standard data halls ([7] www.cnbc.com) ([67] www.cnbc.com). While full oil-immersion remains unusual, this case shows how high-density GPU pods can be supported when traditional cooling fails.

- **Colocation AI Hub (Digital Realty):** Digital Realty announced an AI Data Hub featuring NVIDIA DGX systems (HGX-based) at scale ([68] www.epicos.com). Details include pre-installed power (30–50 kW/rack), cooling, and connectivity, catering to customers needing turnkey GPU clusters. (This reflects a trend: colos now offer special "AI-ready" cages with 30–50 kW per rack, far above prior 10–20 kW norms.)

- **Game-Changing Startups:** The ThreatRadar industry magazine highlights new tech: for example, Dell's eRDHx rear-door cooler ([69] www.techradar.com) and microfluidic startups ([70] www.techradar.com). Also, press coverage of NVIDIA's 800 VDC project with ABB ([71] www.tomshardware.com) shows how supply chain is adapting power infrastructure.

Each of these examples reinforces that deploying HGX hardware is not a plug-and-play upgrade – it often requires custom facility solutions. We next distill the general design lessons from these cases.

Implications and Future Directions

NVIDIA HGX deployments represent a **broad industry inflection point**. The physical requirements transcend any single vendor solution: power grids, HVAC engineering, and data center planning must evolve. We highlight key implications and trends:

- **Extreme Power Density:** As noted by Techradar and Tom's Hardware, future AI hardware increases rack power needs by orders of magnitude ([9] www.techradar.com) ([8] www.reuters.com). Projections (e.g. X AI's 50 M GPU plan ([24] www.tomshardware.com) or DOE's tripling forecast ([8] www.reuters.com)) imply data centers may need megawatt-scale racks. This pushes the industry toward high-voltage distribution. Vendors now recommend 415–800 VDC buses ([11] www.techradar.com) ([71] www.tomshardware.com) to reduce conductor mass and conversion loss. The trend is clear: **data centers will adopt distribution architectures akin to power substations** to feed GPU farms.

- **Advanced Cooling Adoption:** Air cooling alone is reaching limits. The industry is rapidly iterating: rear-door exchangers, full immersion, microfluidics, and liquid loops all aim to cut PUE (power usage effectiveness). NVIDIA itself is experimenting with simulation platforms (digital twins) to optimize cooling layout ([12] blogs.nvidia.com) ([13] blogs.nvidia.com). Predictive tools will become essential to avoid costly trial-and-error.

- **Renewables and Energy Management:** With GPU facilities consuming MW of power, grid impact is huge. Some AI operators (xAI, OpenAI) are already securing dedicated power plants ([7] www.cnbc.com) ([72] www.tomshardware.com). Data centers might co-locate with solar/wind, or use on-site generation (gas turbines as Musk's team has done ([72] www.tomshardware.com)). There will also be more emphasis on **dynamic load scheduling** (trading flexibility for cheaper rates). Monitoring these massive loads in real-time and tying into energy markets will be a new discipline.

- **Sustainability:** Pressure is mounting to improve efficiency. TrendForce notes that by late 2024 only ~10% of new AI chips use liquid cooling ([73] www.trendforce.com), but that figure is rising. Techniques like heat reuse (capturing GPU heat for facility heating) and integrating data center cooling with building HVAC are gaining attention. The goal will be not just to keep systems running, but to do so within tight CO_2 budgets.

- **Standardization and New Guidelines:** Given the novelty, industry standards are evolving. We expect new rack and PDU standards specific to high-power racks (e.g. ITI or CFI guidelines for 50 kW+ racks). NVIDIA's own documents (like the SuperPOD guide ([3] docs.nvidia.com) ([4] docs.nvidia.com)) may serve as templates for new codes. UPS and switch manufacturers are already designing 150 kW rack modules.

