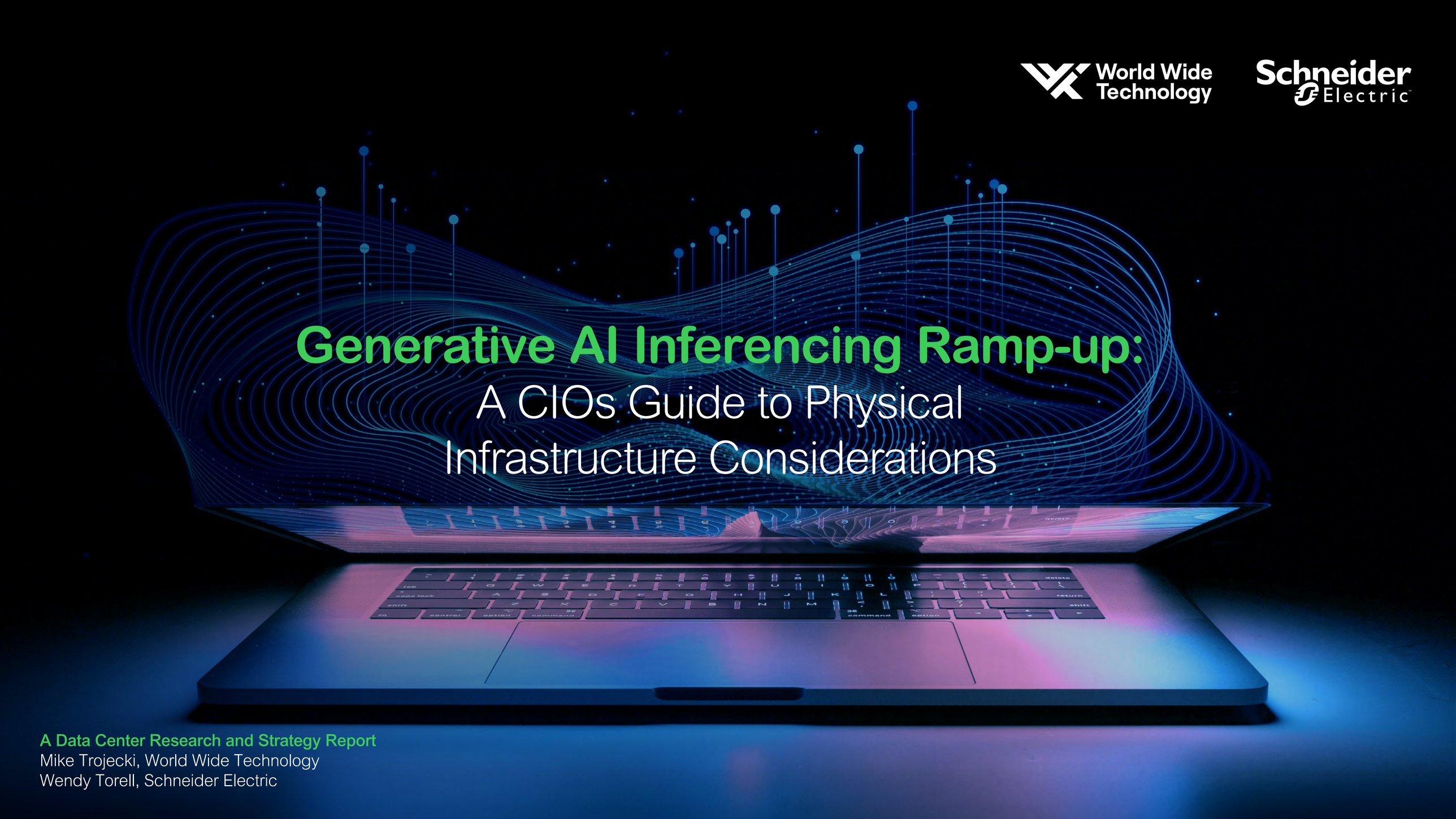


Generative AI Inferencing Ramp-up:

A CIOs Guide to Physical Infrastructure Considerations

A Data Center Research and Strategy Report

Mike Trojecki, World Wide Technology
Wendy Torell, Schneider Electric

FOREWORD

Inference is where generative AI meets the real world. Models are trained behind the scenes, but they become the main attraction in live deployments where they run millions of times. Many industry executives and observers assumed inference would follow a simpler, more predictable deployment path. We've experienced an inflection point: from infrastructure to business model, from sustainability to performance, inference is where value is created - or lost.

And the inference landscape is shifting rapidly. The scale of inference is explosive. Every interaction or application, every embedded use case, drives more cycles. What started at the core is increasingly pushing to the edge. While centralized systems and agentic AI are crucial, there are some cases where distributed and heterogeneous architectures are necessary.

This isn't a technology shift. It's part of the digital transformation journey. Inference alters how we plan capacity, allocate capital, design workloads, and organize operational teams. Cooling, energy, interconnects, latency – all are levers in developing and delivering continued strategic advantage.

Through our collaboration, WWT and Schneider Electric are delivering validated, adaptable solutions that minimize risk and reduce deployment uncertainty. Our goal is to help organizations seamlessly integrate AI into their operations—while maintaining operational efficiency, resilience, and sustainability.

Schneider Electric and World Wide Technology (WWT) developed this report to equip decision-makers with the context and clarity needed to act. To understand the necessary considerations for inference to scale reliably, efficiently, and with agility.

- We specify five types of inference workloads and the unique needs that lay the groundwork for IT and infrastructure
- We address key IT stack considerations for deploying AI inference for three hypothetical models
- We discuss how to deliver physical infrastructure that can be continuously "future-ready."

Inference is core to deploying any AI strategy. It's where AI ambition meets reality, where planned performance can encounter real-world constraints if not effectively considered and planned. Its one of the reasons our deep partnership with WWT is critical. We both believe in planning, testing and proving how our AI solutions drive results for our customers that work. Today.

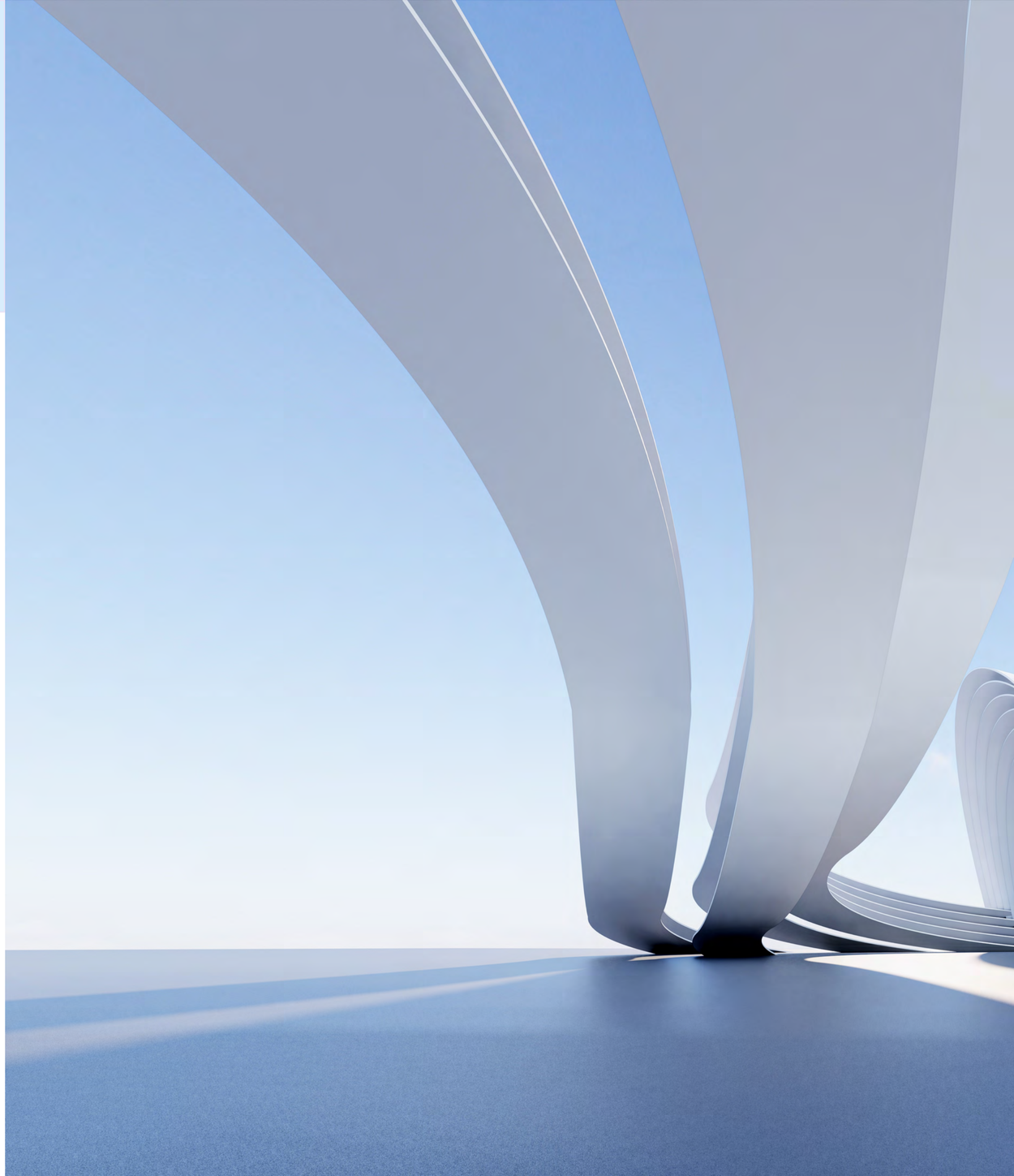
This isn't about chasing trends. As AI improves the way we interact across customer interactions and platforms, delivering better responses and speeding the workflow, it's not its presence that's important. It's about its ability to provide value.

