BUILDING AI FACTORIES:

THE NEW BLUEPRINT FOR DATA CENTERS IN THE BLACKWELL ERA



Executive summary

Data centers are no longer general-purpose compute facilities. They are AI factories, i.e., specialized production systems that convert electricity into intelligence tokens at scale. This is not a branding shift. It is a structural change in what the facility produces, how its performance is measured, and what its physical infrastructure must do.

The old metrics were storage capacity, IOPS, and power usage effectiveness. The new metrics are tokens per second, tokens per watt, and cost per token. Every rack, every coolant loop, every network port, and every power circuit is now evaluated by how much useful AI output it enables per unit of resource consumed.

NVIDIA's GB300 NVL72 systems deliver 50x more tokens per megawatt than the Hopper H100 generation, with 35x lower cost per token. Each rack in a Blackwell deployment consumes up to 142 kW. That density makes air cooling physically impossible and demands co-engineered power, thermal, and network infrastructure from the ground up.

NVIDIA's DSX platform defines how AI factories are designed through digital twins (DSX Sim), operated through unified software (DSX OS), optimized through power management (DSX MaxLPS), and connected to the grid (DSX Flex). Partners including Schneider Electric, Siemens, Jacobs, and Vertiv have built their facility solutions against the DSX specification.

Three groups need different answers from the same build. Executives need a capital case, a timing case, and a competitive case. Engineers need validated topologies, rack density limits, and thermal envelopes. Operators need serviceability, telemetry, failure modes, and control loops. This whitepaper addresses all three, following the path from economic logic to architecture, from power and cooling to operations, from deployment patterns to financial modeling, ending with case studies and a readiness roadmap.

50x	**142 kW**	**300x**
more tokens/MW vs Hopper generation *NVIDIA Blog, 2026*	per rack, GB300 NVL72 Blackwell deployments *Schneider Electric, 2026*	water efficiency gain vs air-cooled legacy systems *NVIDIA Blog, 2025*

KEY PRINCIPLE

The winners will not own the most racks. They will own the most token output per megawatt, the shortest time to first production, and the strongest ability to shape power and cooling around demand. That shifts the board conversation from hardware count to factory throughput.

Section 1: From Data Centers to AI Factories

1.1 The Rise of Intelligence Manufacturing

Traditional data centers stored and retrieved data. They served applications. Performance was measured in uptime and latency. AI factories do something categorically different: they manufacture intelligence and at an unprecedented scale. They take in raw compute power and produce tokens, the output unit of reasoning models, agents, and intelligent systems, continuously, at scale, 24 hours a day.

Think of it the way NVIDIA's CEO, Jensen Huang describes it: in the industrial age, power plants converted energy into electricity. In this AI age, AI factories convert energy into tokens.

The production analogy is precise here! An AI factory is a continuous production line where the raw input is electricity, the processing medium is accelerated compute, and the finished product is inference output measured in tokens per second.

The old data center model optimized compute pools and storage pools as separate assets. The AI factory model co-designs energy, chips, networking, storage, software, and facility systems from the start. A data center optimized for PUE alone is not an AI factory. The latter requires co-design across every layer at a level of integration that legacy facility thinking does not accommodate.

This shift also changes vendor selection. A site that once chose racks, chillers, and switches as separate purchases now needs a co-designed stack. Blackwell rewards tight integration because density, serviceability, and software all move together. A weak point in one layer now drags the whole site down. The practical outcome is a factory mindset: finance, facilities, network, security, and model teams should all review the same capacity plan. The AI factory only works when the business, the chip, and the building obey the same production target.

1.2 Why Blackwell Changes the Design Equation

The NVIDIA GB200 NVL72 and GB300 NVL72 rack-scale systems mark a step-function change in AI compute density. The GB200 NVL72 links 36 Grace CPUs and 72 Blackwell GPUs into a single NVLink domain, consuming 130 to 140 kW per rack. The GB300 NVL72 extends this to approximately 142 kW with Blackwell Ultra while delivering another step in tokens per watt.

Legacy data centers operated at 10 to 20 kW per rack. H100-era hyperscale deployments reached 60 to 80 kW per rack. Blackwell nearly doubles that again. At 140 kW per rack, the thermal and power architecture of the facility is not a secondary design consideration. It is the primary one that determines whether the compute inside it performs as specified.

The performance numbers justify that attention. According to NVIDIA, GB300 NVL72 systems deliver 50x higher throughput per megawatt compared to Hopper H100, translating to 35x lower cost per token and 40x higher revenue potential. For operators, each watt of facility power produces dramatically more useful AI output than two generations ago, but only if the facility is built to support the compute.

1.3 Three Audiences, One Factory

Executives should treat Blackwell as a capacity reset, not a GPU refresh. The competitive advantage goes to organizations that reach high token output per megawatt and the shortest time to first production. That shifts board conversations from hardware count to factory throughput and regional power strategy.

Engineers should see DSX as a validation layer. Simulation turns power, cooling, network, and layout into one design space before capital is committed. The factory rewards tight integration because density, serviceability, and software all move together. A weak point in one layer drags the whole site down.

Operators should see DSX as a control layer. When telemetry, workloads, and building systems share one operating model, the site becomes measurable, steerable, and faster to recover. The practical operating metrics differ from legacy KPIs: token output per watt, cost per token, utilization, and time to first production replace PUE as the primary dashboard.

Section 2: Economic Logic of the Blackwell AI Factory

2.1 From PUE to Tokens Per Watt and Cost Per Token

PUE measures how efficiently a data center delivers power to IT equipment. It was an excellent metric for general-purpose data centers because it captured real overhead inefficiency. It is an inadequate metric for AI factories because it says nothing about what the IT equipment produces.

Two AI factories with identical PUE scores can have radically different business outcomes if one runs Blackwell-class systems and the other runs legacy GPU hardware. The relevant metric is tokens per watt: how many inference tokens do you produce per watt of total facility power consumed.

Cost per token is the commercial derivative. SemiAnalysis InferenceX benchmarks cited by NVIDIA confirm that the GB300 NVL72 delivers the lowest cost per token available. At 50x higher throughput per megawatt versus Hopper, it lets operators produce more intelligence from the same power envelope at lower unit cost.

The executive view should treat AI infrastructure as production capacity. Every megawatt becomes a capacity asset with a token yield. A 50 MW hyperscale site can save over $4 million per year in cooling-related costs by shifting to GB200 NVL72 liquid-cooled systems. The real economic lift comes from higher model throughput, faster inference, higher user concurrency, and lower cost per completed task.

2.2 The AI Factory as a Continuous Production Line

NVIDIA describes the AI factory as a continuous production line: data flows in, models process it, agents act on the output, and applications deliver the result. Agentic AI systems reason, plan, search, retrieve data, write code, and take action, often spawning sub-agents that learn domain-specific skills. This always-on model means any infrastructure event that degrades throughput directly degrades token production and revenue.

A data center that serves request-response workloads tolerates bursty demand. An AI factory serving autonomous agents cannot. Utilization must remain high and consistent. Latency must be bounded. The operating model therefore needs a live view of token output per watt, rack thermal status, grid state, and storage latency.

2.3 CAPEX/OpEx Profiles for Blackwell-Class Clusters

Blackwell AI factory economics require careful planning across three cost centers. On compute, GB200/GB300 NVL72 systems represent a significant per-rack investment. Infrastructure to support them adds $500,000 to $2 million per megawatt of cooling capacity, according to Nautilus Data Technologies.

On operations, NVIDIA reports data centers spend $1.9 to $2.8 million per megawatt per year on energy alone, with cooling accounting for up to 40% of that figure. A Blackwell deployment achieves up to 25x cost savings on cooling energy. The business case rests on two simultaneous levers: higher revenue per rack from throughput, and lower cost per token from efficiency. Both improve together.

Blackwell also changes depreciation logic. Higher output density raises the value of the same footprint, the same grid connection, and the same land parcel. The site gains more economic work from each square foot and each kilowatt. A smart finance team models sensitivity around utilization first, then tariff second, then capex premium third. Underused Blackwell hardware destroys return just as quickly as underused Hopper hardware.

2.4 Hopper vs. Blackwell TCO Comparison

Metric	H100 / Hopper	GB200 NVL72	GB300 NVL72 (Blackwell Ultra)
TDP per chip	700W	1,000W+	1,200W (liquid-cooled)
Power per rack	~70 kW	~130 kW	~140–142 kW
Cooling	Air or DLC	Liquid required	Liquid required
Throughput per MW	Baseline	30x higher	50x higher
Cost per token	Baseline	Lower	35x lower vs H100
Revenue potential	Baseline	40x higher	50x higher
Energy efficiency	Baseline	25x better	30x better
Water efficiency	Baseline	300x better	300x+ better

Figure 1: Hopper vs. Blackwell performance and efficiency comparison. Sources: NVIDIA, Nautilus Data Technologies.