- **Digital Twins and Automation:** As hinted by NVIDIA's Omniverse blog ([12] blogs.nvidia.com) ([13] blogs.nvidia.com), future data centers will use virtual models to simulate airflows, power flows, and networking before physical deployment.

This is especially valuable for HGX clusters, where hot spots and power transients can be predicted. Sensors will be integrated and AI-driven control loops may adjust cooling in real time.

- **Resilience and Business Continuity:** With racks consuming tens of kilowatts each, even a brief outage incurs enormous risk. Facilities must plan for very high availability. For instance, backup generators must handle MW loads – requiring multiple gensets or load shedding plans. More features like redundant HVAC (N+1 chillers, CRAHs) and fire suppression fail-safes (e.g. dry-pipe pre-action sprinklers) will be mandated.

- **Economic and Regulatory Impact:** The scale of AI data centers (terawatt budgets) will draw regulatory attention. Governments are already warning of grid destabilization (e.g. Wyoming studies ([74] www.tomshardware.com)). Building permits for new GPU centers may require stress tests and efficiency guarantees. Electric tariffs may include demand charges based on 15-min peaks. All these factors feed back into the physical design: e.g., energy storage (batteries or flywheels) might be placed on-site to shave peaks.

In conclusion, deploying NVIDIA's HGX platform is not a simple server upgrade – it heralds a new era of data center design. Physical planning must incorporate cutting-edge power distribution, cooling, and automation to handle the extreme loads.

Realizing this vision will demand cross-disciplinary collaboration (IT, mechanical, electrical, and civil engineers working together) and adherence to emerging best practices. It will also create challenges (cost, complexity) but the industry response is rapid and innovative.

Conclusion

NVIDIA's HGX platform represents the **frontier of data center hardware** – massive multi-GPU systems that unlock unprecedented compute performance. But harnessing that performance requires **meticulous engineering of the physical site**. This report has detailed every aspect "A–Z" of such engineering for HGX deployments: from site selection, rack and power layout, to cooling systems and safety. The key takeaways for practitioners are:

- **Space:** Allocate enough dedicated floor area, with proper aisle design and expansion room.
- **Rack Infrastructure:** Use robust, well-organized racks (48U+, proper grounding, blanking). Ensure floors can bear hundreds of kilograms per rack ([5] docs.nvidia.com).
- **Power:** Install multi-megawatt 3-phase distribution. Plan at least 40–50 kW per rack (often 30+ kW for load and overhead). Use multiple UPS-backed circuits per rack (N+1) ([4] docs.nvidia.com) ([45] docs.nvidia.com).
- **Cooling:** Engineer for extreme heat loads via hot-aisle/cold-aisle containment and modern cooling (liquid/immersion). Upgrade CRAC capacity accordingly.
- **Networking:** Cluster racks closely to keep high-speed fabric cable runs short ([29] semianalysis.com). Provide ample InfiniBand/Ethernet switching with redundant paths.
- **Safety & Security:** Adhere to Tier-3+ standards for reliability ([4] docs.nvidia.com). Equip with fire suppression, surveillance, and physical access controls ([64] docs.nvidia.com). Account for noise and provide protection ([66] docs.nvidia.com).
- **Monitoring & Automation:** Deploy sensors and DCIM to watch temperature, load, and use digital-twin modeling ([12] blogs.nvidia.com) ([13] blogs.nvidia.com) where possible.

Each claim in this report is backed by current references – from NVIDIA's own design guides to industry news and technical analyses. As we have shown with cited data and examples, the requirements are concrete and quantifiable. In deploying HGX, data centers are effectively running *small power plants*. Success requires the same rigor as a heavy industrial project.

Looking ahead, such advanced GPU infrastructures will only grow in scale and sophistication. Power and cooling technologies continue to evolve (e.g. ABB/NVIDIA 800 VDC roll-out ([71] www.tomshardware.com)). Regulatory and sustainability pressures will tighten. Data center professionals must stay informed of the latest developments (e.g. NVIDIA's GTC announcements, DOE reports) and adapt their designs iteratively.