Read on, as we believe the emerging shape of inference is central to how you will define your competitive edge.

Yours,



Paul Tyrer
Global VP
AI Ecosystem
Schneider Electric



KEY TAKEAWAYS

1 Not all generative AI inference workloads will be “business as usual”

Once it was commonly held that ultra-high power capacity and ultra-high rack density workloads were primarily for training. Inference would be “business as usual” and would run on commodity gear. Yet, we find that is no longer true: both aspects of AI will need to consider unique physical infrastructure requirements. Some generative AI inferencing workloads will be low capacity and low rack density, but many will not.

2 Model complexity and power requirements depends on use case

A wide spectrum of use cases spans from simple batch-process models to very complex, data intensive, low latency, reasoning, and agentic AI models. Digital twins, computer vision, and edge use cases add complexity, so understanding emerging technological trends is crucial for maintaining a competitive advantage. There is not a “one-size-fits-most” deployment in terms of physical infrastructure (e.g., power, cooling) needs.

3 An open and collaborative ecosystem of partners is important for effective planning and implementation

Given the complex IT stack and rapid pace of technological change, engaging professional services within this ecosystem reduces burden and uncertainty and mitigates technical debt from urgent AI adoption. Partners may offer reference architectures incorporating validated, integrated designs, that accelerate deployment and provide proven paths for generative AI inference physical infrastructure.

4 Inferencing continues to be largely in the cloud

The cloud provides a flexible, cost-effective, and resource-rich environment while reducing operational overhead with faster time to value. This aligns with the needs and cautious approach of many organizations venturing into this evolving landscape.

5 Some workloads have inherent on-premise requirements

Specifically, applications demanding strict data sovereignty/security or requiring ultra-low latency frequently necessitate deployment in an on-premise or local colocation data center due to limitations in cloud environments.

6 When deploying on-premise workloads, future-proofing the physical infrastructure is advised

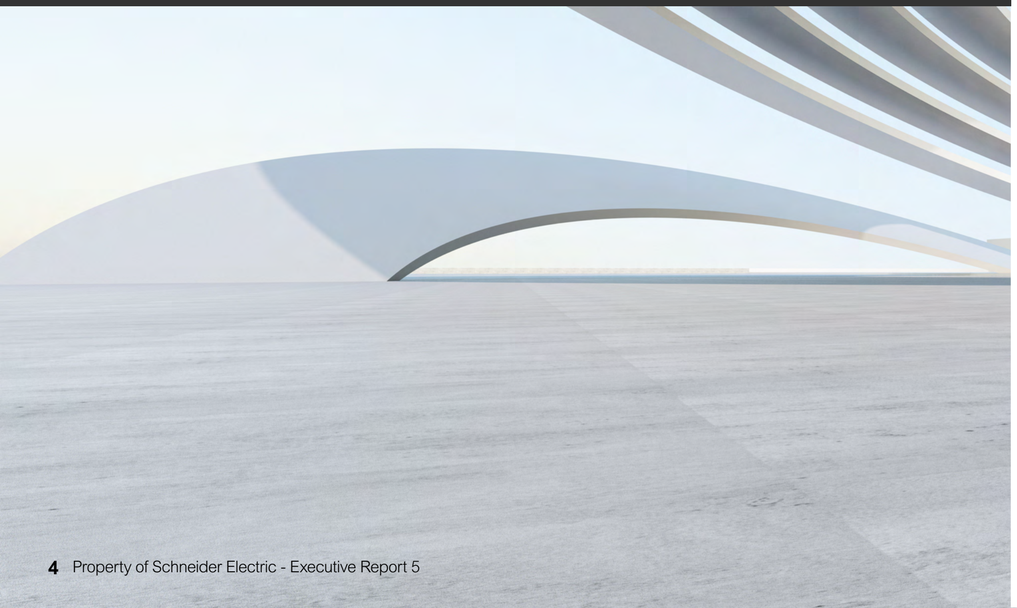
To address the rapid advancements in AI and its growing range of applications, a flexible and modular infrastructure, supported by robust software for visibility and management, is essential. This also includes planning for cooling evolution, like liquid cooling, to handle the higher density of future AI hardware.

RATE THIS REPORT

LET’S GET STARTED

While chatbots are still the most commonly encountered AI applications, generative AI (a.k.a., Gen AI) continues to extend its computational capabilities across domains.

It is reshaping workflows, enhancing creativity, and driving innovation.



> GenAI represents the next landmark opportunity for businesses to sink, swim or soar. Your job as CEO is to figure out how to harness AI to execute the leap in innovation required to build the *next* disruptive business model, product or service.
>
> **Jim Kavanaugh**, Co-Founder and CEO, WWT

2025 has been a pivotal year, as we see Gen AI moving beyond pilot projects into implementation. In industries such as life sciences, finance, government, and manufacturing, executives are deploying sophisticated live inference workloads that provide real-time responsiveness. This enables them to directly answer customer queries on demand, generate tailored patient plans at the point of need, and identify anomalies with immediate effect, significantly enhancing operational efficiency and decision-making.

It's time to consider the best paths for integrating inference AI into the business. Physical infrastructure decisions can no longer be framed around the 2023 assumption that training is extraordinary, inference is business as usual. That assumption does not line up with 2025 knowledge. Some inference jobs do resemble the simple chatbots we've come to expect from AI, but a growing share are approaching power densities once thought to be reserved for training clusters.

Although many CIOs and their boards are actively driving value through prioritized use cases, gaining a full understanding of the infrastructure implications is still crucial to long term planning. This requires evaluation of model complexity, latency, and data policy – then, clearly demonstrating their connection to tangible factors like rack power, cooling topology, and deployment strategies across various environments.

To provide this clarity, we will address three key questions:

1. How can we determine the ideal physical location for these workloads — cloud, colocation, or on-premise?
2. What are the specific power capacity and density requirements driven by different inference workloads?
3. How can organizations proactively build infrastructure with the necessary headroom to smoothly integrate future model upgrades without significant disruptions?

In this executive report, we explore the evolving nature of Gen AI, focusing on inference workloads. We discuss its impact on deployment approaches and the critical physical infrastructure that underpins its growing capabilities.

Distinguishing between the two distinct phases of AI's lifecycle.

Training

Is the computationally intensive process of building the AI model by exposing it to vast datasets, enabling it to learn underlying patterns and relationships. (In White Paper 110, How 6 AI Attributes Change Data Center Design, we dive into these workloads and their physical infrastructure implications). Gen AI discussions have centered largely on the complexities and resource demands of the training phase, and their specialized "training centers" equipped with the immense computational power necessary to develop sophisticated AI models.

