THE 600kW BLUEPRINT

Powering the Next Generation of Hyperscale AI Infrastructure

NEXTDC

where AI lives™

Your blueprint to building AI infrastructure for 600kW+ workloads and beyond

Artificial intelligence (AI) is transforming how we design and operate digital infrastructure. Traditional data centres are evolving into AI Factories, large, software-driven systems built to power model training, inference, and intelligent services at massive scale.

NVIDIA CEO **Jensen Huang** puts it really well:

> *“AI is now infrastructure, and this infrastructure, just like the internet, just like electricity, needs factories... They’re not data centres of the past... They are, in fact, AI factories. You apply energy to it, and it produces something incredibly valuable... called tokens.”*

These environments are designed not just to run AI, but to produce it.

They convert vast streams of power and data into machine learning models, real-time predictions, and token-based intelligence, fueling everything from enterprise search and fraud detection to generative content and autonomous systems. Unlike traditional compute environments, these **AI Factories** are purpose-built to support:

Super dense power delivery

They pack a lot of power into a small space.

Advanced cooling systems

State-of-the-art thermal solutions such as direct-to-chip liquid cooling and immersion cooling, purpose-built for heat-intensive AI training.

Specialised networking

Tailored network fabrics that support GPU-to-GPU communication with ultra-low latency and high bandwidth key for model performance.

AI Factories mark a shift in infrastructure thinking, purpose-built to run massive-scale GPU clusters, consuming 600kW or more per rack. They go beyond traditional data centre design, requiring tightly integrated systems for power, cooling, networking, and performance optimisation, including:

Ultra-fast networking

Programmable, low-latency infrastructure, designed for real-time inference and distributed training.