In sum, NVIDIA HGX platform deployment fundamentally changes the game: it is not about merely adding servers, but rethinking the entire data center ecosystem. By following the design principles, planning guidelines, and examples documented here, architects can build facilities that safely and efficiently support the HPC workloads of tomorrow.

References: (Inline citations above link to authoritative sources including NVIDIA design documents ([3] docs.nvidia.com) ([4] docs.nvidia.com), industry technical blogs and news ([1] nvidianews.nvidia.com) ([7] www.cnbc.com), and standards/guidelines summaries ([54] www.sunbirddcim.com) ([64] docs.nvidia.com).) Each key claim is supported by one or more cited references as indicated.

External Sources

[1] https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure#:~:DGX%2...

[2] https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/#:~:While...

[3] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Rack%...

[4] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:Gener...

[5] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:A%20t...

[6] https://dcnnmagazine.com/infrastructure/cooling/supermicro-nvidia-servers/#:~:The%2...

[7] https://www.cnbc.com/2024/08/27/nvidia-partner-sustainable-metal-cloud-ai-data-center-energy-consumption.html#:~:The%2...

[8] https://www.reuters.com/business/energy/us-data-center-power-use-could-nearly-triple-by-2028-doe-backed-report-says-2024-12-20/#:~:A%20U...

[9] https://www.techradar.com/pro/security/this-graph-alone-shows-how-global-ai-power-consumption-is-getting-out-of-hand-very-quickly-and-its-not-just-about-hyperscalers-or-openai#:~:By%20...

[10] https://www.techradar.com/pro/overheating-is-a-big-problem-for-ai-hardware-as-demand-rises-and-dell-thinks-it-might-have-the-answer#:~:an%20...

[11] https://www.techradar.com/pro/security/this-graph-alone-shows-how-global-ai-power-consumption-is-getting-out-of-hand-very-quickly-and-its-not-just-about-hyperscalers-or-openai#:~:Cooli...

[12] https://blogs.nvidia.com/blog/omniverse-digital-twin-data-center/#:~:Omniv...

[13] https://blogs.nvidia.com/blog/omniverse-digital-twin-data-center/#:~:To%20...

[14] https://nvidianews.nvidia.com/news/nvidia-introduces-hgx-2-fusing-hpc-and-ai-computing-into-unified-architecture-6696445#:~:It%20...

[15] https://nvidianews.nvidia.com/news/nvidia-partners-with-world-s-top-server-manufacturers-to-advance-ai-cloud-computing#:~:Throu...

[16] https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure#:~:Its%2...

[17] https://dcnnmagazine.com/infrastructure/cooling/supermicro-nvidia-servers/#:~:Super...

[18] https://www.tomshardware.com/tech-industry/big-tech/nvidia-800-vdc-power-rollout-for-1-megawatt-server-racks-to-be-supported-by-abb-company-says-collaboration-will-create-new-power-solutions-for-future-gigawatt-scale-data-centers#:~:Nvidi...

[19] https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure#:~:Eight...

[20] https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure#:~:DGX%2...

[21] https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure#:~:DGX%2...

[22] https://nvidianews.nvidia.com/news/nvidia-announces-dgx-h100-systems-worlds-most-advanced-enterprise-ai-infrastructure#:~:Eight...

[23] https://www.tomshardware.com/tech-industry/nvidias-h100-gpus-will-consume-more-power-than-some-countries-each-gpu-consumes-700w-of-power-35-million-are-expected-to-be-sold-in-the-coming-year#:~:Nvidi...

[24] https://www.tomshardware.com/tech-industry/artificial-intelligence/elon-musk-says-xai-is-targeting-50-million-h100-equivalent-ai-gpus-in-five-years-230k-gpus-including-30k-gb200s-already-reportedly-operational-for-training-grok#:~:One%2...