Section 3: Full-Stack AI Factory Architecture

3.1 The Five Layers of the AI Factory

NVIDIA's AI factory architecture organizes into five co-designed layers. Each depends on the others. Optimizing any single layer in isolation creates constraints in adjacent layers. The stack works only when the top and bottom layers align: a model team that asks for more tokens per second must accept the power, cooling, network, and storage consequences of that request.

- Energy Layer: Grid interconnect, substation, power distribution, UPS, generators. At Blackwell density, AI campuses require 100 MW to 1 GW of power. Energy availability is now a hard constraint on where AI factories can be built, often trumping land cost and fiber availability.
- Chips and Compute: Grace Blackwell GB200/GB300 GPUs, NVLink for scale-up within the rack, and the Grace CPU for CPU-side workloads. The unit of compute deployment is the NVL72 rack, not the individual server. This rack-scale model changes procurement, installation, and serviceability.

- Infrastructure: Networking (InfiniBand or Spectrum-X Ethernet), storage, and the physical facility including cooling, power distribution, and structural engineering. Infrastructure most often limits the performance of the compute above it.
- Models: Foundation models, fine-tuned enterprise models, and the training and inference pipelines. Model quality and inference optimization directly determine output value per watt.
- Applications: End-user and enterprise applications that consume inference output. Applications define query volume, latency requirements, and the workload shape the factory must serve.

Security and isolation belong in the architecture, not as a launch gate. NVIDIA addresses secure AI at factory scale through Confidential Computing and DOCA microservices within DSX. For multi-tenant or regulated deployments, the implications include workload isolation, identity-aware access, audit trails, and strict control over derived artifacts such as embeddings, tags, and model snapshots. The factory is not a compute cluster. It is a governed production system.

3.2 Core Compute: GB200/GB300, HGX vs. NVL72, NVLink Fusion

The GB200 NVL72 and GB300 NVL72 are rack-scale systems for trillion-parameter model inference and training. They differ from HGX server-level GPU configurations in a critical way: the NVL72 connects 72 GPUs within a single rack using NVLink Switch technology, creating a unified pool of GPU memory accessible as a single logical unit.

For workloads requiring massive model parallelism, running 405B or 671B parameter models in real time, the NVL72's unified memory architecture outperforms disaggregated server configurations. For smaller workloads or organizations prioritizing flexibility over peak throughput, HGX B200 configurations offer a less capital-intensive entry point.

NVLink Fusion widens the design space further for semi-custom systems, which matters when a campus needs to mix standard NVIDIA platforms with partner ASICs or custom silicon under a single fabric. The B200 GPU in liquid-cooled configurations delivers 20 PFLOPS FP4 performance at 1,200W TDP, that single number drives the entire facility engineering discussion.

3.3 Network Fabric: InfiniBand vs. Spectrum-X Ethernet

Network fabric is one of the most consequential architectural decisions. InfiniBand delivers the lowest latency and highest bandwidth for scale-out training where all-reduce communications between GPU nodes dominate performance. Spectrum-X Ethernet is tuned for AI inference at scale, with NVIDIA's Adaptive Routing and RoCE delivering competitive performance with the operational simplicity of Ethernet.

For large-scale training clusters at gigawatt-class AI factories, InfiniBand remains the performance leader. For inference-heavy deployments or Ethernet-native enterprises, Spectrum-X is a viable and increasingly competitive alternative. The decision depends on latency sensitivity, operational expertise, and whether existing staff have Ethernet or InfiniBand experience.

3.4 Storage for Trillion-Parameter and Multi-Agent Workloads

AI factory storage requirements differ from enterprise storage in two ways: access patterns (sequential large-block reads for model loading and checkpoint writing) and throughput demand (multiple GPU nodes demanding peak I/O simultaneously during training runs).

Multi-tier storage is the standard architecture. Hot NVMe flash serves active model artifacts, embeddings, and live checkpoints. Warm tiers handle datasets needing regular access. Cold tiers preserve archives and compliance copies. For agentic systems, the retrieval layer becomes part of the product surface, so vector stores and metadata pipelines deserve the same governance as the model itself. A design that treats storage as passive capacity will stall during checkpoint storms and retraining cycles.

Section 4: The NVIDIA DSX Blueprint

4.1 What Is DSX?

NVIDIA DSX is the full-stack platform that defines how AI factories are designed, built, and operated. Formally released in March 2026, DSX integrates validated reference designs, digital twin simulation, power optimization software, and a unified operational control layer into a single framework.

DSX is not a product you buy and deploy. It is a framework and ecosystem. NVIDIA defines the reference architecture and software interfaces. Ecosystem partners including Schneider Electric, Siemens, Jacobs, Vertiv, and Trane Technologies build their facility solutions against the DSX specification. The old approach treated rack, power, cooling, and operations as separate procurement events. DSX turns those choices into a single co-design problem.

> **In the age of AI, intelligence tokens are the new oil. DSX helps operators maximize the output of every watt they put into the factory.**

4.2 DSX Sim: Simulate First, Deploy Correctly

DSX Sim uses NVIDIA Omniverse digital twin technology to simulate the complete AI factory before any physical construction begins. Engineering and construction partner Jacobs uses DSX Sim to design the building, optimize compute density, and maximize token throughput against power constraints, all in simulation. The Omniverse DSX Blueprint brings facilities teams, OEMs, and operations groups into one shared model.

Partners including Cadence, PTC, Siemens, and Dassault Systemes integrate their facility design tools with the Omniverse DSX Blueprint, creating a live configurator that automates rack-to-facility design with validated accuracy. When teams discover a layout error in a digital twin, the fix costs engineering time. When they discover the same error after already poured concrete and advance ordered switchgear, the fix costs months. That gap alone justifies the simulation workflow for large builds.

4.3 DSX MaxLPS: Tokens Per Watt Optimization

DSX MaxLPS is the power management layer. It optimizes token output per watt within fixed power envelopes and performs real-time demand response for grid programs. MaxLPS monitors GPU utilization, cooling plant load, power distribution capacity, and workload scheduling. When power headroom is

available, it increases throughput. When the facility approaches limits, it applies intelligent workload shaping, not simple throttling, to maximize token production within constraints.

Phaidra, the AI-powered industrial automation company, has integrated DSX MaxLPS to operate the entire AI factory as a single unit of compute at scale. This demonstrates what AI-controlled facility optimization looks like in practice: continuous, autonomous adjustment across workloads, power, and cooling to keep tokens per watt at maximum.

Max-Q thinking changes the load profile. Instead of chasing maximum theoretical draw at all times, the factory sets a power envelope and tunes compute against the cap. That approach matters in regulated utility zones, in sites with congested interconnects, and in campuses that buy power under demand charges.

4.4 DSX Flex: AI Factory as Grid Asset

DSX Flex links the AI factory to the grid. The software receives utility signals, load shedding, demand response, pricing events, then adapts workload behavior across the facility. Energy leaders Emerald AI, GE Vernova, Hitachi, and Siemens Energy use the DSX reference architecture to unlock grid capacity and model substation behavior before site selection.

A facility that responds to grid signals can serve the grid during stress events and recover value during calm periods. The AI factory shifts from one-way consumption to two-way infrastructure. That role matters in regions where power access, not land, blocks the next phase of expansion. For the CFO, DSX Flex means a more resilient power contract and potentially a stronger license to expand.

4.5 DSX Exchange: Unifying IT and OT

AI factories generate continuous telemetry from thousands of sensors across compute, networking, power, cooling, and building systems. DSX Exchange provides the common communication layer that connects IT systems (servers, switches, storage) with operational technology (cooling plant controls, power distribution management, building management systems) and digital twin software in real time.

Without DSX Exchange, IT and OT operate in silos. Compute operators do not see facility power capacity in real time. Facilities teams do not see GPU utilization in real time. DSX Exchange collapses that separation, enabling teams to monitor, enforce policies, and optimize performance across the full AI factory from a single operational plane. The facility becomes software-defined, reading the grid, sensing thermal state, mapping workload demand, and adjusting in real time.

Section 5: Power Architecture for Blackwell-Class Density

5.1 Grid Interconnect and Campus-Scale Power Planning

AI factory campuses at gigawatt scale require power planning that begins at the transmission grid, not at the substation. A 1 GW AI campus requires grid interconnect agreements, new substation construction, and potentially new transmission infrastructure, all requiring permits, utility negotiations, and multi-year timelines.