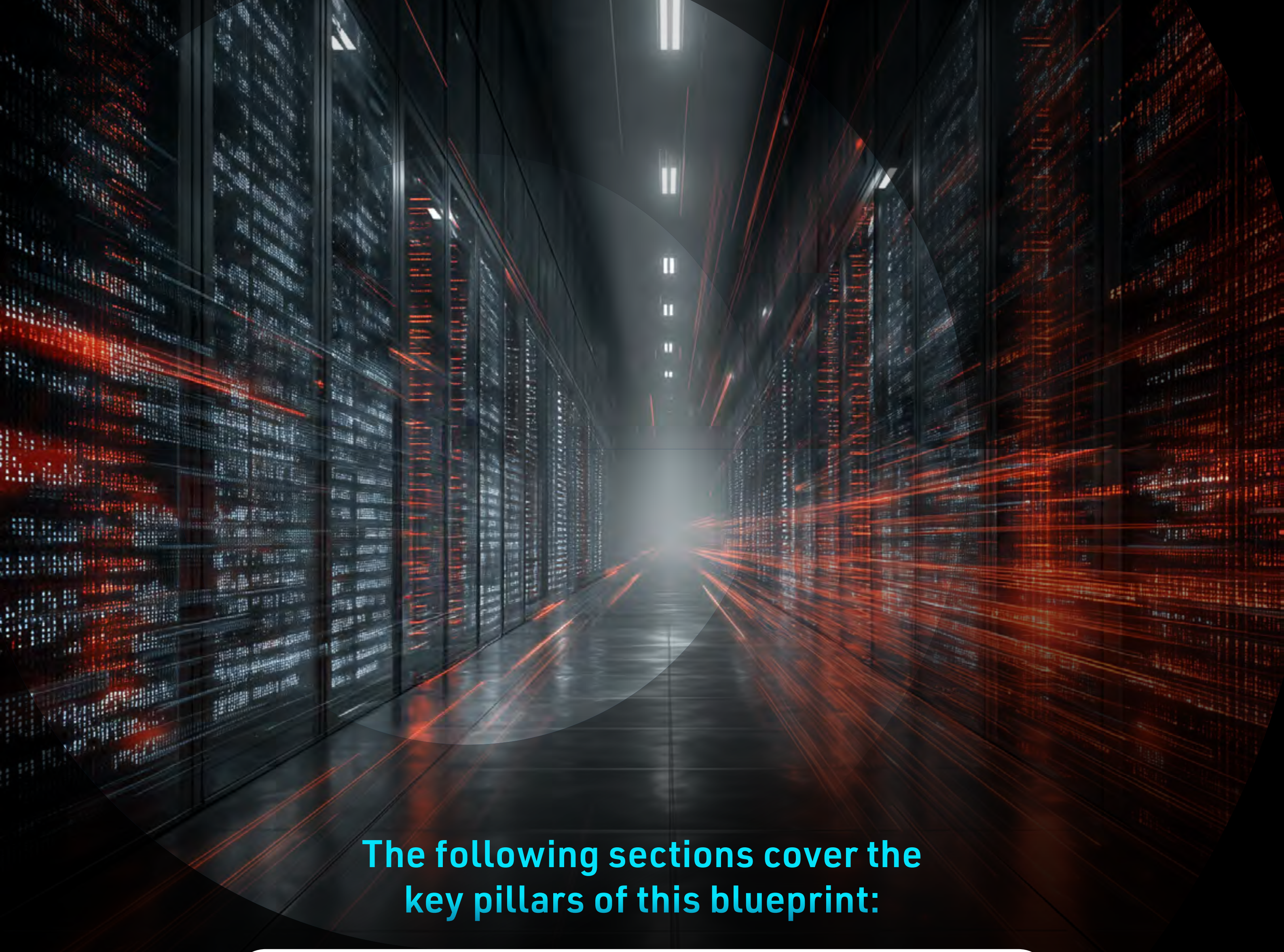
AI-optimised layouts

The physical arrangement is designed perfectly for AI workloads.

Proximity to data and users

Strategically located near data sources and end users to minimise latency, meet data sovereignty requirements, and improve user experience.

As AI transforms industries and economies, building the physical foundation for this change isn't just an option anymore; it's absolutely crucial. This plan explains what's involved in engineering these next-generation, hyperscale AI factories from the ground up.



The following sections cover the key pillars of this blueprint:

Why traditional data centres can't meet **AI's demands**

The computing needs of modern AI, from training large-scale models to generating real-time token-based outputs — aren't just more demanding; they are fundamentally different and accelerating at an extraordinary pace. Traditional data centres, designed for a previous era of general-purpose computing, cannot scale to support the specialised requirements of AI Factories: dense compute, intelligent orchestration, and low-latency inference at global scale.

Power and thermal density

AI accelerators consume significantly more power and generate far more heat than traditional servers. Conventional air-cooling systems and legacy power delivery infrastructure are insufficient to manage racks running at 50kW, 100kW, and increasingly, up to 1MW per rack. This intense energy consumption presents major thermal challenges and places immense pressure on power grids and sustainability targets.

Hyperscalers must prioritise sites with access to substantial, reliable, and ideally renewable energy sources, while implementing advanced cooling systems such as direct-to-chip or immersion liquid cooling. Moreover, the physical form factor of AI clusters introduces additional design constraints, including heavier racks, liquid cooling manifold integration, and precise airflow zoning. These infrastructure requirements are difficult—and often impossible—to retrofit into traditional data centre designs not purpose-built for AI.

Interconnectivity requirements

Training and inference at scale demand ultra-high-speed, low-latency communication between accelerators and nodes. AI Factories use high-bandwidth topologies such as NVLink, NVSwitch, and RoCE fabrics, far exceeding what standard Ethernet networks in enterprise data centres can support. These interconnects are vital for distributing workloads and avoiding I/O bottlenecks during model training and real-time inference.

Operational complexity

AI infrastructure requires a different operational paradigm. Instead of managing virtual machines, operators must orchestrate tightly coupled accelerators, monitor thermal telemetry in real time, and maintain high availability for extremely dense compute clusters. These specialised systems require automation, cooling intelligence, and power management strategies far beyond those in traditional IT stacks.

Extreme computational intensity

AI workloads especially during model training require quadrillions of calculations and benefit from massively parallel processing. Specialised hardware like Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) are significantly more efficient than general-purpose CPUs for these tasks, but they come with steep power, cooling, and orchestration demands.

Phased workload demands

Pre-training

The initial training of large-scale AI models on massive datasets is extraordinarily compute-intensive. It requires tightly integrated clusters of accelerators, advanced interconnects, and sustained energy and cooling, often over days or weeks to process trillions of parameters.

Fine-tuning

Once a foundation model is trained, fine-tuning adapts it to specific tasks, domains, or data sources. While less demanding than pre-training, it still requires considerable compute, particularly in high-performance sectors like medicine, law, and technical AI.

Inference (Test-Time Scaling)

Deploying trained models in production to deliver real-time outputs requires low-latency, high-throughput infrastructure. Inference must scale to handle millions of concurrent requests for latency-sensitive use cases such as autonomous vehicles, conversational AI, recommendation systems, and live language translation.

Conclusion

Traditional data centres fall short not only in capacity, but in capability. They were not built to sustain the thermal loads, rack densities, or high-throughput communication demands of AI workloads. Meeting these challenges requires a wholesale reinvention of infrastructure, one designed from the ground up for AI's computational intensity, energy requirements, and architectural complexity. The AI Factory emerges from this need: a purpose-built environment reimagined from the rack to the grid.