Inference

On the other hand, is the phase where the trained model is put to work, generating predictions, insights, or content based on new, unseen data. As Gen AI matures and transitions from exploration to practical application, the significance of the inference phase is growing. It is during inference that organizations realize the tangible value of their AI investments, deploying models to address real-world challenges and create new opportunities. However, industry uncertainty and emerging questions arise around deployment strategies.



The wide spectrum of Gen AI applications

The landscape for Gen AI is shifting initial assumptions about deployment and revealing a wide spectrum of applications. The early view centered on massive, centralized training facilities complemented by separate, lighter weight "inference centers".

However, this view, like the technology, has changed as AI models have become more varied and sophisticated.

Inference models are moving beyond simple text outputs toward complex reasoning. They work with diverse data types (multi-modal capabilities). They tackle increasingly intricate problems. They employ techniques like chain-of-thought prompting to decompose multifaceted problems by breaking them down into logical, sequential steps. This evolution is fueled by relentless technological advancements across hardware, software, and algorithms. These advancements are leading to greater efficiency, such as more processing per unit of energy or tokens per watt, and are broadening the scope of potential use cases.

As AI becomes more capable — especially with the rise of "agentic AI" designed to make autonomous decisions—it is likely to have a significant impact on workflows. NVIDIA's CEO Jensen Huang opened CES 2025 by declaring it the "Year of AI Agents," projecting that these autonomous programs represent a "multi-trillion-dollar opportunity" and heralding an "Age of AI Agentics" with a new digital workforce.[1] This transformation involves AI systems that perceive their environment, set goals, make decisions, and execute tasks with minimal human intervention. However, this iterative and autonomous approach to problem-solving causes a significant increase in computational requirements during inference.

What is the WWT AI proving ground?

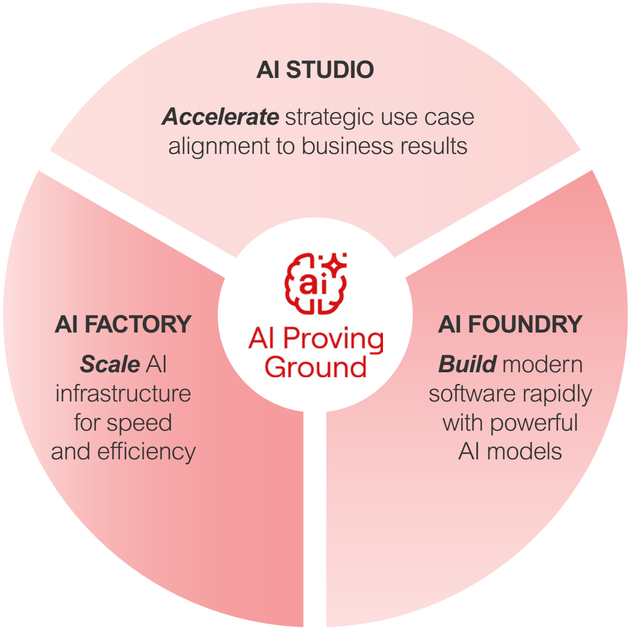
The AI Proving Ground (AIPG) provides state-of-the-art access to the world's leading AI technologies. Powered by our Advanced Technology Center, this unique lab environment accelerates client ability to learn about, test, train and implement AI solutions.

The World Wide Technology AI Proving Ground serves as an environment for organizations to explore Gen AI's potential without committing to significant capital expenditures on on-premise infrastructure. This model allows businesses to:

- Experiment with various models and use cases.
- Model your own environment.
- Scaling resources up or down as needed.

The AI Proving Ground is at the center of WWT's practical approach to aligning business needs to the outcomes-focused pillars of Gen AI, including WWT's AI Studio, AI Foundry, and AI Factory.

Figure 1: AI Proving Ground framework.





Meaningful use cases are growing

While many organizations are still in the proof-of-concept and testing phases, a cautious yet systematic deployment is underway.

Life sciences

Life Sciences companies are increasingly using data, machine learning processes and technology to drive and accelerate value. However, silos of compute as well as suboptimal data and code practices are reducing scientist productivity, ROI and drug program success. WWT and partners including NVIDIA are addressing these challenges by using data strategy services, code optimization practices and purpose-built AI infrastructure to deliver cost-optimized, hybrid cloud machine learning (ML) platforms.

Use cases for life sciences

- R&D focus areas (algorithm-driven)
- Computational chemistry
- Structural biology
- Genomics

Business-area driven

- Clinical trials
- Supply chain
- Commercialization

Scaling AI in Life Sciences





Financial institutions

Leading banks are realizing real improvements in employee productivity and operational resilience through AI – but success comes from focusing on pragmatic use cases, strong governance and user adoption rather than hype. Workplace services teams should prioritize a handful of scalable AI initiatives (e.g., virtual service agents, intelligent automation in workplace processes) paired with governance and upskilling to drive near-term results while laying the groundwork for broader transformation.

Key insights: What's working now vs. what's emerging

What's working now (high-impact use cases from 2024)

- AI virtual assistants for IT support
- Process automation and "copilots" for employees
- Enhanced workplace intelligence platforms
- AI-enhanced training and onboarding

What's emerging (trends to watch in 2025)

- GenAI moving beyond pilot stage
- AI-augmented employee experience platforms
- Predictive workplace analytics
- Focus on governance, risk and skills

Real-world examples in the banking industry

The following examples show how banking organizations are using AI to enhance workplace services.

- **JPMorgan Chase** – Enterprise AI assistant at scale.
- **Banco Bradesco** – GenAI for employee efficiency
- **Wells Fargo** – AI-augmented workplace training and productivity
- **Standard Chartered** – AI-enhanced employee experience platform
- **HSBC** – AI-powered knowledge management

AI and Automation in Banking: Transforming Workplace Services

Meaningful use cases are growing

Emerging use cases also include government. At the same time, data centers are changing as well, as this case study indicates.

Government

Smart Cities and Infrastructure:

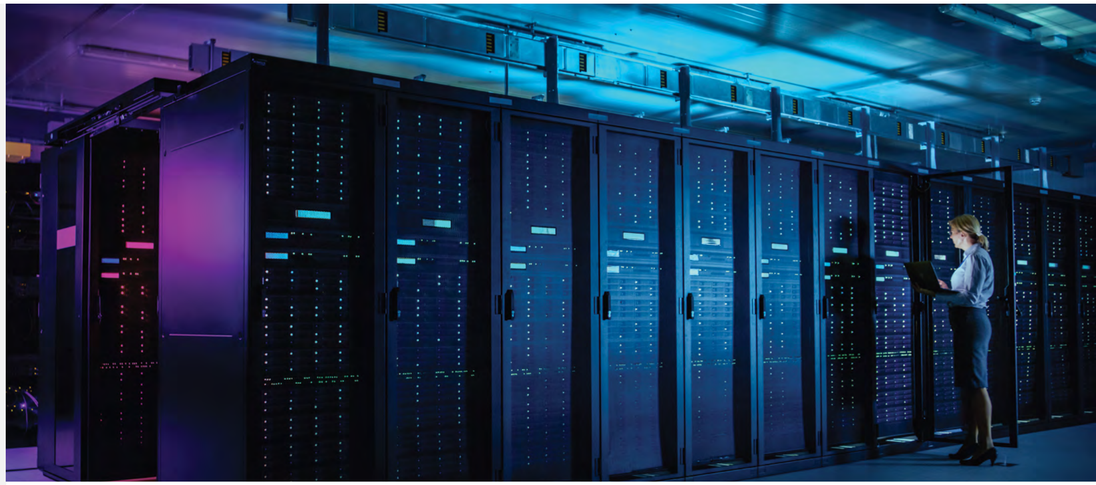
- **Optimizing traffic flow** through AI-powered systems that analyze real-time data, predict congestion, and adjust traffic lights, potentially reducing congestion and emissions.
- **Enhancing public safety** by leveraging AI-powered cameras and sensors to monitor city spaces and provide real-time alerts, potentially allowing for faster responses to dangerous situations or suspicious activities.
- Using digital twins to model and simulate urban environments for better planning and decision-making, such **as predicting flooding during storms or optimizing public transportation routing**.

Examples include the AMIGOS project in Las Rozas, Madrid, using AI-driven cameras and sensors for city planning, and the ELABORATOR initiative in Issy-les-Moulineaux, France, employing AI to manage intersections and prioritize cyclists.