Site selection is now primarily a power availability question. It is whether 100 MW to 1 GW of reliable, affordable power can be delivered within a competitive timeline. Grid interconnect lead time, transformer procurement, switchgear layout, and protection coordination now drive project schedule. Utility interconnect queues in constrained regions run months to years, making early engagement with transmission operators a first-year priority, not a pre-construction one.

5.2 The 800 VDC Power Path

NVIDIA has already moved the conversation toward 800 VDC for future AI factories and 1 MW IT racks. Traditional 54V in-rack distribution does not scale cleanly to megawatt-class rack design. Power conversion losses, copper weight, and distribution complexity all rise as density rises. 800 VDC architecture supports existing data centers while providing a path to a fully 800 VDC future. Less copper, fewer conversion stages, and more modular power paths improve efficiency and simplify assembly.

NVIDIA reports that newer power supply designs with energy storage can smooth AI load spikes and reduce peak grid demand. AI workloads do not draw power like office loads, they ramp, synchronize, and concentrate demand. The power train needs fast response and stable control logic to absorb those behaviors without tripping protection systems.

5.3 Power Distribution Inside the Factory

Inside the AI factory, medium-voltage power distribution from the substation steps down through transformers to low-voltage distribution at the data hall level. At Blackwell densities, AI factories increasingly use medium-voltage direct-to-rack (MVDC) architectures that reduce power conversion losses compared to traditional AC distribution chains.

Per-rack power distribution units must handle 140 to 160 kW of peak load. Busway distribution systems replace traditional cable runs because they are easier to reconfigure as rack layouts change during facility lifecycle. Campus design should anticipate future rack classes: a 100 MW site built with no headroom for tomorrow's density will struggle to absorb later GPU generations. Oversize conduit, distribution, and routing paths early enough to avoid a painful rebuild.

5.4 Designing for Max-Q Operation

Max-Q in the AI factory context refers to operating at the peak-efficiency point on the tokens-per-watt curve, not maximum power draw. GPU performance scales sub-linearly with power above a certain threshold. Slightly reducing the power cap often yields minimal throughput reduction while significantly reducing facility power load and cooling requirements.

DSX MaxLPS implements this optimization automatically. Your power distribution and cooling plant must be designed to support both the maximum power draw scenario for burst workloads and the Max-Q operating point for sustained production runs. Sizing only for Max-Q risks thermal runaway when workloads push toward peak power.

5.5 Resilience Architecture

AI factory power resilience follows the same principles as enterprise data center resilience but at significantly higher absolute power levels. UPS systems must bridge the gap between grid failure and generator startup, typically 10 to 30 seconds. At 100 MW campus scale, UPS battery capacity is a substantial capital investment and must be sized for the compute load, not just facility overhead.

Generator capacity at AI factory scale increasingly uses distributed generator farms rather than centralized units. Battery energy storage systems are also entering the design, shifting the meaning of resilience from backup alone to operational flexibility. A site that can shift workload, taper non-critical jobs, and participate in grid services has a stronger business case than a site that idles when the grid needs relief.

Operators should also plan for harmonics, transient events, and load sequencing. AI workloads do not draw power like office loads. They ramp, synchronize across thousands of GPUs simultaneously, and concentrate demand into short windows tied to training step boundaries or inference traffic bursts. The power train needs fast response and stable protection coordination logic so it absorbs those behaviors without nuisance tripping. At 800 VDC and beyond, cabling weight and conversion loss become strategic line items: fewer conversion stages and more modular power paths improve efficiency and simplify installation, which favors campuses that standardize the electrical spine across every hall rather than customizing each one.

Section 6: Thermal Architecture, Liquid-First Design

6.1 Why Air Cooling Is Obsolete at Blackwell Density

Air cooling reached its practical limit at around 40 kW per rack in production environments. Today's hyperscale AI facilities support over 135 kW per rack, making air cooling physically incapable of removing heat at the rate it is generated. The thermal differential that air cooling achieves at the chip surface is approximately 60 degrees Celsius. Liquid cooling reduces that to 17–20 degrees Celsius for single-phase direct-to-chip implementations.

Water has over 23 times the thermal conductivity of air and stores 3,000 times more heat per unit volume, according to Gartner's liquid cooling glossary. At Blackwell power densities, the choice is not air versus liquid cooling. The question of whether to deploy liquid cooling at all has been answered by the hardware.

DESIGN REQUIREMENT
At 1,200W TDP per GPU and 72 GPUs per NVL72 rack, air cooling is not an option for Blackwell-class deployments. Direct-to-chip liquid cooling is a prerequisite, not an optimization.

6.2 Direct-to-Chip Liquid Cooling Fundamentals

Direct-to-chip liquid cooling routes coolant through cold plates mounted directly on GPU dies and CPU packages. Coolant leaving the cold plates passes through a technology cooling system loop to a coolant distribution unit (CDU). The CDU uses a liquid-to-liquid heat exchanger to transfer heat from the technology loop to the facility cooling loop. The facility loop rejects heat through cooling towers, dry coolers, or heat exchangers depending on climate and design.

A key advantage is that the facility cooling loop operates at significantly warmer temperatures than air-cooled chillers require. In many climates, Blackwell deployments can reject heat to ambient air through dry coolers without mechanical refrigeration, eliminating chiller energy costs entirely during moderate temperature periods.

Reliability engineering changes substantially with liquid cooling at every level of the stack. Leak detection, isolation valves, materials compatibility, maintenance procedures, and pump redundancy all rise sharply in importance compared to a legacy air-cooled hall. Schneider Electric's RD110 and RD111 materials stress the role of delta-T management, CDU placement, and serviceability in liquid cooling blueprints. Operators who spent years managing fans and filters must now treat pumps, manifolds, and coolant chemistry as first-class maintenance objects.

6.3 GB200/GB300 NVL72 Thermal Specifics

The GB200 NVL72 system generates up to 130 kW of thermal load per rack at peak operation. The GB300 NVL72 raises this to approximately 142 kW. These are not peaks to be managed around, they are the continuous operating state for a fully utilized system running sustained AI inference workloads.

Managing this load requires a CDU per rack or per small cluster, with flow rates and delta-T engineered to match the load. Typical design targets use a supply temperature of 25 to 40 degrees Celsius and a delta-T of 10 to 15 degrees Celsius, achieving coolant return temperatures of 35 to 55 degrees Celsius.

Maintaining thermal headroom is critical. When GPU junction temperatures approach throttle thresholds, chips reduce clock speed to manage temperature. This throttling directly reduces tokens per second. An AI factory that consistently throttles due to inadequate cooling produces fewer tokens per watt than one with proper thermal headroom, defeating the entire economic case for Blackwell deployment.

A mature liquid-cooled site treats thermal telemetry as a live business signal. Supply temperature drift, return temperature drift, and pressure loss all affect token output. When the control room sees those signals early, the team corrects the issue before the GPUs throttle and the revenue curve bends downward.

6.4 Water Efficiency: A 300x Improvement

One of the most significant but underreported advantages of Blackwell liquid cooling is water efficiency. NVIDIA reports that GB200 NVL72 achieves 300x greater water efficiency than traditional air-cooled architectures. Direct-to-chip liquid cooling with closed-loop heat rejection eliminates evaporative water consumption in many configurations. In climates where free cooling is available for significant portions of the year, Blackwell facilities approach near-zero water consumption.

For AI factory projects in water-constrained geographies, this efficiency gain is not just an environmental benefit, it is a permit enabler. Regulators in water-scarce regions increasingly scrutinize data center water

consumption. A 300x reduction in water intensity makes Blackwell deployments viable in locations where legacy air-cooled facilities would face permitting rejection.

Section 7: Reference Designs for Grace Blackwell Facilities

7.1 The Role of Reference Designs

A reference design is an end-to-end validated architecture for building and operating AI factory infrastructure. It specifies not just individual components but their integration: how CDUs connect to facility water systems, what pipe diameters are required at which flow rates, how power distribution units connect to rack-level power, and how monitoring systems communicate with the DSX operational layer.

Reference designs reduce deployment risk and accelerate time to operation. Without them, every AI factory deployment requires custom engineering for each component integration, a process taking months and introducing errors. The old habit of buying hardware first and inventing the facility around it later no longer works at Blackwell density. The rack is too dense and the thermal risk too high.

Reference designs play three roles. First, they give the procurement team a known bill of materials and an approved integration pattern. Second, they give the engineering team validated assumptions for coolant flow, delta-T, filtration, and service clearances. Third, they give the operations team a maintenance model from day one.

7.2 Schneider Electric RD110 and RD111 for GB300 NVL72

Schneider Electric's RD110 and RD111 reference designs are co-developed with NVIDIA and purpose-built for Grace Blackwell GB300 NVL72 platforms, supporting up to 142 kW per rack.