AI token throughput: the new measure of success

In the realm of AI Factories, traditional metrics such as storage capacity or standalone network bandwidth are insufficient to gauge performance. A more pertinent metric has emerged: AI token throughput, the rate at which an AI system generates output tokens during inference. This metric encapsulates the system's ability to deliver real-time predictions and content generation, serving as a direct indicator of its intelligence production capacity.

What is a token?

In AI, particularly with Large Language Models (LLMs) like ChatGPT, a token is a fundamental unit of text or code that the model processes. It's often a word, but it can also be a part of a word, a punctuation mark, or even a space. For example, if you make a request to ChatGPT, such as "Tell me a story about a brave knight," the words "Tell," "me," "a," "story," "about," "a," "brave," and "knight" would likely each be treated as individual tokens. However, tokenisation isn't always one-to-one with words; a word like "running" might be broken into "run" and "##ning" (two tokens), or a common phrase might be represented by a single token.

To provide a comprehensive view of AI system performance, token throughput is often considered alongside other key indicators:

Time to First Token (TTFT)

Measures the latency between input submission and the generation of the first output token, crucial for responsiveness.

Tokens Per Second (TPS)

Indicates the rate of token generation, reflecting the system's overall throughput and processing speed.

Time Per Output Token (TPOT)

Represents the average time taken to generate each token, impacting the user experience.

Goodput

Focuses on the volume of useful output delivered within acceptable latency thresholds, ensuring quality and efficiency.

Elevated token throughput directly correlates with an AI Factory's capacity to handle extensive, concurrent inference workloads efficiently. Achieving this necessitates optimised hardware configurations, such as high-performance GPUs or TPUs, and advanced software strategies, including model parallelism and efficient batching techniques.

While factors like energy efficiency, cost management, and scalability are crucial, AI token throughput stands out as the definitive measure of an AI Factory's effectiveness. It encapsulates the facility's core mission: transforming data into actionable intelligence at scale, thereby driving innovation and competitive advantage across industries.

New priorities for data centres: **the 5S of AI Factories**

The rise of AI Factories is making IT and data centre leaders completely rethink their priorities. In the past, data centres were planned around things like overall size (square metres, total power) and how cheap they were to run. Today, five key priorities, which we call the 5S, have become crucial for the large cloud providers (hyperscalers) building and running AI infrastructure:

- Speed ✓
- Scale ✓
- Sovereignty ✓
- Sustainability ✓
- Security ✓

Speed

In the world of AI Factories, 'speed' means several things. First, it's about time-to-value – how quickly can you train a new AI model or add more capacity when demand suddenly increases? Hyperscalers now compete on how fast they can set up new GPU clusters or launch AI services. Cloud-native AI platforms focus on quick setup and minimal hassle; for example, offering GPU capacity by the hour, ready with AI frameworks, so development teams can innovate rapidly. Executives need to ensure their infrastructure (and partners) can deploy at "hyperspeed" both in getting hardware ready and moving data quickly. High-performance connections (low-delay networks, locations close to users) are also vital, as model training and AI predictions happen in real-time. Simply put, if your AI Factory can't keep up with the speed of experimentation and user demand, innovation will move elsewhere.

Scale

AI workloads that used to run on a few servers now need thousands of GPUs working at the same time. 'Scale' isn't just about having big data centres; it's about smoothly expanding within and across different facilities. Hyperscale AI Factories must support huge amounts of computing power (petaflops to exaflops), millions of simultaneous AI model queries, and training runs involving trillions of parameters. This requires designs that are modular and can be easily copied. For instance, NVIDIA's reference AI Factories are built from "pods" or blocks of GPUs that can be cloned and connected by the hundred. Cloud providers talk about "availability zones" dedicated to AI, and "AI regions" appearing where there's plenty of power. The goal is to expand AI computing almost like a utility, adding more AI Factory space with minimal disruption. Scale also means having a global presence: hyperscalers like AWS, Google, and Alibaba are expanding AI infrastructure to more regions to serve local needs while balancing workloads worldwide. If an AI service suddenly needs ten times more capacity — maybe because of a hit app or a breakthrough model, the infrastructure has to scale in days, not months. That's why NVIDIA gives its partners a five-year roadmap: building enough AI-ready power and space takes time. Leading data centre operators are already planning for expansions of over 100 megawatts (MW) to make sure growth never holds back innovation.