Smart Cities: Integrating AI and open data

Digital Twins and Urban Planning: Designing Smarter, More Inclusive Cities





Scott Data case study

Scott Data saw an opening in the market—and ran at it. Most data centers couldn't meet the power and cooling demands of modern AI workloads. Instead of retrofitting, Scott Data repositioned entirely, shifting from traditional colocation to GPU as a service.

The move wasn't speculative. It was deliberate. Working with World Wide Technology, Scott Data co-designed a new architecture built around high-density compute. They started with what they had—existing infrastructure—and ran a proof of concept to validate the approach before scaling.

The result: An on-demand GPU platform that lets clients tap AI-ready compute without the capital outlay of building their own. It's more than capacity. It's speed to capability.

Now Scott Data is scaling to meet demand in sectors that need AI performance but don't want to be in the data center business—agriculture, healthcare, finance, and others. It's not just a service pivot. It's a full shift in business model, designed for a software-defined, workload-driven future.

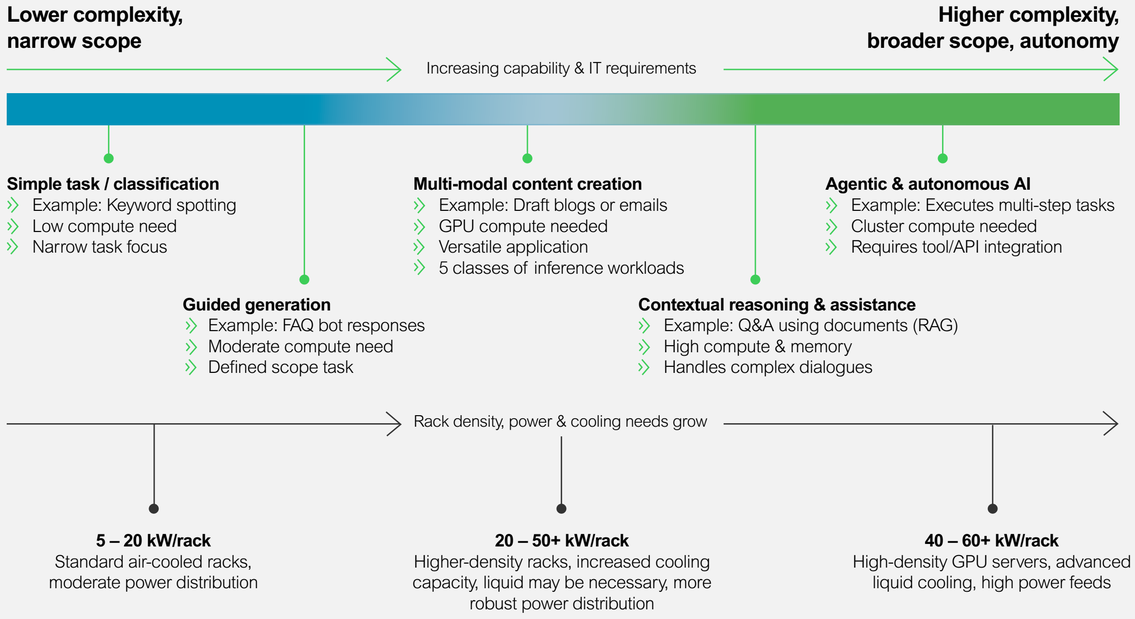
Scott Data Embraces AI Transformation to Offer GPU as a Service

Figure 2 illustrates the vast spectrum of complexity and purpose within Generative AI inference workloads, leading to dramatically different underlying IT requirements, and therefore physical infrastructure requirements. The figure focuses on model complexity, but it's important to note there are other drivers to the IT (and physical infrastructure) requirements, such as number of concurrent users, latency requirements, model accuracy, etc. We discuss these later in detail.

A complex agentic system demands significantly more resources than simpler applications. Simply put, this is because they perform chain-of-thought reasoning, call external APIs (tools), and iterate until a goal is met. Each extra step drives more tokens, more GPU cycles, and therefore more watts. This variation dictates everything from compute hardware (e.g., a single GPU versus a large cluster of them), network bandwidth for data and model transfer, memory footprint, and the sophistication of software orchestration frameworks.

Figure 2

The spectrum of model complexity, 5 classes of inference workloads



Beyond purely digital agents, the trajectory of AI extends towards physical embodiment in robotic systems. These robotic AI applications, also known as "physical AI", while often leveraging the foundational capabilities demonstrated in agentic models, introduce an entirely new layer of IT requirements. The need for real-time sensor data processing, low-latency control signals, and robust integration with physical actuators adds significant complexity to the infrastructure. Furthermore, the training of these systems often involves simulations and reinforcement learning in complex physical environments, demanding specialized compute and potentially large-scale simulation platforms.

Beyond computational needs, other critical factors significantly impact IT requirements. The size and nature of the datasets used for fine-tuning or retrieval-augmented generation (RAG)[2], as well as the data locality and security policies, can heavily influence storage and network considerations. Latency requirements for real-time applications, regardless of model size or whether they are purely digital or embodied, can also necessitate specific infrastructure choices and potentially edge deployments. Furthermore, the choice of deployment environment (cloud, colocation, on-premise) and the need for scalability to handle varying user loads or the deployment of numerous robotic units are crucial determinants of the overall IT architecture.

Consequently, inference workloads span a wide range of infrastructure solutions. This ranges from application-specific chips that are hyper-optimized for particular models / use cases to "AI factories" requiring the latest and most powerful hardware for agentic models. The deployment of models trained on proprietary data through techniques like RAG, along with the use of compressed and more specialized models, further diversifies the inference landscape. This can influence the optimal balance of compute, memory, and network resources. Finally, the compute needs for inference are escalating, requiring more tokens, processing power, and real-time responsiveness even outside of the initial training phase.



The Cloud provides flexibility, risk mitigation, and access to resources

Wider adoption of Gen AI faces two key deployment barriers: skills gaps and infrastructure constraints. The limited availability of skilled professionals to develop, deploy, and manage these advanced systems poses a significant challenge. Simultaneously, the demanding infrastructure requirements, particularly concerning compute resources, power consumption, and cooling, present considerable hurdles, especially for on-premise and hybrid deployments.

Organizations often conduct initial experiments, pilots, and early implementations in the cloud to mitigate infrastructure investment risks and address the need for specialized skills. Using partners who can deliver testing environments and expertise provides an effective option to explore before commiting. This allows for a more cautious approach, deferring significant hardware commitments until the business value and workload characteristics are better understood.



Reasons for "Cloud Smart" approach

1 Cost flexibility & risk reduction

- Pay-as-you-go pricing model
- Eliminate large upfront cost expenditures
- Minimize risks of over-provisioning and technology obsolescence

2 Access to specialized AI hardware & expertise

- Access to GPUs/TPUs
- Managed AI services
- Leverage cloud provider's specialized knowledge and support

3 Access to continuous innovation

- Access to the latest pre-trained and customizable models
- Integrated AI development tools and frameworks
- Enable continuous adoption of advancements in AI/ML

4 Scalability & elasticity

- Scale resources easily and on-demand
- Adapt to new technologies quickly
- Enable rapid iteration and experimentation

5 Accelerated time to value

- Faster development cycles
- Rapid prototyping and POCs
- Focus core business resources on innovation

Cost flexibility & risk mitigation:

Many companies are experimenting with Gen AI to prove its value within their specific contexts; the cloud offers a lower barrier to entry for these initiatives, minimizing upfront investment. Cloud servers have essentially served as a lower risk "test bed" for organizations to explore Gen AI's potential without committing to significant capital expenditures on on-premise infrastructure. See also page 6 for more information on WWT Proving Ground. The pay-as-you-go model allows businesses to experiment with various models and use cases, scaling resources up or down as needed.

Access to specialized AI hardware and expertise:

Moreover, cloud providers often have better access to the limited supply chain of specialized AI hardware, such as high-performance GPUs and TPUs, crucial for both training and inference of advanced models. Beyond hardware, cloud platforms offer managed AI services and a wealth of expertise and support to help organizations navigate the complexities of Gen AI. For instance, services like Azure OpenAI Service[3] allow businesses to integrate powerful language models into their applications without managing the underlying infrastructure, enabling rapid deployment of generative AI solutions.