RD110 targets standard deployment configurations where the primary heat rejection path uses Schneider's EcoStruxure power management integrated with NVIDIA DSX. RD111 addresses deployments with extended thermal requirements or facilities in climates requiring additional cooling plant capacity.

Both reference designs specify CDU placement relative to rack rows, required pipe diameters for primary and secondary coolant loops, manifold configurations for multi-rack deployments, filtration system specifications, isolation valve placement for maintenance access, and integration points with Schneider EcoStruxure and NVIDIA DSX Exchange.

For brownfield teams, the reference design becomes a retrofit contract. It tells the owner where the plant needs reinforcement, where the CDU belongs, and where the floor or wall path must change. That clarity cuts design churn and speeds board approval.

7.3 Scaling from a Single NVL72 Rack to Multi-Hall Campuses

A single NVL72 rack is the base deployment unit. AI factories scale by adding racks to a cooling pod, pods to a data hall, and halls to a campus. Each scaling step requires specific engineering considerations.

At the pod level (typically 4 to 12 racks), CDU capacity and the primary coolant loop must be sized for the aggregate rack load. At the hall level, chiller plant or dry cooler capacity must match the sum of all pod loads

with appropriate redundancy. At the campus level, power substation capacity, cooling tower capacity, and network backbone bandwidth must all scale together.

DSX Sim models this scaling accurately, allowing engineers to validate that cooling plant capacity will meet thermal loads at each scale increment before committing capital to equipment procurement. The most effective campuses preserve the same rack class, coolant standard, and network topology across generations so the operations team does not relearn the site each quarter.

Section 8: Lessons from H100, Pathfinding for Blackwell

8.1 H100 Thermal and Power Profile vs. Blackwell

The H100 SXM5 GPU operates at 700W TDP. Dense H100 rack deployments reached 60 to 80 kW per rack. Nautilus Data Technologies notes that Blackwell variants push 1,200W TDP in liquid-cooled configurations, a 70% increase in per-chip power draw and near-doubling of per-rack thermal load. Organizations that mastered H100 liquid cooling infrastructure are in the strongest position to scale Blackwell efficiently.

8.2 What H100 Deployments Proved

The H100 deployment cycle proved three things that directly inform Blackwell planning:

- Air cooling hits hard limits at scale. Companies deploying thousands of H100 GPUs discovered that maintaining optimal sustained performance requires single-phase direct liquid cooling. Facilities that attempted air cooling at H100 density faced persistent thermal throttling that degraded throughput and made SLA compliance difficult.
- Thermal management directly impacts performance outcomes. Chips at optimal temperatures maintain higher clock speeds longer. Even small improvements in thermal management compounded into meaningful performance gains across long training runs. The same principle applies at Blackwell, amplified by the higher TDP.
- Infrastructure decisions during H100 deployment determine Blackwell readiness. Organizations that deployed H100 with sub-optimal cooling to save capital face complete retrofitting for Blackwell. Those that built proper liquid cooling for H100 accommodate Blackwell with incremental upgrades to pumps, CDUs, and pipe capacity.

Teams who learned liquid cooling during H100 deployments developed stronger operating muscle memory than teams who waited. The first liquid-cooled generation teaches the exact habits that keep Blackwell stable: startup sequencing, maintenance windows, contamination control, and incident response.

8.3 Required Upgrades for Blackwell Transition

- Pump and flow rate capacity: Blackwell's higher thermal load requires higher coolant flow rates. Existing pumps must be rated for the increased flow requirement, or you need upgraded pump modules.
- Pipe diameter: Higher flow rates require larger diameter primary coolant piping to manage pressure drop across the loop. Undersized piping creates flow restrictions that reduce cooling capacity below GPU load demands.

- CDU capacity per rack: H100-era CDUs rated for 60–80 kW per rack cannot handle 140 kW. You need CDUs rated for the Blackwell thermal envelope with appropriate safety margin.
- Facility cooling plant: Cooling systems sized for H100 rack densities need capacity expansion before Blackwell deployment. Audit current capacity utilization before assuming it can absorb the increased load.
- Structural loading: High-density Blackwell racks and coolant skids increase floor loading. Floor reinforcement and anchoring strategy need review before procurement locks the design.

8.4 Using H100 Sites as Brownfield Launchpads

Nautilus recommends auditing current H100 liquid cooling installations to document what worked and what required retrofitting. That institutional knowledge transfers directly to Blackwell planning. Calculate your thermal capacity headroom: if your facility runs at 80% of cooling capacity on H100 workloads, expanding to Blackwell density likely requires facility upgrades taking six to twelve months.

For brownfield owners, the strategic question is not whether Blackwell fits. The question is where the weakest point in the existing plant sits. Some sites have enough utility capacity but weak cooling. Some have strong cooling but weak switchgear. Some have structure and power but lack routing space for liquid infrastructure. A good H100 baseline gives enough operational evidence to answer those questions before a larger capex cycle begins.

Section 9: Mechanical, Electrical, and Plumbing Blueprint

9.1 White-Space Layouts for NVL72 Racks

White-space planning changes in a Blackwell factory. The floor is no longer a generic open area with racks in rows. The layout must preserve service corridors, coolant routing, cable management, and maintenance swing space. Vertiv and Schneider recommend aisle widths and row lengths chosen for access as much as for density, especially critical when CDUs must be serviced in place without interrupting adjacent production racks.

The most effective campuses preserve the same rack class, coolant standard, and network topology across halls so operations teams do not relearn the site each quarter. Design variance creates maintenance variance. Maintenance variance creates cost. Standardization lowers both.

9.2 Equipment Room Design

The equipment room holds CDUs, heat exchangers, pumps, valves, sensors, and control panels. ASHRAE guidance on data center cooling recommends treating the equipment room with the same rigor as the white space: access paths, isolation valves, and maintenance corridors deserve as much attention as cable trays.

When liquid cooling and high-density power share the same white space, the layout must prevent one system from blocking service to the other. Mechanical design and electrical design now intersect more tightly than before. A facility that ignores this intersection during design typically pays for it through downtime and service friction.

9.3 Structural Considerations

High-density racks, coolant skids, and upstream plant gear increase floor loading and vibration risk. Seismic requirements matter in many regions. A useful planning rule: design for the load class the site will carry after the second generation refresh. If that load exceeds the original plan, reinforce now instead of staging a second expensive retrofit later.

9.4 Safety, Leak Detection, and Maintenance Access

The plumbing layer deserves the same rigor as the electrical layer. Leak detection, drip management, isolation valves, pressure transducers, and maintenance bypass paths all need to exist before the first rack goes live. Facilities that treat coolant as a side utility usually pay through downtime and service friction. The coolant chemistry and material compatibility plan should match the system life cycle.

Operators need a policy for fill, treatment, sampling, and replacement. They also need vendor standards for seals, tubing, and corrosion prevention. These are mundane topics until a leak or contamination event arrives, then they become uptime topics.

Maintenance access deserves equal weight in the design phase. CDUs, manifolds, and isolation valves should be reachable without powering down adjacent racks, and service technicians need clear swing space to pull a failed pump or replace a filter without disturbing live production. Facilities that retrofit access paths after commissioning routinely find that the most failure-prone components sit in the hardest-to-reach locations, simply because access was treated as an afterthought rather than a design input from day one.

9.5 Integrating Building Management Systems with DSX

The BMS should not sit in isolation from the AI stack. A strong design exposes cooling state, alarms, and power events to the same operational layer that tracks workload health via DSX Exchange. A useful operating pattern places the BMS, the cluster scheduler, and the AI optimization layer in a shared decision loop. When the cooling system sees a supply-temperature change, the workload layer should see the same signal. When the power train hits a demand event, the scheduler should know whether a delay, migration, or cap event protects service levels.

Structural engineers should also ask: what load class will the site carry after the second generation refresh? If the answer exceeds the original plan, the team should reinforce now instead of staging a second expensive retrofit later.

Section 10: Network and Storage Fabric for Trillion-Parameter Models

10.1 Training Cluster Topologies

Large-scale AI training uses all-reduce communication patterns where every GPU must exchange gradient information simultaneously. This demands a network topology with high bisection bandwidth and low congestion. Fat-tree topologies using NDR InfiniBand at 400 Gbps per port remain the standard for demanding training clusters. Dragonfly topologies offer cost-effective long-range connectivity between AI factory pods or between geographic sites but introduce some non-uniformity in all-reduce performance that training frameworks must account for.