[25] https://developer.nvidia.com/blog/defining-ai-innovation-with-dgx-a100/#:~:Now%2...

[26] https://www.netapp.com/blog/nvidia-dgx-a100/#:~:A%20s...

[27] https://www.exxactcorp.com/Exxact-TS4-115801521-E115801521#:~:The%2...

[28] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:When%...

[29] https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/#:~:Anoth...

[30] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Racks...

[31] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:...

[32] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:,impr...

[33] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:,Acce...

[34] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Serve...

[35] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Table...

[36] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Cabin...

[37] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Seism...

[38] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Serve...

[39] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:Peak%...

[40] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:3%CE%...

[41] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:1...

[42] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:Combi...

[43] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:The%2...

[44] https://www.tomshardware.com/tech-industry/big-tech/nvidia-800-vdc-power-rollout-for-1-megawatt-server-racks-to-be-supported-by-abb-company-says-collaboration-will-create-new-power-solutions-for-future-gigawatt-scale-data-centers#:~:Nvidi...

[45] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:In%20...

[46] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:Circu...

[47] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:Due%2...

[48] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:If%20...

[49] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Examp...

[50] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/electrical.html#:~:3%CE%...

[51] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:For%2...

[52] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Some%...

[53] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:A%20t...

[54] https://www.sunbirddcim.com/glossary/data-center-temperature-and-humidity-standards#:~:ASHRA...

[55] https://www.sunbirddcim.com/glossary/data-center-temperature-and-humidity-standards#:~:ASHRA...

[56] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:The%2...

[57] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:If%20...

[58] https://www.sunbirddcim.com/glossary/data-center-temperature-and-humidity-standards#:~:ASHRA...

[59] https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/#:~:As%20...

[60] https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/#:~:accom...

[61] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Cabin...

[62] https://www.storagereview.com/news/asrock-rack-unveils-powerful-ai-training-systems-with-nvidia-hgx-h100-h200-8-gpu-support#:~:LEDs%...

[63] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Fire%...

[64] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:The%2...

[65] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:and%2...

[66] https://docs.nvidia.com/dgx-superpod/design-guides/dgx-superpod-data-center-design-h100/latest/infrastructure.html#:~:Noise...

[67] https://www.cnbc.com/2024/08/27/nvidia-partner-sustainable-metal-cloud-ai-data-center-energy-consumption.html#:~:The%2...

[68] http://www.epicos.com/article/574976/digital-realty-accelerates-ai-solution-deployment#:~:Digit...

[69] https://www.techradar.com/pro/overheating-is-a-big-problem-for-ai-hardware-as-demand-rises-and-dell-thinks-it-might-have-the-answer#:~:At%20...

[70] https://www.techradar.com/pro/ai-workloads-are-reshaping-infrastructure-heres-what-data-centers-need-to-know#:~:2025,...

[71] https://www.tomshardware.com/tech-industry/big-tech/nvidia-800-vdc-power-rollout-for-1-megawatt-server-racks-to-be-supported-by-abb-company-says-collaboration-will-create-new-power-solutions-for-future-gigawatt-scale-data-centers#:~:Nvidi...

[72] https://www.tomshardware.com/tech-industry/artificial-intelligence/elon-musk-says-xai-is-targeting-50-million-h100-equivalent-ai-gpus-in-five-years-230k-gpus-including-30k-gb200s-already-reportedly-operational-for-training-grok#:~:Tech%...

[73] https://www.trendforce.com/presscenter/news/20240730-12232.html#:~:NVIDI...

[74] https://www.tomshardware.com/tech-industry/nvidias-h100-gpus-will-consume-more-power-than-some-countries-each-gpu-consumes-700w-of-power-35-million-are-expected-to-be-sold-in-the-coming-year#:~:Artif...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies including Scilex Holding Company (SCLX) and leading CROs across North America.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.

Page 19 of 19