Access to continuous innovation of models, platforms, and tools:

In some instances, the cloud may represent the only avenue for accessing the latest, most advanced AI models, as model providers often release their newest innovations first or exclusively through their cloud platforms. These "AI factories" house the most advanced and computationally intensive AI models and can achieve high efficiency through economies of scale and optimized infrastructure. Content libraries, models, and frameworks such as NVIDIA's NGC catalog[4] provide access to 600 GPU optimized models, containers and scripts across use cases including NLP, computer vision, and speech recognition. Coreweave offers access to NVIDIA GPUs to facilitate training and deployment. All of the tech giants offer access: AWS, Google, IBM, Microsoft, Oracle, and META. This enables organizations access to powerful tools to stay apace with technological advancements available in integrated AI tools and platforms.

Scalability and elasticity:

Given AI's rapid technological evolution, organizations often reduce risk by leveraging the agility and scalability of cloud platforms. The Gen AI ecosystem itself is still in a state of flux, with new models, tools, and best practices emerging frequently. This makes the flexibility of the cloud particularly attractive as it enables the scaling of resources, up or down, to adapt to these changes and support rapid iteration.

Faster time to value:

The cloud provides a flexible and resource-rich environment that contributes to a faster time to value with readily available infrastructure and services. This allows for quick development and deployment of PoCs, enabling companies to focus on business outcomes and accelerate their innovation cycles.

Overall, the cloud, whether it be traditional cloud or specialized GPU cloud, also known as GPU-as-a-Service (GPUaaS)[5], provides a flexible, cost-effective, and resource-rich environment with reduced operational overhead and faster time to value, aligning with the needs and cautious approach of organizations. For these reasons, we anticipate this strategy will continue to be leveraged for the foreseeable future.

While concerns about networking bottlenecks have historically existed, advancements in cloud networking infrastructure are increasingly addressing these issues[6], making the cloud a viable option for even demanding AI workloads. On the flip side, there are times, when on-premise or local/private cloud or colocation are the only avenue.



Some use cases are still better suited to on-premise

While cloud deployment makes sense for many applications, there are some Gen AI applications deployed on-premise, with the IT stack closer to the point of use, either on-premise or at a local colocation site. In practice, three key factors tend to guide this decision. These can be thought of as a sequence of gates. A single yes at any gate inclines the workload to land locally.



Reasons for on-premise, close to point of use

1 Ultra-low latency requirements

- Real-time AI inference
- Avoids cloud network latency

2 Data security, compliance, regulations

- Greater control over data access
- Adherence to regulatory requirements
- Safeguard company reputation
- Mitigate potential financial and legal liabilities

3 Business & financial model preference

- Cost and control benefits
- Safeguards valuable IP
- Maintain competitive edge

Ultra-low latency & high criticality requirements

For applications requiring real-time processing and ultra-low latency, edge or on-premise deployments might be necessary to minimize delays. This is non-negotiable when dealing with high-speed, critical physical operations where any delay is unacceptable. Below are four examples of such applications.

Real-time surgical assistance and robotic surgery:
Gen AI models analyze live video feeds, medical imaging, and patient data during surgical procedures to provide surgeons with real-time insights, guide robotic surgical arms, or even perform certain autonomous tasks under supervision. Patient safety is paramount. Delays in image processing, AI-driven guidance, or robotic control could have severe consequences during surgery. For instance, if an AI is assisting with identifying critical anatomical structures or guiding a robotic arm, any lag could lead to errors with potentially life-threatening outcomes. Millisecond-level latency is crucial for precise and safe surgical interventions.[7]

Smarter factory robotics:
Gen AI can rapidly generate new, coordinated paths for teams of robots working closely together. Running this AI locally is essential to prevent collisions (avoiding potential fires or worker injuries) and to maximize efficiency in dynamic environments - split-second timing cannot tolerate cloud communication delays.[8]

Perfecting complex manufacturing:
In processes like precision metal 3D printing or semiconductor fabrication, Gen AI can integrate real-time adjustments to machine settings based on live sensor data. This delivers perfect quality and prevents costly material waste or defects, requiring immediate, on-site AI control that cloud latency would disrupt.[9]

High-frequency algorithmic trading:
In high-frequency trading, even a few milliseconds of delay can mean the difference between a profitable trade and a significant loss. Competitors with faster systems have a distinct advantage. Sending data to and from a cloud environment would introduce delays that render the trading algorithms ineffective. Co-location within or very near exchange data centers (often considered a form of on-premise for this purpose) is essential.[10]

Data security, compliance, regulations

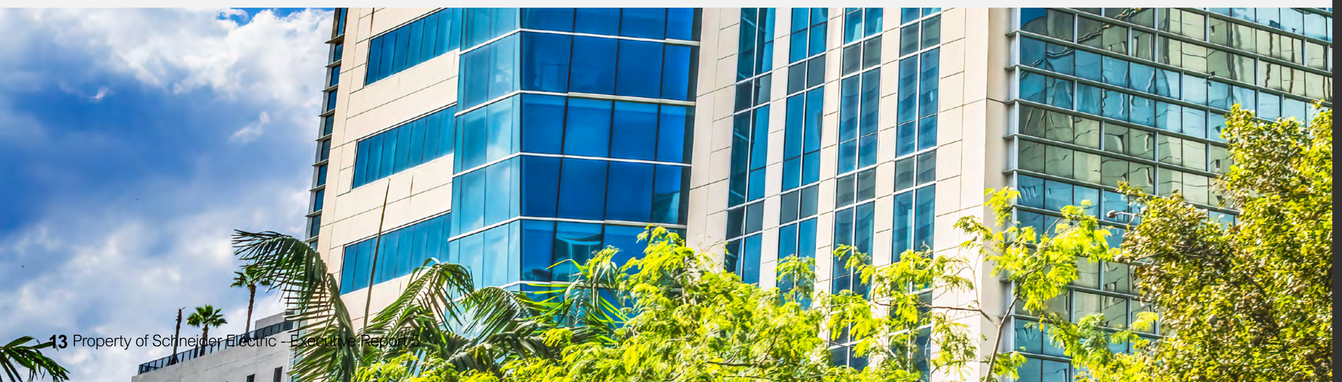
When dealing with highly sensitive information such as classified intelligence, protected health records, or confidential financial data, the risks associated with transmitting and storing this data in a third-party cloud environment can be significant. NIST highlights various security and privacy challenges inherent in public cloud environments that can pose significant risks for highly sensitive information such as classified intelligence, protected health records, or confidential financial data.[11] An on-premise solution provides greater control over:

- data access
- security protocols
- physical infrastructure
- adherence to strict regulatory frameworks like HIPAA, GDPR, or defense-related classifications

This helps safeguard an organization's reputation and mitigates potential financial and legal liabilities. Stringent country data sovereignty laws, requiring that sensitive data must reside within national borders, can also necessitate on-premise or local deployments to ensure compliance with these regulations and avoid cross-border data transfers. Citizen data would be a prime case for localized or collocated inference workloads. Two example applications help to illustrate on-premise deployment use cases:

Military/defense secure intelligence analysis and threat assessment:
Gen AI models are used to analyze classified intelligence data, including satellite imagery, signals intelligence, human intelligence, and open-source information. Military and defense organizations have strict protocols and infrastructure designed to protect this information. This involves air-gapped networks and heavily secured physical locations. Cloud environments, even FedRAMP-certified ones, might not meet the most stringent security requirements for top-secret data.

Real-time analysis of medical imaging for critical diagnostics:
Gen AI models analyze medical images (like MRI, CT scans, X-rays) in real-time within a hospital or clinic setting to assist radiologists and clinicians in making critical diagnoses. For certain highly regulated or uniquely sensitive patient data applications, hospitals might adhere to internal security protocols and infrastructure that they believe offer the most stringent protection, potentially exceeding their comfort level with even HIPAA-certified cloud environments.