10.2 Latency, Bandwidth, and Congestion Control

Inference workloads have different network requirements from training. For inference, end-to-end latency per request is often more important than raw throughput. Agentic AI workflows, where a single user request triggers dozens of sequential model calls, amplify this sensitivity. A 10-millisecond delay per model call becomes 100 to 300 milliseconds of added latency across a multi-step agent workflow.

The hidden network issue is not only bandwidth. Congestion control, telemetry, and topology awareness matter just as much. A fabric that behaves well under average traffic but fails during synchronized all-reduce phases will slow the whole training run. A good AI factory measures fabric health at the same cadence as power and cooling, with network observability living close to the scheduler so congestion triggers workload adjustment before users feel the degradation.

10.3 Storage Architectures for High-Throughput Workloads

Training a trillion-parameter model generates checkpoint files of hundreds of gigabytes at regular intervals. Checkpoint write bandwidth requirements of multiple terabytes per second are realistic, demanding parallel distributed storage systems designed for write-intensive AI workloads.

For trillion-parameter and multi-agent workloads, data locality becomes a business issue. If the retrieval layer sits far from the inference layer, latency rises and token cost rises with it. If the storage layer lacks fast parallel reads, checkpoint and restart times stretch. The best designs place compute, storage, and network policy in one planning loop, then use telemetry to keep the loop honest after launch.

10.4 Multi-Cluster and Multi-Region Connectivity

For multi-cluster deployments, the network layer also shapes business continuity. A site with common network policy across clusters can move workloads between pods, facilities, or regions with less pain. That flexibility helps during maintenance, local outages, and growth phases. AI workload portability requires consistent fabric policy across sites, if policy diverges, job migration introduces new failure modes.

Section 11: AI Factory Operations, DSX OS and Run-Time Control

11.1 DSX OS as the Facility Operating System

DSX OS is the operational layer of the NVIDIA DSX platform. It provides unified cluster operations across compute, networking, storage, power, and cooling. Think of it as the operating system for the physical AI factory: just as a server OS schedules processes and manages resources, DSX OS schedules workloads and manages physical resources across the facility in real time. Once the physical AI factory comes online, the digital twin built with DSX Sim becomes an active operations tool, continuously reflecting the current state and enabling operators to test configuration changes in simulation before applying them to production.

11.2 Real-Time Optimization Across Power, Cooling, and Workloads

The distinctive feature of AI factory operations versus legacy data center operations is the tight coupling between workload scheduling, power management, and cooling control. In a conventional data center, these domains operate largely independently. In an AI factory, they must co-optimize in real time.

A workload shift that moves compute load from one facility zone to another changes local thermal load, which requires the cooling plant to redistribute coolant flow, which may change the power draw of the cooling plant itself. DSX Exchange ensures all three domains see changes simultaneously. DSX MaxLPS determines the optimal power allocation given current cooling headroom and grid conditions. DSX OS implements the workload scheduling that results from that optimization.

11.3 Token-Per-Watt Observability and AI Agents

Operations in an AI factory look very different from a legacy data hall. Phaidra's work demonstrates AI agents orchestrating cooling, power, and workload management to maximize tokens per watt. AI agents flag anomalies, predict maintenance windows, tune setpoints, and suggest workload placement before thermal or electrical limits hit. The operator still owns the final decision, but the loop becomes faster and more informed.

Token-per-watt reporting belongs next to uptime reporting. If a workload loses throughput while the platform reports a healthy cluster, the site still fails the business. The operating team should set separate alerts for thermal drift, power drift, and token drift, a facility can look healthy in one domain and fail in another.

A mature operations desk should run on thresholds, not intuition. Thresholds define when to cap power, when to shift workloads, when to enter maintenance mode, and when to escalate. A good dashboard shows whether the token stream rises because the model improved or because the plant got cooler. The difference matters to finance and to engineering.

Section 12: Deployment Patterns and Migration Paths

12.1 Greenfield Gigawatt-Class AI Factory Campuses

Greenfield campuses suit the largest builds. They allow utility interconnect design, substation placement, water strategy, road access, and hall orientation to fit the AI factory from the start. The NVIDIA AI Factory Research Center at Digital Realty's Manassas, Virginia site demonstrates this approach, facility design was validated in DSX Sim digital twin before ground was broken, and the build-out process is captured as a scalable, repeatable recipe.

The downside is lead time. Greenfield campuses at gigawatt scale require years from site selection to first production. Campus design should anticipate future rack classes. A site built around today's density with no headroom for tomorrow's density will struggle to absorb later GPU generations.

12.2 Brownfield Upgrades and Retrofit Strategies

Brownfield retrofits move faster because the building shell already exists. The challenge is compromise. Existing structural limits, legacy switchgear, and water systems force a more careful design. The best Blackwell migrations isolate the highest-value hall first. That hall becomes a proving ground for power, liquid cooling, and operations. After the first hall stabilizes, the next hall reuses the same pattern with fewer surprises.

12.3 Modular and Prefabricated Approaches

Modular builds sit between greenfield and brownfield. Prefabricated AI blocks from partners like Vertiv and Bechtel lower project risk and let the business align capex with demand. Prefabricated pods, repeatable mechanical blocks, and staged capacity additions compress construction timelines by shifting assembly work from the building site to controlled factory environments. For organizations needing to deploy Blackwell capacity rapidly, modular approaches can reduce time to first production from years to months.

12.4 Deployment Model Decision Matrix

Deployment Model	Capital	Timeline	Best For	Key Risk
Greenfield campus	Very high	3–5 years	Hyperscale, new geography	Lead time, power permits
Brownfield retrofit (H100-ready)	Moderate	6–18 months	H100 sites with liquid cooling	Structural limits
Brownfield rebuild	High	18–36 months	Legacy air-cooled sites	Compromise & cost overrun
Colocation	Low–moderate	3–9 months	Capex avoidance, fast start	Vendor dependency
Modular/prefab blocks	Moderate	6–18 months	Uncertain demand, fast deploy	Higher per-unit cost

Figure 2: Blackwell deployment model comparison.

The decision tree is direct. Greenfield wins when power, land, and strategic scale matter most. Brownfield wins when time and sunk assets matter most. Modular wins when demand is uncertain or when the organization needs a fast first step with low structural commitment. DSX and digital twin design reduce risk in all three cases because simulation reveals the cost of each choice before steel and concrete lock the path.

Section 13: Sustainability, Water, and Regulatory Considerations

13.1 Beyond PUE: Tokens Per Joule as the Sustainability Metric

Sustainability reporting for AI factories is evolving beyond PUE. The metric that aligns infrastructure efficiency with business output is tokens per joule: how many tokens of inference output does the facility produce per joule of energy consumed. Blackwell's 50x throughput-per-megawatt advantage over Hopper directly improves tokens per joule. For organizations with energy consumption commitments or science-based targets, this improvement is material.

The IEA's Energy and AI report frames the macro picture: AI workloads are among the fastest-growing segments of electricity demand globally. The organizations that measure tokens per joule rather than PUE alone will have a defensible narrative with regulators, investors, and utility partners.

13.2 Water Efficiency and Site Selection

The global liquid cooling market is projected to grow to $21.3 billion by 2030, driven partly by water efficiency advantages over legacy air-cooled architectures. Direct-to-chip liquid cooling with closed-loop heat rejection eliminates evaporative water consumption in many configurations. Water strategy should begin

with local context: some regions face water stress, others face grid stress. The site design needs to solve the local constraint first.

13.3 DSX Flex as a Sustainability Lever

DSX Flex adds a practical sustainability mechanism. If the factory can shift work in response to price signals or grid stress, the site reduces peak demand and improves grid compatibility. When renewable generation creates surplus capacity, DSX Flex absorbs that energy through increased compute throughput. The important outcome is not only ESG narrative, it is a more resilient power contract and a stronger license to expand.

A sustainability report looks stronger when token output joins energy and water data. That metric set shows business value and resource discipline in one frame. It also gives regulators and local stakeholders a clearer view of what the campus produces relative to what the campus consumes.

13.4 Regulatory and Permitting Landscape

AI factory projects face regulatory scrutiny on four fronts: energy consumption, water use, land use, and local grid impact. In the European Union, the Energy Efficiency Directive requires large data centers to report energy consumption, PUE, and renewable energy percentage. Environmental review, zoning, and grid interconnect can all extend project timelines by months or years.

Early community and utility engagement lowers the chance of surprise. A factory-scale AI campus needs a permit strategy as much as a technology strategy. On-site renewable supply, storage, and flexible load logic belong in the design brief, not as an annual report exercise, but as active elements of the site's power contract and expansion roadmap.