Sovereignty

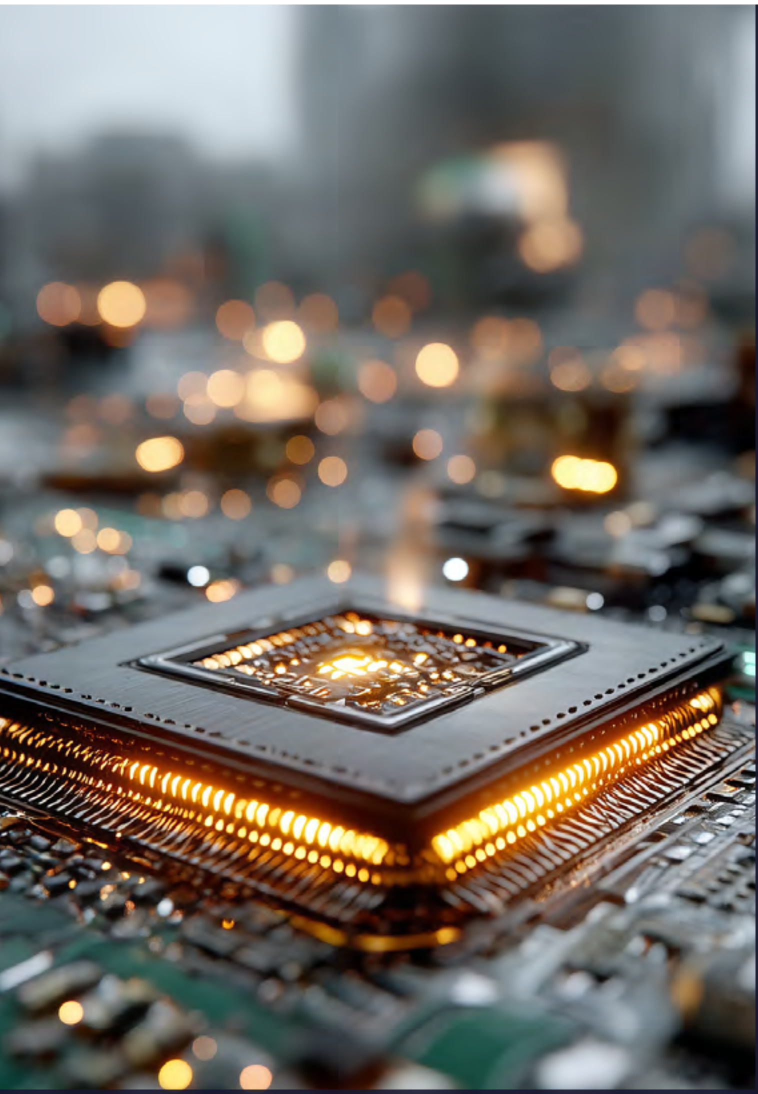
Data sovereignty and infrastructure sovereignty have become critical in the age of AI. As AI systems are used in sensitive areas, from healthcare diagnoses to national security, where data and models are stored, and under whose laws, is a major concern. Hyperscalers must navigate a complex set of regulations that increasingly demand certain data remains within national borders, or that AI workloads are processed in locally controlled facilities for privacy and strategic reasons. The recent push for "sovereign cloud" offerings in Europe and elsewhere reflects this trend. For AI Factories, sovereignty can mean choosing data centre locations to meet legal requirements and customer trust. It's no longer just about technical specifications, but also about geopolitical and compliance positioning. For example, European cloud users might prefer (or be required by law) to use AI infrastructure hosted in the EU by EU-based providers. In China, AI infrastructure must be locally hosted due to strict data laws. Even within countries, some government or enterprise workloads demand sovereign-certified facilities, those checked for handling classified data or critical infrastructure roles. Where your AI infrastructure lives isn't just a technical choice, it's a competitive one. Latency, compliance, and sustainability are all shaped by location. Leading data centre operators choose sites based on a strategic mix of low latency, data sovereignty, and energy resilience. In practice, this means hyperscalers are investing in regions they previously left to partners and partnering with local data centre specialists to ensure sovereign coverage. The AI Factory revolution won't be a one-size-fits-all global solution; it will be a network of regionally tailored hubs that balance global scale with local control.

Sustainability

The power-hungry nature of AI has put sustainability at the heart of the conversation. Company boards and governments are increasingly scrutinising the energy and carbon footprint of AI operations. A single large AI training run can use as much electricity as hundreds of homes; scaled across many runs, AI could significantly impact company and national energy goals. Hyperscalers are acutely aware that any perception of AI as "wasteful" or environmentally harmful could lead to regulatory or public backlash, not to mention the direct impact on energy costs. Therefore, the new mantra is "performance per watt" and designing for efficiency from scratch. Leading cloud data centres are committing to 100% renewable energy (through solar, wind, hydro, or even emerging nuclear partnerships) to power AI Factories. They're also adopting advanced cooling to reduce waste; for example, liquid cooling can drastically cut cooling power overhead and even allow heat reuse, improving PUE (Power Usage Effectiveness) dramatically. Every aspect of facility design is under the microscope for sustainability, from using sustainable building materials to implementing circular economy principles for hardware (recycling and reusing components). Importantly, hyperscalers are now reporting metrics like "carbon per AI inference" or "energy per training run" as key performance indicators. The next generation of data centres will be judged not just on capacity, but on efficiency. Boards, regulators, and customers are asking harder questions: Where is the energy coming from? How efficient is your data centre? What are you doing to minimise environmental impact? The next generation of AI infrastructure won't be judged by performance alone, it will be evaluated by sustainability, transparency, and the ability to scale responsibly. To remain competitive (and compliant), AI Factories must be sustainable by design, aligning with global net-zero ambitions and corporate ESG commitments. Sustainability is no longer a nice-to-have Corporate Social Responsibility (CSR) item; it's a core design principle and differentiator in the AI era.