Business/financial model preference

Organizations anticipating consistently high and predictable workload demands may find the long-term cost of ownership more attractive than variable cloud expenses. This is particularly true for businesses managing extremely large, proprietary datasets, where recurring cloud egress fees can become a significant financial burden. Consider a major film studio that uses Gen AI for advanced special effects rendering. They might have petabytes of uncompressed video footage and proprietary 3D models. Running their AI inference in the cloud would mean constantly uploading and downloading these massive datasets, incurring exorbitant egress fees for every frame rendered or every model refined. Keeping these workloads on-premise allows them to process and move this data internally, avoiding those recurring data transfer costs. Furthermore, some organizations prioritize capital expenditure over operational expenditure or prefer the strategic control and ownership inherent in managing their own infrastructure.

Beyond cost and control, the need to safeguard valuable intellectual property and maintain a competitive edge through innovative AI models can also strongly favor on-premise deployments. Keeping sensitive model architectures, training data, and unique AI innovations within their own secure environment offers enhanced protection against unauthorized access or data leakage. This preference for ownership and control, driven by both financial considerations and the strategic imperative to protect IP, can make the upfront investment of an on-premise solution a more aligned business strategy, despite potential technical debt risks.

WWT best practices recommendation

Balance cyber risk and opportunity with AI

Whether or not your organization has a firm grasp on the secure and responsible use of Gen AI, employees in many organizations are already using these tools in their day-to-day work. Current challenges include integrating AI projects into existing processes and systems and then securing them. Many organizations are seeking to leverage the power of AI and take advantage of the opportunities it presents while also balancing the new risks these tools introduce.

When striking the proper balance of risk and opportunity in AI, your organization is positioned to:

- Increase speed and efficiency in operations
- Augment staff and resources to help attain any competitive advantages
- Improve resilience if an AI tool is compromised
- Support compliance with data privacy regulations
- Help maintain reputation and customer trust
- Reduce the risk of costly remediation and recovery

WWT emphasizes the importance of data security, privacy, and compliance in the context of Gen AI.

Securing AI throughout the enterprise

As AI use cases grow in scope and number worldwide, new attack surfaces and AI-specific threats have been documented and are increasing. Organizations using AI systems are implementing security measures to manage and secure the use of AI across the enterprise. Common security measures include strong user authentication, a centralized AI governance framework, regular audits to identify unauthorized AI deployments, and additional controls.

To secure the use of AI, organizations typically benefit from:

- Evaluating overall use of AI broadly, including AI systems used internally or via SaaS. This may include specific AI models, data science tools, shared MLOps and data analysis platforms.
- Evaluating the approach to assessing the potential vulnerabilities in AI models and related systems.
- Assessing current AI security capabilities, such as data governance, model management, vulnerability management, and red and blue team exercise
- Identifying defenses against AI attacks, such as prompt injection, data poisoning, model theft, adversarial examples, and other emerging threats.
- Reviewing existing plans for AI and AI security.



Using AI to improve enterprise cybersecurity

The increasing amount and complexity of data and threats, combined with talent shortages in cybersecurity, have created significant challenges for many cybersecurity teams. Gen AI has the potential to help manage threats, evaluate IT ecosystems and operations, increase speed and efficiency, augment staff and resources, and boost productivity.

To take advantage of AI in cybersecurity, organizations may consider:

- Assessing the current use of AI for cybersecurity defenses.
- Reviewing plans and use cases for AI in security.
- Evaluating and prioritizing recommendations based on value, complexity and feasibility.
- Building a roadmap for AI security improvement based on established practices and frameworks.

> Our goal is to break down those silos and align strategy across the C-suite. Security isn't just the CISO's job anymore. It's an executive priority that touches everything from customer data to global compliance.
>
> **Chris Konrad**, Vice President of Global Cyber at WWT

Considerations in IT stack deployment

For those requiring on-premise deployment, the decision around the appropriate IT stack for AI inference is multifaceted. It's important to align the chosen infrastructure with the specific demands and limitations of the AI application(s) to deliver optimal performance, cost-efficiency, and compliance.





Table 1 outlines key IT infrastructure considerations for deploying AI inference models, highlighting how these needs vary across a spectrum defined by model complexity and latency demands. However, several crucial variables impact the optimal IT stack, including the following. Understanding these factors is paramount for executives to make informed decisions about technology investments and resource allocation for their AI initiatives.

- sophistication of the AI model itself
- number of concurrent users accessing it
- required speed of response (latency)
- level of accuracy needed for the application
- any constraints imposed by data sensitivity and regulations

The table presents three distinct scenarios across the spectrum from **Figure 2**: simple batch process inference, near real-time inference of moderate complexity, and complex real-time reasoning/agentic models. As we move from left to right, generally both model complexity and the demand for low latency increase. For instance, a simple batch process for overnight reporting requires less powerful compute and has relaxed latency requirements compared to a complex fraud detection system needing to analyze transactions in milliseconds. The number of concurrent users also plays a significant role; a model serving a large number of simultaneous requests will necessitate more scalable compute, memory, and networking resources than one serving only periodic or well-scheduled requests.

Table 1
Key IT stack considerations for deploying AI inference models for three hypothetical models

IT stack consideration	Simple batch process inference model (Low complexity)	Near real-time inference (Moderate complexity)	Complex real-time reasoning or agentic inference (High complexity)
Compute hardware	**Cost-effectiveness for large offline tasks:** Prioritize hardware optimized for high throughput in batch processing. Consider options like high-core-count CPUs or cost-optimized GPUs. Scale based on dataset size and processing time.	**Sustained processing & responsiveness for moderate concurrency:** Focus on hardware that can handle continuous data streams with reasonable latency for a moderate number of concurrent users. Consider multi-core CPUs and potentially mid-range GPUs.	**Low latency & high throughput for high concurrency:** Focus on hardware that delivers rapid processing for immediate responses to a potentially large number of concurrent users. Consider powerful GPUs, specialized AI accelerators, and edge computing.
Memory and storage	**High capacity, moderate speed for large datasets:** Emphasize storage solutions for large training and inference datasets. Cost-effective options are key. Memory needs driven by batch size.	**Fast access for streaming data and moderate concurrency:** Prioritize memory and storage with relatively low latency to handle continuous data ingestion and processing for a moderate user load. Consider fast SSDs and sufficient RAM.	**Fast access, moderate capacity for complex operations and high concurrency:** Prioritize low-latency memory and storage for quick data retrieval and processing for potentially many concurrent users. Consider fast SSDs and ample RAM.
Networking	**Reliable throughput for data transfer:** Ensure sufficient bandwidth for moving large datasets. Latency is less critical for batch processing.	**Moderate latency & good bandwidth for streaming data:** Ensure sufficient bandwidth for continuous data flow and reasonable latency for timely processing for the expected user base.	**Low latency & high bandwidth for real-time interaction:** Critical for immediate responses to a potentially large number of concurrent users and continuous data streams. Robust and high-speed networking is essential.
Model deployment and serving tools	**Scalable batch processing:** Focus on tools that efficiently deploy and execute batch inference jobs on large datasets. Consider workflow management and scheduling.	**Stream processing & continuous deployment with scaling:** Utilize tools supporting continuous data streams and frequent model updates, with the ability to scale based on data volume and user demand.	**Low-latency serving & monitoring for high concurrency:** Prioritize tools enabling fast and reliable deployment for real-time inference, with robust monitoring and scaling capabilities to handle numerous concurrent users.
Deployment environment	**Likely in the cloud (cost-optimized) but on-premise/colo for strict data requirements:** Cloud often provides cost-effective and scalable options. However, strict data sensitivity or regulations might necessitate on-premise or colocation regardless of model complexity.	**Cloud-based stream processing (scalable) but on-premise/colo for strict data requirements:** Cloud offers robust, scalable services. However, strict data sensitivity or regulations might require on-premise or colocation.	**Cloud, on-premise, or colo:** Ultra-low latency requirements and/or strict data sensitivity/regulations might necessitate on-premise or colocation solutions. Otherwise, the cloud offers scalability and managed services.

Accuracy needs can also influence the IT stack. More complex models often aim for higher accuracy but come with increased computational costs. Conversely, simpler models might be sufficient for applications with lower accuracy requirements, allowing for more cost-effective infrastructure. Even a simple model that uses chain-of-thought inferencing will require more compute resources compared to quick responses. Critically, data sensitivity and regulations can override other considerations. Even a simple, low-latency model might require deployment in an on-premise or colocation environment if it handles confidential data subject to strict compliance rules.