Section 14: Risk, Resilience, and Multi-Generation Planning

14.1 Planning for Hardware Generations: Blackwell to Rubin and Beyond

The GPU hardware generation cycle runs approximately two years. Blackwell is current. NVIDIA's Vera Rubin platform extends that performance curve further. The NVIDIA Vera Rubin DSX AI Factory reference design confirms the forward trajectory: facilities must be designed with thermal and power headroom beyond current Blackwell requirements to accommodate Rubin and subsequent generations without facility reconstruction.

A practical planning rule: design your cooling plant for 150% of current Blackwell load. Design your power distribution for 200% of current load. Reserve space, power, and pipe paths for one more generation than the current build needs. The incremental capital cost of this headroom is small compared to the cost of facility reconstruction when the next GPU generation arrives.

14.2 Supply Chain Risk and Vendor Diversification

AI factory supply chains are concentrated. A small number of GPU manufacturers, advanced cooling system providers, and high-radix switch vendors supply the components for essentially all Blackwell deployments.

Supply chain disruptions at any of these vendors create deployment delays that are difficult to mitigate in the short term.

Mitigation strategies include maintaining supplier relationships with multiple cooling system vendors, pre-ordering infrastructure components with long lead times well ahead of planned GPU deployment dates, and working with ecosystem partners who carry inventory for rapid deployment. Reference designs help because they reduce dependence on one-off parts and custom engineering paths.

Procurement teams should also track lead time separately for each major component category rather than treating the project as a single critical path. GPU allocation, CDU manufacturing, switchgear fabrication, and transformer procurement each run on independent schedules with different bottlenecks. A project plan that assumes all four arrive on a synchronized timeline is making an optimistic assumption that rarely survives contact with real supply chains. Building schedule buffer around the longest-lead item, and sequencing site readiness work so it does not block on the shortest-lead item, keeps the overall project timeline realistic.

14.3 Cyber-Physical Security for AI Factories

AI factories present a new class of cyber-physical security challenge. DSX Exchange integration between IT and OT creates network paths between compute systems and facility control systems. A compromised cooling plant control system can cause GPU throttling or thermal damage. Security architects must apply zero-trust principles to DSX Exchange integrations: authenticate every service-to-service connection, segment IT and OT networks with explicit allow-listed communication paths, and audit all facility system access in real time.

The threat model should include both external attackers and internal mistakes. A faulty policy, a bad firmware update, or a misrouted control command can hurt the same plant. Strong authentication, network segmentation, change control, and rollback procedures belong in the base design. CISA guidance on critical infrastructure security applies directly to AI factory cyber-physical environments.

14.4 Business Continuity and Multi-Site Resilience

Business continuity should include workload portability. If one site loses utility service or cooling headroom, another site should absorb key workloads where possible. Uptime Institute's resilience framework recommends the AI factory portfolio become more resilient when capacity and data flow across multiple sites rather than depending on one campus alone.

Resilience also includes operational knowledge. A site that documents startup, shutdown, failover, and maintenance routines reduces the risk from knowledge loss when staff changes. That is especially important when the site uses liquid cooling and AI-assisted control, the process must survive turnover, not only hardware refresh.

Section 15: Financial and Strategic Considerations

15.1 The Board-Level Financial Case

The financial case for AI factories rests on three variables: token yield per watt, speed to production, and competitive positioning. Blackwell improves all three when the design aligns with the workload. A practical Blackwell-era financial model uses three layers.

- **Layer 1 - Direct utility cost**: The easiest to measure. Data centers spend $1.9–$2.8 million per megawatt per year on energy. Blackwell's 30x energy efficiency improvement against Hopper directly reduces this figure per token produced.
- **Layer 2 - Capex premium for liquid cooling and control systems**: Liquid cooling infrastructure adds $500K–$2M per MW. This is offset by the utility savings achieved over a 5-to-7-year depreciation cycle.
- **Layer 3 - Revenue uplift from more tokens per megawatt and faster deployment:** This is the largest lever and the hardest to model without site-specific workload data. Conservative planning uses 10x throughput improvement; base cases use 20–30x.

A month of lost time to market can cost far more than a small capex premium. In fast-moving AI markets, the value of earlier production often exceeds the value of a slightly lower build cost. The CFO should compare Blackwell not only against Hopper but also against delay.

15.2 Campus-Scale Financial Model

Campus Size	IT Load (80%)	Blackwell Racks	Annual Power at $0.08/kWh	Planning Note
100 MW	80 MW	~563 racks	~$70M	80% IT share, 142 kW per rack
300 MW	240 MW	~1,690 racks	~$210M	80% IT share, 142 kW per rack
1,000 MW (1 GW)	800 MW	~5,634 racks	~$701M	80% IT share, 142 kW per rack

Figure 3: Campus economics table. Utility only. Excludes compute capex, land, tax, labor, and financing.

15.3 Token Economics Sensitivity

Scenario	Hopper Baseline	Blackwell Planning Case	Board-Level Implication
Throughput	1.0x	10x to 30x	Output density rises sharply
Energy efficiency	1.0x	10x to 25x	Utility cost per token falls
Water efficiency	1.0x	~300x	Cooling footprint drops
Token cost	1.0x	0.05x to 0.15x	Unit economics improve substantially

Figure 4: Token economics sensitivity. Use local workload measurements before final investment approval.

Financial models should include a downside case. If utilization lags or tariff rates rise, the board needs to know the floor, not only the upside. Sensitivity work should test four shocks: electricity price shock,

utilization shock, model mix shock, and delay shock. A Blackwell factory usually survives one shock well. The danger comes from two or more simultaneously.

> **The cleanest board KPI is cost per million tokens. That metric sits at the intersection of model choice, hardware choice, power price, and utilization. It allows finance and engineering to speak the same language.**

Strategic positioning matters because infrastructure becomes product differentiation. A company with a stronger AI factory can launch new services sooner, support more users, and train better models with less friction. That strength compounds across product lines and markets.

The strongest investment thesis links Blackwell to revenue speed, not only to utility savings. Better inference, better model service, and faster iteration all feed the top line. When the board sees the full chain from infrastructure to revenue, the capex case becomes easier to defend than when the conversation stays limited to utility cost comparisons.

Section 16: Case Study Vignettes

16.1 Greenfield Gigawatt Campus, Digital Realty Manassas

A national cloud and AI services provider secures power in a market with strong transmission access and builds a multi-hall campus around DSX simulation. The design uses a standardized rack class, modular substations, direct liquid cooling, and a common control fabric across halls. The NVIDIA AI Factory Research Center at Digital Realty's Manassas, Virginia site is the real-world validation of this approach: the build-out is captured as a scalable, repeatable recipe using the Omniverse DSX Blueprint.

Executives gain capacity scale and a clear token economics story. Engineers gain repeatable build blocks validated in simulation. Operators gain one maintenance pattern across the campus. The greenfield approach proves scale and establishes the template for subsequent halls.

16.2 Brownfield H100 to Blackwell Retrofit

A colocation operator already running H100-class liquid cooling upgrades the most valuable halls first. The project replaces select CDUs, expands pipe diameter, modernizes switchgear, and adjusts control logic for Blackwell thermal density, following Nautilus Data Technologies' documented upgrade protocol.

Executives gain faster time to market because the building shell already exists. Engineers gain a lower-risk migration path. Operators gain practical experience with the same coolant and control model they will need for later refresh cycles. The brownfield approach proves reuse and establishes operational muscle memory for the Blackwell transition.

16.3 Modular Phased Deployment

A regional enterprise with a volatile demand profile chooses prefabricated AI blocks rather than committing to a full campus. The team deploys one pod, validates the token economics, then adds capacity in phases. The architecture uses DSX Sim before each stage and DSX Flex to react to utility conditions.

Executives gain capex discipline, buying only the next block of capacity, then measuring the payoff before the next block ships. Engineers gain a controlled expansion path. Operators gain a stable service model with fewer surprises. The modular approach proves discipline and keeps capital aligned with product demand.

16.4 Sovereign AI Factory

A public sector consortium builds a national AI infrastructure platform for sensitive workloads. The design favors strong data sovereignty, controlled operations, and a predictable power contract. The factory still uses Blackwell-class density and liquid cooling, but governance, isolation, and auditability sit at the center of the design.

The executive goal is strategic autonomy. The engineering goal is repeatability within strict security boundaries. The operations goal is trustworthy service under public accountability. This case proves control and establishes the template for regulated and government AI factory deployments.

PATTERN ACROSS ALL FOUR CASES:

The executive team wins when time to value shortens. The engineering team wins when the design repeats cleanly. The operations team wins when the facility has one operating model rather than a patchwork of exceptions.