Security

With AI becoming a backbone for everything from financial services to autonomous vehicles, the security of AI infrastructure is paramount. Here we mean both cybersecurity and the physical security and resilience of infrastructure. On the cyber side, AI workloads often involve valuable training data (which could include personal data or proprietary information) and models that are intellectual property worth billions. Protecting these from breaches is critical; a compromised AI model or a disrupted AI service can cause immense damage. Hyperscale AI Factories are targets for attackers ranging from lone hackers to state-sponsored groups, all seeking to steal AI technology or sabotage services. This means investing in robust encryption (for data when it's stored and when it's moving), secure access controls, continuous monitoring powered by AI itself, and isolated compute environments (to prevent one client's AI environment from affecting another's in multi-tenant clouds). On the physical side, downtime is unacceptable; an AI Factory outage could halt operations for an organisation or even knock out critical infrastructure (imagine if an AI-driven power grid or hospital network fails). Therefore, AI data centres are built with extreme redundancy and hardened against threats. Many pursue Tier IV certifications for fault tolerance, and features like on-site backup power for days, multi-factor access controls, and even EMP or natural disaster protection in some cases. Additionally, supply chain security has emerged as a concern: ensuring that the chips and software powering AI are free from backdoors or vulnerabilities (which also links back to sovereignty). Security by design is a must. As one NEXTDC customer put it, their clients "rely on the ability to run AI-powered applications without interruption, for as long as they need," so having a partner that can guarantee uptime and flexibility is crucial. In practice, hyperscalers are choosing colocation providers and designs that emphasise robust risk management – from certified physical security controls to comprehensive compliance with standards (ISO 27001, SOC 2, etc.). In the AI Factory age, a security breach or prolonged outage isn't just an IT issue, it's a critical organisational risk. Therefore, security and resilience permeate every layer of the 5S model, underpinning speed, scale, sovereignty, and sustainability goals with a foundation of trust and reliability.



In summary, these 5S priorities are shaping decisions at the highest levels. Hyperscaler CIOs and CTOs are now asking:

Can our infrastructure deploy new AI capacity fast enough (**Speed**)?

Can it grow to the scale we'll need next year and five years from now (**Scale**)?

Do we have the right locations and partnerships to meet data jurisdiction and governance needs (**Sovereignty**)?

Are we minimising our environmental impact and energy risk even as we expand (**Sustainability**)?

And can we guarantee security and resilience end-to-end so that our AI services never falter (**Security**)?

The AI Factory era demands a holistic approach. Success will come from excelling across all five dimensions, rather than optimising for just one. In practice, this means designing data centre solutions that are agile and fast, massively scalable, locally available and compliant, green and efficient, and resilient by design and uncompromising on security. That's a tall order – but it's exactly what the leading innovators are now building.

The foundational components of an AI Factory

Building an AI Factory requires a holistic rethinking of digital infrastructure, focusing on highly specialised components:

The compute layer

GPUs and accelerators

These are the engines of the AI factory. Rack-scale architectures with dense, multi-GPU configurations are optimised for both training and inference.

AI-specific processors

Beyond GPUs, integrating AI-specific processors such as Google's Tensor Processing Units (TPUs), optimised for machine learning tasks, can offer significant performance and efficiency benefits.

Modular architectures

Incorporating modular systems like NVIDIA's MGX platform provides flexibility and scalability, allowing for tailored configurations that meet specific AI workload requirements.

Advanced liquid cooling

Given the extreme power consumption and heat generation of AI accelerators, sophisticated liquid cooling systems (e.g., direct-to-chip, immersion cooling) are essential for thermal management, allowing for higher density and sustained performance.

Reference architectures

Blueprints like NVIDIA DGX POD or SuperPOD provide validated designs for on-premises AI infrastructure deployment, streamlining the construction of dedicated AI factories.