A robust ecosystem of technology providers supports these diverse needs. For example, Dell Technologies[12] provides hybrid AI infrastructure solutions, enabling deployment where data resides for security and compliance. NVIDIA provides cutting-edge GPU technology crucial for accelerating both training and inference across the complexity spectrum[13]. Furthermore, major cloud providers like Amazon Web Services (AWS)[14] and Microsoft Azure[15] offer a comprehensive suite of services, from cost-effective compute and storage for batch processing to low-latency, scalable platforms for real-time applications, empowering organizations to implement AI across this entire spectrum.

Five essential tips for AI planning and integration

When embarking on the integration of AI into your operations, it's important to approach the process with a strategic mindset. These quick tips are designed to guide you through the complexities of AI deployment, ensuring that your technological advancements align with both immediate and long-term objectives. From engaging IT and facilities, to considering energy demands and sustainability, these insights will help you navigate the evolving landscape of AI technology efficiently.

Special thanks to Mike Parham (WWT), and Victor Avelar (Schneider Electric).

1 Be aware of your long-term goals as well as your immediate goals

Always keep an eye on the horizon. When planning, it's imperative to address not only your immediate needs but also potential future requirements. As you consider your growth strategy, remember that future needs might alter your trajectory. Ensuring you have a long runway for growth will help you adapt to changes as they arise.

2 Engage IT first and facilities afterward

When contemplating the integration of AI into your operations, start by clarifying your goals and desired outcomes to IT leaders as a first step. Next, evaluate where your AI operations should physically take place. These decisions will determine the amount and density of compute resources required on-site. Once you have a solid plan, your facilities management can perform its critical role initiatives. Facilities engagements are important for designing and implementing the necessary infrastructure modifications to support your future AI needs effectively.

3 Stay informed of impending changes

Many OEMs and vendors have new technologies and tools coming out which can change a lot of AI dynamics. Stay close to the sources of truth and be aware of implications of new technologies. Partners such as Schneider and WWT, industry blogs, thought leadership articles, and AI-focused events are invaluable resources that will keep you abreast of what's on the horizon but also what's working now.

4 Consider energy conservation and sustainability

With the increased focus on sustainable computing and energy efficiency in past years, businesses have made great strides. Yet, AI is projected to drive energy demands to an all new high. Be sure to consider things such as increased use of infrastructure software management to drive efficiency as AI will need more power than previously imagined.

5 Work with trusted partners with vast resources and world-class expertise and toolsets

Schneider Electric, a WWT facilities infrastructure (FIT) partner, helps ensure our clients' data centers are ready to handle AI workloads. Put another way, they make sure facilities have the necessary power and cooling to support your AI workloads.

WWT's value-add includes a dedicated team of AI experts with extensive industry knowledge and hands-on experience in design and deployment, focusing on the physical layer that underpins the IT stack.

se.com

For on-premise needs, future-ready the physical infrastructure

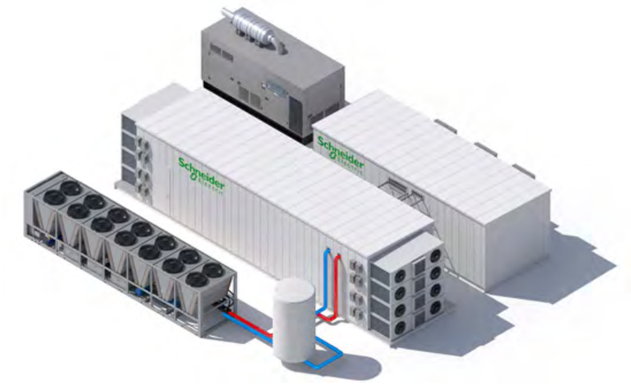
Given that AI stacks are an emerging application, we recommend engaging equipment vendors regarding your application and specific AI power profile. Making data center power and cooling systems future-proofed, flexible, and scalable is crucial given the quickly evolving demands from AI and other high-density computing. This involves a shift towards more adaptable infrastructure. Best practices in physical infrastructure planning and design to help future proof your data center environments include the following:

Design with modular and prefabricated systems

Utilizing modular, prefabricated, and pre-integrated power and cooling solutions streamlines deployments and enhances scalability. They allow organizations to adapt to new requirements without extensive overhauls of existing systems. Modular designs facilitate accelerated, incremental upgrades, so infrastructure can grow alongside technological advancements. It also minimizes upfront capital expenditure and allows for just-in-time capacity expansion. The purpose-built modular infrastructure is customized to specific use cases and can then be scaled into repeatable clusters for deployment as AI needs grow.

Examples include Schneider Electric's micro data centers and prefabricated modules (see **Figure 3**). They also address situations where space is at a premium or simply not available, where you can deploy prefabricated modules outside your building. Or in cases when it would be too disruptive to retrofit a production data center.

Figure 3 - Prefabricated data center design



Embrace liquid cooling

While liquid cooling is often highlighted as a key solution for supporting Gen AI training clusters, those deploying inferencing workloads should be ready for it too. Traditional air-cooling techniques will be suitable in many cases, but not all. The high thermal design power (TDP) of specialized AI chips drives this need. While not all inferencing workloads will require such advanced cooling methods, the more sophisticated complex reasoning models likely will. Cooling infrastructure designs should be adaptable, enabling the transition from solely air-cooling to a hybrid cooling solution (with liquid cooling) as the need arises over time. Liquid cooling comes with the added benefit of efficiency gains, which helps deliver a future-proofed design. See White Paper 133, Navigating Liquid Cooling Architectures for Data Centers with AI Workloads, for more on incorporating liquid cooling into your infrastructure.

Design adaptable power distribution systems

Design power distribution systems with the flexibility to handle much higher power draws per rack than current averages. Implementing busway systems rather than traditional conduit and wire offers easier reconfigurations and increased capacity without significant re-cabling. When specifying rack PDUs, it is best practice to specify ones that can accommodate the higher rack densities of future IT equipment generations. For example, 240 V 125 A (non-IEC) rPDU provides 71.9 kW. Using two of these can support 143.8 kW. Refer to WP126, Retrofitting Existing Power Systems for AI Clusters for more details on impact of high density AI on power distribution.

Implement data center software for real-time visibility and control

Having clusters of high-power density and liquid-cooled IT alongside traditional air-cooled IT means that certain software functions become more critical. Software tools include DCIM, EPMS, BMS, and digital electrical design tools. These tools collectively address the challenges of design uncertainty and operational risk in this dynamic environment. See White Paper 110, How 6 AI Attributes Change Data Center Design, for more details on the importance of implementing software management into your AI designs.

Leveraging vendor expertise

Partnering with experienced vendors provides best practices for model selection and deployment aligned with business goals, providing smoother workflow integration and enhanced performance across software, hardware, and physical infrastructure. This includes physical infrastructure vendors/partners. Use vendors that provide comprehensive solutions, software, and service capabilities. Choose vendors who offer the opportunity to test and get hands-on with the technologies. Since the power and cooling needs differ across deployments, collaborating with a partner who can create a tailored (engineer to order or configure to order) solution is essential. Being able to examine advanced proofs of concept for selected technologies leads to better decisions.

Crucially, these partnerships often unlock valuable, validated AI cluster reference designs that provide your team with some options for deploying your physical infrastructure. These are pre-architected blueprints from proven deployments – significantly accelerating the design and implementation of efficient, scalable Gen AI inference infrastructure optimized for power and cooling, while reducing uncertainty and risk. These can be leveraged as a complete design, or oftentimes, a subset. You may not need to build an entire data center, but you can very much still leverage the designs for specifying the power and cooling distribution of a pod of IT gear. For instance, Reference Design 108 can be used as guidance for how to distribute power to high-density AI racks, using 100% rated 800 A breakers to provide 575 kW of busway capacity.

Next steps

Gen AI inferencing is gaining momentum, with an increasing number of pilot projects underway and production deployments being implemented. To stay on track, businesses should take the following four proactive steps:

1

Assess each AI application's deployment strategy individually

Have your teams (or partners) conduct thorough assessments of factors such as data sensitivity, regulatory requirements, latency needs, cost management, in-house skill sets, service level agreements, and scalability to determine the most appropriate deployment method for each AI application. Leverage the traditional cloud or neoclouds for applications that benefit from the latest technology resources and scalability they offer. When appropriate based on data sensitivity, ultra-low latency demands, or business philosophy, deploy on-premise or in a local colocation facility, near the point of use.