Section 17: Implementation Roadmap

17.1 Blackwell AI Factory Readiness Checklist

Use this checklist to assess your organization's readiness. Score each item: 1 (not started), 2 (in progress), 3 (complete). Areas scoring below 80% are likely deployment blockers.

Domain	Readiness Item	Score
Site & Power	Confirmed grid interconnect capacity for planned Blackwell load	1/2/3
Site & Power	Substation capacity and timeline assessed	1/2/3
Site & Power	800 VDC power path evaluated for new builds	1/2/3
Site & Power	Power distribution topology designed for 140+ kW per rack	1/2/3
Thermal	Liquid cooling infrastructure specified or contracted	1/2/3
Thermal	CDU capacity per rack or pod validated for GB300 load (142 kW)	1/2/3
Thermal	Pipe diameters and flow rates specified for Blackwell TDP	1/2/3
Thermal	Thermal telemetry integrated with DSX OS dashboard	1/2/3
Infrastructure	Structural floor loading confirmed for NVL72 rack weight	1/2/3

Infrastructure	Network fabric (IB or Spectrum-X) topology designed	1/2/3
Infrastructure	Storage architecture sized for checkpoint and ingestion loads	1/2/3
DSX Platform	DSX Sim digital twin created and facility layout validated	1/2/3
DSX Platform	DSX Exchange IT/OT integration points specified	1/2/3
DSX Platform	DSX OS and MaxLPS operational runbooks defined	1/2/3
DSX Platform	DSX Flex demand response configured with utility partner	1/2/3
MEP	Leak detection and isolation valve plan complete	1/2/3
MEP	BMS integration with DSX Exchange defined	1/2/3
Operations	Model lifecycle governance and retraining pipeline in place	1/2/3
Operations	Token-per-watt dashboard and alerting thresholds defined	1/2/3
Operations	Staged rollout plan with monitoring thresholds defined	1/2/3
Risk	Brownfield audit of H100 infrastructure complete (if applicable)	1/2/3
Risk	Cyber-physical security architecture reviewed	1/2/3
Risk	Multi-generation capacity headroom (150%+) planned	1/2/3

17.2 Sequencing: From Simulation to Commissioning

1. Site and power assessment (Months 1–3). Confirm power availability, site capacity, grid interconnect timeline, and 800 VDC path viability.
2. DSX Sim digital twin design (Months 2–5). Model facility layout, thermal performance, and network architecture before committing to construction.
3. Infrastructure procurement (Months 3–8). Order long-lead items: CDUs, high-capacity cooling plant, power distribution equipment, and GPU systems.
4. MEP design and BMS integration (Months 4–8). Complete mechanical, electrical, and plumbing blueprint. Define BMS-to-DSX integration points.
5. Facility construction and mechanical installation (Months 5–14). Build or retrofit the physical facility, install cooling plant, power distribution, and structural systems.
6. Compute and network deployment (Months 10–16). Install NVL72 racks, connect cooling and power, deploy networking fabric.
7. DSX OS commissioning and staged rollout (Months 14–18). Commission DSX Exchange and DSX OS, run load tests, then roll out AI workloads in stages while monitoring thermal and power performance.
8. Go-live validation (Month 18+). Verify that the plant handles real load, real user traffic, and real failure recovery before naming the site production-ready.

17.3 Aligning Infrastructure, Model, and Product Teams

AI factory projects fail when infrastructure, model development, and product teams operate on separate timelines. Resolve this with a shared production contract before any capital is committed: infrastructure teams commit to specific power, cooling, and network capacity at defined points in time. Model teams commit to workload profiles. Product teams commit to launch timelines that align with facility readiness milestones.

The KPIs that align all three teams are tokens per second, tokens per watt, and cost per token. Every architectural decision, from CDU sizing to network topology to model quantization, should be evaluated against its impact on those three numbers. The business does not win because the facility has a lower PUE on a slide. The business wins because the facility produces more useful output with less friction.

Blackwell-era infrastructure is not a short-term upgrade cycle. It is the opening step in a new class of economic infrastructure. The firms that win will treat AI factories as long-lived production assets with token yield, utility strategy, and operational control at the center. The firms that lose will keep measuring the wrong thing: they will optimize the room while missing the factory, optimize the rack while missing the network, optimize the model while missing the plant. The market now rewards the teams that build and operate the whole system as one coherent production line, from grid interconnect to inference output, rather than the teams that excel at any single layer in isolation.

> **The firms that win will plan the factory before they plan the rack order. They will validate the site before they pour concrete. They will design for tokens per watt, not only for watts per rack.**

Appendix A: Financial Model Assumptions and Campus Economics

The following model uses conservative planning assumptions for board review. Total campus power refers to the utility envelope, not only the IT load. IT load uses 80% of the campus envelope. Blackwell rack density uses 142 kW per rack. Electricity sensitivity uses three rates so the finance team can map local tariff reality onto the model.

Campus	IT Load	Blackwell Racks	Power at $0.04/kWh	Power at $0.08/kWh	Power at $0.14/kWh
100 MW	80 MW	~563	$35.0M	$70.1M	$122.6M
300 MW	240 MW	~1,690	$105.1M	$210.2M	$367.9M
1 GW	800 MW	~5,634	$350.4M	$700.8M	$1,226M

Figure 5: Campus financial model. Utility cost only. Excludes compute capex, land, tax, labor, and financing.

A deeper model should track three costs beyond utility: the liquid cooling premium, the higher-end electrical distribution premium, and the labor and maintenance premium tied to specialized systems. Those costs rise in Blackwell builds, but each one should be compared against the revenue value of higher token output and the avoided cost of delay.

Appendix B: Token Economics Sensitivity

Blackwell changes token economics in two ways: it raises output per rack and per megawatt, and it lowers cooling overhead relative to the workload envelope. A sensible finance model tests 10x, 15x, and 20x token output improvement cases rather than assuming full marketing claim performance.

Scenario	Hopper Baseline	Blackwell Planning Case	Board-Level Implication
Throughput per MW	1.0x	10x to 30x	Output density rises sharply
Energy efficiency	1.0x	10x to 25x	Utility cost per token falls significantly
Water efficiency	1.0x	~300x (NVIDIA claim)	Cooling footprint drops dramatically
Token cost (normalized)	1.0	0.05 to 0.15	Unit economics improve 7–20x
Revenue potential per rack	Baseline	40–50x	Same footprint produces more value

Figure 6: Token economics sensitivity analysis. Conservative and base cases for board-level modeling.

Sensitivity work should test four shocks: electricity price shock, utilization shock, model mix shock, and delay shock. The cleanest board KPI is cost per million tokens, that metric sits at the intersection of model choice, hardware choice, power price, and utilization. When executives and engineers both use that number, infrastructure decisions and product decisions finally speak the same language.

Glossary of Key Terms

This glossary defines the core vocabulary used throughout this whitepaper. Teams spanning facilities, network engineering, finance, and AI platform roles benefit from a shared definition set before reviewing the technical sections in depth.

AI Factory

A purpose-built facility that converts electricity into AI inference output (tokens) as its primary product, measured by tokens per second, tokens per watt, and cost per token rather than legacy data center metrics like PUE alone.

Tokens Per Watt

The number of inference tokens an AI factory produces per watt of total facility power consumed. The primary efficiency metric for Blackwell-era infrastructure, replacing PUE as the dominant design target.

Cost Per Token

The fully-loaded cost (capex amortization plus opex) to produce one unit of AI inference output. The commercial metric that ties infrastructure design directly to AI product economics.

NVL72

NVIDIA's rack-scale architecture connecting 72 GPUs (and in GB200/GB300 configurations, 36 Grace CPUs) into a single NVLink domain with unified memory access, functioning as one logical compute unit rather than 72 independent servers.

TDP (Thermal Design Power)

The maximum sustained heat a chip is expected to generate under normal operating load, used to size cooling infrastructure. Blackwell GPUs reach 1,200W TDP in liquid-cooled configurations, compared to 700W for H100.

CDU (Coolant Distribution Unit)

The mechanical unit that transfers heat from the chip-level technology cooling loop to the facility cooling loop via a liquid-to-liquid heat exchanger, while regulating flow rate, pressure, and supply temperature.

Delta-T (ΔT)

The temperature difference between coolant entering and leaving a cooling loop. A key design and monitoring parameter for liquid cooling systems; typical Blackwell designs target a delta-T of 10 to 15 degrees Celsius.

DSX

NVIDIA's full-stack platform defining how AI factories are designed (DSX Sim), operated (DSX OS), power-optimized (DSX MaxLPS), grid-integrated (DSX Flex), and unified across IT/OT systems (DSX Exchange).