Networking architecture

Advanced interconnects

Implementing high-speed interconnects like NVIDIA's NVLink and InfiniBand facilitates low-latency, high-bandwidth communication between compute nodes, essential for efficient AI model training and inference.

High-performance ethernet fabrics

Specialised Ethernet fabrics provide the backbone for large-scale data transfer and communication between compute nodes, ensuring high throughput and minimal latency.

Data Processing Units (DPUs)

These specialised processors offload networking, storage, and security tasks from GPUs, freeing up valuable compute resources to focus solely on AI workloads.

Software-Defined Networking (SDN)

Adopting SDN provides dynamic network management, allowing for optimised data flow and resource allocation tailored to AI workloads.

Storage architecture

Optimised for high-speed data ingestion

AI models require access to vast datasets for training and inference, demanding storage systems capable of ultra-fast data ingestion to prevent bottlenecks.

Tiered storage solutions

Implementing a tiered approach, combining high-speed NVMe storage for active datasets with scalable object storage for archival data, can optimise performance and cost-efficiency.

Distributed storage systems

Scalable and reliable distributed file systems or object storage are essential for managing immense volumes of AI data and models, facilitating efficient data sharing and serving.

Data versioning and lineage

Incorporating data versioning and lineage tracking ensures reproducibility and accountability in AI model development, facilitating better model management and compliance.

Data reuse and feedback loops

The architecture should support continuously feeding data generated by AI applications back into the system to refine and improve model performance, creating a virtuous cycle of intelligence.

Integration with enterprise storage

Seamless integration with existing enterprise data lakes and storage systems allows organisations to leverage their current data assets effectively.

Security and compliance

Integrated security frameworks

Embedding security at every layer of the AI Factory, from hardware to application, is crucial. Solutions like Cisco's Secure AI Factory with NVIDIA emphasise the importance of integrated security measures to protect data and AI models.

Compliance and governance

Establishing robust compliance and governance protocols ensures that AI operations adhere to regulatory standards and ethical guidelines, fostering trust and reliability.

Key characteristics **defining an AI Factory**

Beyond its components, an AI Factory possesses distinct characteristics that enable accelerated AI development:

Specialised hardware

Purpose-built with powerful accelerators (GPUs, TPUs) designed specifically for AI computations, drastically speeding up model training and inference.

Scalable and resilient infrastructure

Designed to handle massive datasets and complex AI models with elasticity, ensuring solutions can be developed and deployed rapidly and reliably.

Modular and composable infrastructure

Adopting a modular approach allows for flexibility and scalability, building infrastructure that can be easily reconfigured to accommodate different AI workloads, from training to inference. Components can be scaled independently, optimising resource utilisation and cost-efficiency.

Advanced software stack

Provides access to a comprehensive suite of tools, including machine learning libraries, MLOps platforms, data visualisation, and model deployment pipelines, streamlining the entire AI lifecycle.

Operational sophistication and MLOps

Beyond hardware, AI Factories are defined by their advanced operational frameworks. They integrate Machine Learning Operations (MLOps) platforms to automate and manage the entire AI lifecycle, from data ingestion and model training pipelines to deployment, monitoring, and continuous retraining. This demands highly automated processes, sophisticated telemetry, and real-time observability, alongside a unique blend of engineering talent skilled in both infrastructure and AI workflows.

Continuous learning and feedback loops

AI Factories are designed to support continuous learning cycles. This involves implementing systems that feed real-time data back into the AI models to refine and improve their performance over time and establishing automated processes for retraining models as new data becomes available, ensuring AI systems evolve with changing data patterns.

Sustainability and energy efficiency

Given the significant energy demands of AI workloads, sustainability is a key consideration. This involves selecting hardware components that offer high performance per watt to reduce overall energy consumption and implementing advanced cooling technologies, such as liquid cooling, to manage heat effectively and further improve energy efficiency.

Robust security and data governance

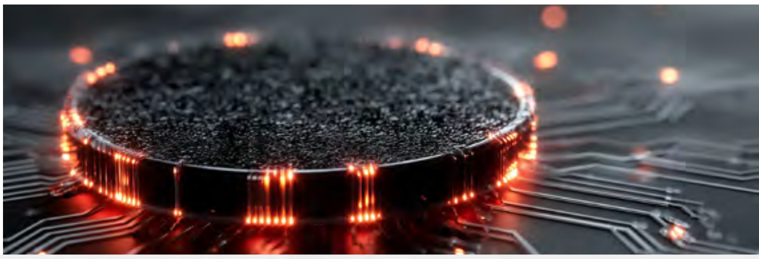
Ensuring the security and compliance of AI operations is paramount. Given that AI models are trained on and process vast, often sensitive datasets, AI Factories require exceptional security protocols and stringent data governance. This includes implementing comprehensive data privacy and security measures, such as encryption, strict access controls, and audit trails, designed to protect sensitive information and the intellectual property embedded within AI models. It also involves establishing robust governance frameworks to ensure compliance with regulatory requirements, including the Australian Privacy Act, APPs, and international standards like GDPR. These controls foster trust, uphold ethical principles, and ensure your AI systems remain secure, transparent, and legally sound.

Integration with enterprise systems

For AI Factories to deliver maximum value, seamless integration with existing enterprise systems is crucial. This is achieved by utilising API-driven architecture to connect AI capabilities with business applications, facilitating real-time decision-making and process automation, and ensuring AI systems can access and process data from enterprise data lakes, enhancing the breadth and depth of insights generated.

Concentrated expertise

Fosters innovation by bringing together multidisciplinary teams of data scientists, machine learning engineers, and domain experts who collaborate to develop and deploy AI solutions.



Unleashing potential: the benefits of AI Factories

Investing in AI Factories unlocks significant strategic advantages for organisations:

Faster time-to-market

Accelerates the development and deployment of AI solutions, crucial for staying competitive and responsive to market changes.

Improved efficiency and cost-effectiveness

Optimises the AI development process through specialised infrastructure, reducing total cost of ownership for AI-intensive workloads.

Enhanced innovation

Creates a dedicated environment where experimentation and creativity thrive, enabling the rapid development of breakthrough AI-capabilities.

Increased agility

Empowers organisations to pivot quickly by rapidly deploying AI-driven strategies in response to emerging opportunities or threats.

Sustainable competitive edge

Establishes long-term differentiation by building proprietary AI capabilities in-house.

Resilience and availability

Delivers mission-critical reliability with high availability, fault tolerance, and workload flexibility.

Greener AI at scale

Enables energy-efficient and cooling-efficient AI operations that align with enterprise sustainability strategies and ESG (Environmental, Social, and Governance) objectives.

Data sovereignty and compliance

Maintains full control over sensitive data and ensures compliance with complex cross-border regulatory frameworks.

Talent magnetism

Serves as a centre of excellence that attracts world-class AI and engineering talent.

AI Factories in action: where software and infrastructure converge

As organisations race to build more intelligent systems, two foundational forces are converging:

- Software platforms that embed intelligence across the organisation, from operations to customer experience
- Specialised infrastructure engineered for high-performance model development and deployment, including large language models and other frontier AI systems

This section explores how leading organisations are building and integrating the critical components of the modern AI Factory, from platform-driven pioneers like Uber, Airbnb, and Netflix to infrastructure trailblazers such as Tesla, Meta, and Microsoft.

Operational platforms: embedding intelligence at scale

The very concept of the modern AI Factory was shaped by hyperscalers like Google, AWS, Alibaba, Tencent, and ByteDance. These companies seamlessly combined massive-scale infrastructure with intelligent software platforms to deliver intelligence at global scale.

Their early investments didn't just modernise data centres, they transformed them into software-driven engines of decision-making and automation, setting the pace for the broader industry.

Here's how platform-first pioneers are putting intelligence to work inside their organisations:

Uber

Michelangelo

Uber's ML platform, Michelangelo, enables real-time predictions and optimisations across the company. Supporting over 10 million predictions per second, it powers:

Intelligent dispatching

Matches riders with drivers using real-time data

Dynamic pricing

Adjusts fares based on live demand and supply

Route optimisation

Delivers efficient routing for drivers across Uber, Uber Eats, and freight

Airbnb

Bighead

Airbnb's Bighead platform supports:

Fraud prevention

Identifying and mitigating malicious activity

Search ranking

Ordering listings to match user intent

Customer personalisation

Enhancing guest experience using behavioural data

Netflix

Metaflow

Netflix created Metaflow to streamline the ML lifecycle:

Recommendation engines

Personalised content delivery

Workflow orchestration

Managing complex training and deployment pipelines

Rapid experimentation

Supporting fast iteration across teams

Physical AI factories: the infrastructure backbone

While software platforms bring intelligence to life across the enterprise, they rely on a new class of physical infrastructure, built for speed, density, and the demands of large-scale AI workloads.

This is the foundation of the AI Factory: purpose-built, high-performance environments that support the full lifecycle of modern models, from training and fine-tuning to inference at scale.

Today's leaders are deploying liquid-cooled, multi-megawatt systems optimised for the operational realities of machine learning, where performance is measured in token throughput, workload scalability, and the ability to run continuously at extreme density.