Digital Twin

A simulated, software-based replica of a physical facility used to validate power, cooling, network, and structural design decisions before construction begins. DSX Sim is built on NVIDIA Omniverse digital twin technology.

Max-Q

An operating mode that targets the peak-efficiency point on the tokens-per-watt curve rather than maximum power draw, trading a small amount of throughput headroom for significantly improved energy and thermal stability.

MVDC (Medium-Voltage Direct Current)

A power distribution architecture that delivers medium-voltage DC power closer to the rack, reducing the number of AC-DC conversion stages and associated losses compared to traditional AC distribution chains.

800 VDC

An emerging power distribution standard for AI factories operating at megawatt-class rack densities, replacing legacy 54V in-rack distribution to reduce copper weight, conversion losses, and distribution complexity.

Brownfield vs. Greenfield Deployment

Brownfield refers to retrofitting or upgrading an existing facility (commonly an H100 site) for Blackwell-class density. Greenfield refers to building an entirely new facility, optimized from the ground up for Blackwell or later generations.

Fat-Tree Topology

A network topology providing full bisection bandwidth between all leaf switches, commonly used in InfiniBand-based AI training clusters to support all-reduce communication patterns with minimal congestion.

All-Reduce

A collective communication pattern in distributed AI training where every GPU exchanges gradient information with every other GPU simultaneously, demanding extremely high bisection bandwidth and low network latency.

PUE (Power Usage Effectiveness)

The ratio of total facility power to IT equipment power, used historically to measure data center overhead efficiency. PUE remains a useful overhead metric but does not capture AI output, which is why tokens per watt has become the primary efficiency benchmark for AI factories.

Time to Day One

The elapsed time between production approval and the first day an AI factory is live and serving production workloads. A key planning metric distinct from construction completion, since commissioning, staged rollout, and load validation all occur after the physical build is finished.

A Note on Methodology

This whitepaper synthesizes publicly available technical material from NVIDIA, Schneider Electric, Nautilus Data Technologies, and supporting industry, research, and standards bodies including IEA, EPRI, NIST, CISA, ASHRAE, and Uptime Institute. Figures attributed to NVIDIA reflect NVIDIA's published performance claims for Blackwell-class systems relative to the Hopper generation; readers should validate workload-specific figures against their own benchmarks before finalizing investment decisions. Financial models in the appendices use illustrative assumptions for board-level directional planning and exclude site-specific variables such as local tariffs, land costs, labor rates, and financing terms.

Frequently Asked Questions

What is an AI factory and how does it differ from a traditional data center?

An AI factory is a purpose-built facility that manufactures intelligence tokens continuously at scale. Its performance is measured in tokens per second, tokens per watt, and cost per token, not in storage IOPS or server uptime. The AI factory model co-designs energy, chips, networking, storage, software, and facility systems from the start rather than treating them as separate procurement events.

Why is liquid cooling required for Blackwell deployments?

The GB200 and GB300 NVL72 systems consume up to 142 kW per rack, with GPUs operating at 1,200W TDP. Air cooling has a practical limit of roughly 40 kW per rack. At Blackwell density, direct-to-chip liquid cooling is physically required, not an option.

What is NVIDIA DSX and which components does it include?

NVIDIA DSX is the platform defining how AI factories are designed, built, and operated. It includes: DSX Sim (Omniverse-based digital twin simulation), DSX OS (operational software for cluster and facility management), DSX MaxLPS (tokens-per-watt optimization), DSX Flex (grid demand response), and DSX Exchange (IT/OT integration layer).

What specific infrastructure upgrades does an H100 site need for Blackwell?

Nautilus Data Technologies documents the required upgrades: upgraded pump and flow rate capacity, larger diameter primary coolant piping, CDUs rated for 140+ kW per rack (vs 60–80 kW for H100), expanded facility cooling plant capacity, and in many cases structural floor reinforcement. A full thermal headroom audit of the existing H100 infrastructure is the required first step.

What are the Schneider Electric RD110 and RD111 reference designs?

Schneider Electric's RD110 and RD111 reference designs are co-developed with NVIDIA and specify complete power and cooling architectures for Grace Blackwell GB300 NVL72 platforms at up to 142 kW per rack. They define CDU placement, pipe specifications, manifold configurations, and integration with Schneider EcoStruxure and NVIDIA DSX Exchange.

What is the 800 VDC power path and why does it matter?

NVIDIA and partners are moving AI factory power architecture toward 800 VDC distribution to address the inefficiency of traditional 54V in-rack systems at megawatt-class rack densities. EPRI research supports this direction. 800 VDC reduces copper weight, conversion stages, and distribution complexity while improving efficiency as rack density increases.

How does Blackwell compare to Hopper in energy and water efficiency?

The GB200 NVL72 delivers 25x more energy efficiency and 300x more water efficiency than legacy air-cooled architectures, according to NVIDIA's Chill Factor analysis. The GB300 NVL72 extends this to 30x energy efficiency improvement and 50x higher throughput per megawatt. At $0.08/kWh, a 50 MW hyperscale site saves over $4 million annually in cooling-related energy and water costs.

What is the difference between InfiniBand and Spectrum-X for an AI factory build?

Both are NVIDIA-engineered fabrics built for AI, but they serve different priorities. InfiniBand delivers the lowest latency and highest bandwidth for tightly-coupled training where all-reduce dominates performance. Spectrum-X Ethernet gives Ethernet-native enterprises AI-grade performance with lower operational learning curve. Large training-focused gigawatt campuses typically default to InfiniBand; inference-heavy or Ethernet-native enterprise deployments increasingly choose Spectrum-X.

How should a board model the financial case for Blackwell versus delaying the investment?

The financial case should weigh three layers together: direct utility cost (the easiest to measure), the capex premium for liquid cooling and stronger distribution, and the revenue uplift from higher tokens per megawatt and faster time to production. A month of lost time to market in a fast-moving AI market often costs more than a modest capex premium. Boards should model a downside case as well as the upside, testing electricity price shocks, utilization shocks, model mix shocks, and delay shocks independently, since a Blackwell factory typically tolerates one shock well but is more vulnerable when two or more occur simultaneously.

What is the practical difference between DSX Sim and traditional CFD modeling?

Traditional computational fluid dynamics (CFD) modeling typically validates thermal behavior for a single rack or row in isolation. DSX Sim extends that concept to the entire facility as one connected digital twin, simultaneously modeling power distribution, cooling plant behavior, network cabling, and structural layout together. Ecosystem partners such as Cadence, PTC, Siemens, and Dassault Systemes integrate their

facility design tools directly into the Omniverse DSX Blueprint, so power, thermal, and layout decisions are validated as one system rather than as separate engineering exercises run in sequence.

Sources Referenced

NVIDIA Blog: AI Factories, The New Infrastructure of Intelligence (May 2026)

NVIDIA Solutions: AI Factories Architecture Overview

NVIDIA AI Factory Product Page

NVIDIA DSX Platform: AI Factory Design, Simulation, and Operations

NVIDIA Newsroom: DSX Gives Infrastructure Builders the Playbook for AI Factories

NVIDIA Blog: Blackwell Platform Water Efficiency (April 2025)

NVIDIA Blog: Omniverse DSX Blueprint, Gigawatt-Scale AI Factories

NVIDIA Press Release: Vera Rubin DSX AI Factory Reference Design (March 2026)

NVIDIA GB200 NVL72 Product Page

NVIDIA GB300 NVL72 Product Page

NVIDIA H100 Product Page

NVIDIA Blackwell Architecture Overview

NVIDIA Vera Rubin Architecture

NVIDIA NVLink Switch System

NVIDIA InfiniBand Networking

NVIDIA Spectrum-X Ethernet

NVIDIA Liquid Cooling

NVIDIA Grace Blackwell Superchip

Schneider Electric Blog: Liquid Cooling Reference Designs for AI Data Centers (January 2026)

Nautilus Data Technologies: What H100 Deployments Are Teaching Us About the Blackwell Era (April 2026)

Digital Realty: NVIDIA AI Factory Research Center, Manassas VA

MarketsandMarkets: Data Center Liquid Cooling Market ($21.3B by 2030)

IEA: Energy and AI Report

EPRI: Data Center Power and 800 VDC Research

NREL: Renewable Energy and Data Centers

NIST Cybersecurity Framework

CISA: Critical Infrastructure Security

Uptime Institute: Data Center Resilience and Sustainability

ASHRAE: Data Center Cooling Standards

Phaidra: AI Agents for Data Center Operations

Vertiv: AI Factory Infrastructure Solutions

Bechtel: AI Campus Construction and Engineering

WEKA: Parallel Storage for AI Workloads

DDN: AI Storage and Checkpoint Solutions

NVIDIA Omniverse Developer Center