Microsoft + xAI

Colossus Supercomputing Infrastructure

Location: Memphis, TN (primary); Atlanta, GA (secondary)

Scale:

- 200,000+ GPUs deployed (H100/H200 mix)
- Scaling to 1 million GPUs
- 150 MW approved (Memphis site)

Purpose: Training and inference for xAI's Grok model family, integrated into Microsoft's Azure ecosystem

Architecture:

- Liquid cooling
- 400 GbE (NVIDIA Spectrum-X)
- Tesla Megapacks, renewable-backed grid
- All-flash storage

Status: Operational; expansion through 2026

Partners: Microsoft, xAI, BlackRock, MGX (UAE), NVIDIA (technical advisor)

Source: AI Infrastructure Partnership (AIP), March 2025. "How Musk's $30B Microsoft pact is reshaping AI infrastructure."

Meta

AI Research SuperCluster (RSC)

Location: Undisclosed Meta data centres (U.S.)

Scale:

- 6,080 A100 GPUs (current)
- Expanding to 16,000 GPUs; ~5 exaflops compute

Purpose: Training foundational models for NLP, vision, speech, and metaverse

Architecture:

- InfiniBand (1,600 Gb/s)
- AIRStore pipeline
- 175 PB FlashArray, 46 PB cache, 10 PB FlashBlade
- Full-stack encryption and anonymisation

Status: Operational

Partners: NVIDIA, Penguin Computing (SGH), Pure Storage

Source: Meta AI Research announcement, January 2022. "Introducing the AI Research SuperCluster."

Tesla

Dojo and Cortex AI Supercomputers

Location: Palo Alto, CA; Buffalo, NY; Giga Texas (Cortex)

Scale:

- Dojo: 3,000 Tesla D1 chips per Exapod; up to 7 Exapods
- Cortex: 50,000+ H100 GPUs in 2024; scaling to 100,000
- Target: 90,000+ H100-equivalent GPUs by 2024 end

Power:

- Dojo: 2.2 MW per cabinet (tested)
- Cortex: Water-cooled cluster; high-density AI loads

Purpose:

Training FSD, Optimus (robotics), multimodal video models

Architecture:

- Tesla D1 custom chips; D2 in production with TSMC
- Stainless steel racks; unified software + hardware stack

Status:

- Dojo operational since 2022
- Cortex fully deployed Q4 2024; further roadmap in place

Partners: Tesla, TSMC (Dojo), NVIDIA (Cortex)

Sources: Tesla AI Day, TechCrunch, CNBC, IEEE Spectrum (2019–2025). "Tesla's Dojo: A Timeline."

The strategic imperative

The AI Factory is emerging as the new blueprint for enterprise infrastructure, uniting software intelligence with purpose-built environments designed to scale. From real-time platforms to sovereign-grade systems, it delivers on five core demands of modern digital capability: speed, scale, sustainability, security, and sovereignty.

This isn't about keeping up. It's about building the infrastructure that puts you ahead.

The AI Factory imperative: build for the intelligence economy

The shift from traditional data centres to purpose-built AI Factories marks a critical inflection point for organisations seeking to fully realise the potential of artificial intelligence. This is not a linear upgrade, it's a foundational transformation of compute infrastructure, network architecture, and operational readiness to support the demands of next-generation intelligence workloads.

By investing in AI Factories, forward-looking enterprises equip themselves to handle the exponential growth in AI models, data volumes, and power density. They gain the strategic capability to innovate faster, compete smarter, and lead in an economy defined by intelligence.

This new era requires decisive action.
Success hinges on infrastructure partners who can deliver across the "5S" dimensions, Speed, Scale, Sovereignty, Sustainability, and Security, without compromise.

Your AI Factory starts here - with NEXTDC

The intelligence economy isn't on the horizon, it's already transforming industries. To lead in this new era, you need infrastructure that's not just ready for AI, but purpose-built for it.

NEXTDC's high-performance, AI-optimised data centre platform delivers the density, sovereignty, and sustainability your workloads demand today and tomorrow.

Let's build your AI Factory.

Connect with our specialists today to design the infrastructure that defines your next competitive advantage.

NEXTDC
where the cloud lives™

136 398
sales@nextdc.com
nextdc.com

This document is correct at the time of printing and is for presentation purposes only. This document does not constitute an offer, inducement, representation, warranty, agreement or contract. All information contained in this document (including all measurements, photographs, pictures, artist's impressions and illustrations) is indicative only and subject to change without notice. NEXTDC Limited, its employees, representatives, consultants and agents make no representations or warranties as to the accuracy, completeness, currency or relevance of any information contained in this document and accept no responsibility or liability whatsoever for any discrepancy between the information contained in this document and the actual data centres or services provided by NEXTDC Limited or for any action taken by any person, or any loss or damage suffered by any person, in reliance upon the information contained in this document. © 2025 NEXTDC Limited ABN 35 143 582 521.

HS05_2025_040725_01