2

Translate AI requirements into physical infrastructure requirements

Map the specific IT demands of your AI workloads into tangible needs for an on-premise physical setup. Essentially, it means figuring out how much power is needed, what the IT rack count and power density is, and what type of cooling the IT equipment requires. Validate these requirements with your vendors and partners in parallel with IT equipment planning to avoid deployment delays.

3

Conduct an AI-readiness assessment of your physical infrastructure

This comprehensive evaluation should involve your vendors, integrators, and the use of specialized software tools like DCIM, EPMS, BMS, and digital electrical design tools to accurately gauge your site's ability to accommodate an AI solution. Assess current power and cooling capacity, rack density capabilities, and available space. When significant gaps in capacity or lack of space are identified, consider prefabricated modular solutions. Furthermore, if the existing grid lacks sufficient capacity for expansion, explore innovative power solutions like those offered by AlphaStruxure. This thorough assessment ensures your physical environment can adequately and efficiently host the AI solution.

4

Upskill on liquid cooling technologies

Many facility operators have not had direct experience with liquid cooling yet. But with TDPs of chips continuing to rise, and the growing spectrum of AI applications that will require these specialized chips, your facility operators will be faced with this technology. Prepare ahead of time through training courses, industry conferences and workshops, pilot programs, and partnerships with your technology providers.

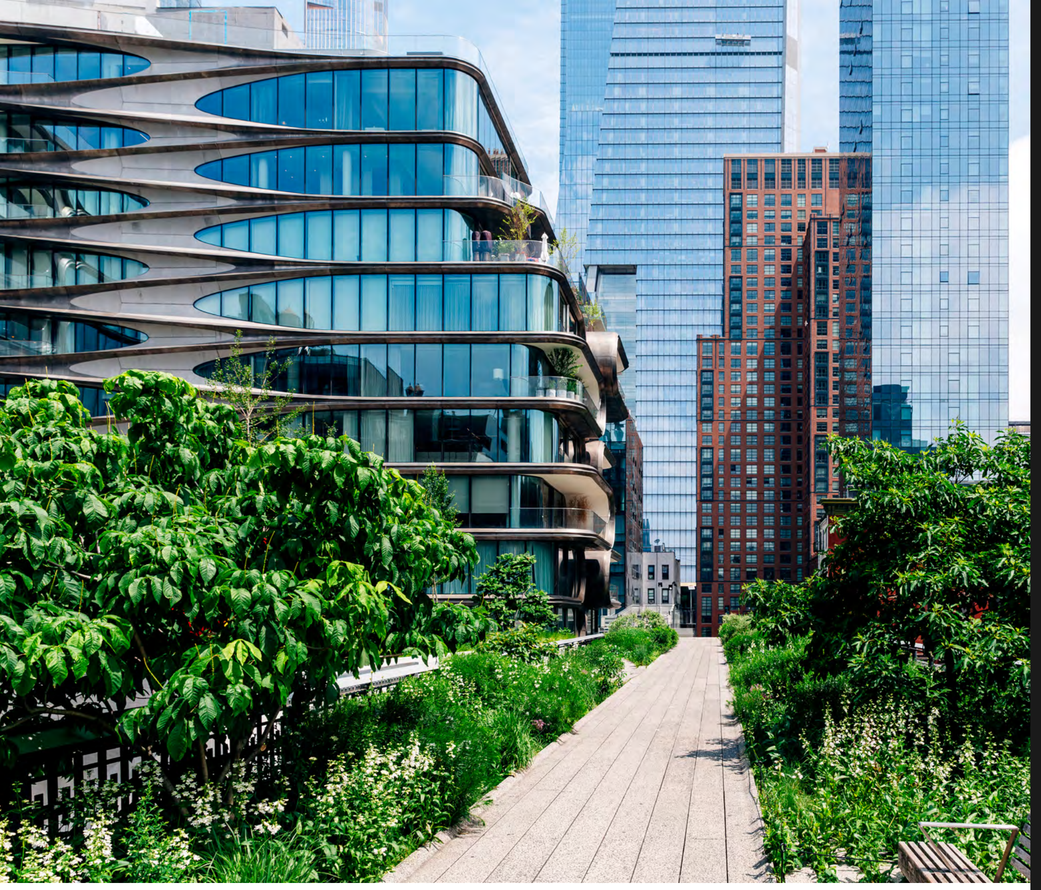


Endnotes

1 From AI assistants to Agentic AI: Collaborating in the Age of Intelligence, CDO Times, Accessed link on August 05, 2025
2 What is RAG (Retrieval-Augmented Generation), AWS, Accessed link on August 05, 2025
3 Azure Open AI Service, Accessed link on August 05, 2025
4 NGC catalog, NVIDIA
5 What is GPU-as-a-Service (GPUaaS) or GPU Cloud?, WWT, Accessed link on September 04, 2025
6 Cloud providers are continuously investing in and improving their networking infrastructure (e.g., faster interconnects, RDMA over Converged Ethernet - RoCE, enhanced network topologies).
7 Surgeons Provide Clarity on Applications for Generative AI in Patient Care, American College of Surgeons, Accessed link on August 05, 2025
8 Understanding Latency in AI: What It Is and How It Works, Galileo, Accessed link on August 05, 2025
9 How Generative AI is Transforming the 3D Printing Industry, Medium, Accessed link on August 05, 2025
10 High-frequency Trading, Colocation, and the Limits of the Speed of Light, Lime Trading, Accessed link on August 05, 2025
11 NIST Special Publication 800-144, "Guidelines on Security and Privacy in Public Cloud Computing"
12 Dell Technologies Unveils Infrastructure Innovations Built to Power Modern AI-Ready Data Centers, Dell, Accessed link on August 05, 2025
13 NVIDIA hardware for inference wokloads, Accessed link on August 05, 2025
14 Amazon SageMaker inference, AWS, Accessed link on August 05, 2025
15 What is Azure AI model inference?, Microsoft, Accessed link on August 05, 2025



se.com



THIS DOCUMENT IS TO BE CONSIDERED AS AN OPINION PAPER PRESENTING GENERAL AND NON-BINDING INFORMATION ON A PARTICULAR SUBJECT. THE ANALYSIS, HYPOTHESIS AND CONCLUSIONS PRESENTED THEREIN ARE PROVIDED AS IS WITH ALL FAULTS AND WITHOUT ANY REPRESENTATION OR WARRANTY OF ANY KIND OR NATURE, EITHER EXPRESS, IMPLIED OR OTHERWISE.

© 2025 Schneider Electric. All rights reserved.

Life Is On | Schneider Electric

Authors



Mike Trojecki

Sr. Director,
AI Practice
World Wide Technology
LinkedIn

Mike Trojecki leads AI infrastructure and go-to-market (GTM) strategy at WWT, helping enterprises turn AI ambitions into operational results. He works with clients to build scalable systems that integrate compute, storage, networking, and data pipelines into production-grade AI. Mike partners with CIOs, CTOs, and business leaders to align AI investments with performance, security, and cost goals. Across various sectors, he shares knowledge of best practices and emerging trends. His experience adds value to clients in (insert core sectors). His work with executives is making AI—especially generative AI—more usable, governable, and profitable. From optimizing workflows to designing effective, efficient and sustainable infrastructures, Mike is working at the forefront of increasing AI's value and practical application. He's known for cutting through the hype to bring clarity and executional excellence to organizations. Mike is a veteran of the US Air Force supporting communications for Air Force One and the White House.



Wendy Torell

Senior Research Analyst
Data Center Research & Strategy
Schneider Electric
LinkedIn

Wendy Torell is a Senior Research Analyst in Schneider Electric's Data Center Research & Strategy group bringing 30 years of data center experience. Her focus is analyzing and measuring the value of emerging technologies and trends: providing practical, best practice guidance in data center design and operation. Beyond traditional thought leadership, she championed and leads development of interactive, web-based TradeOff Tools. These calculators help clients quantify business decisions, while optimizing their availability, sustainability, and cost of their data center environments. Her deep background in availability science approaches and design practices helps clients meet their current and future data center performance objectives. She brings a wealth of experience across Schneider Electric's broad portfolio and with the market at large. She holds a BS in Mechanical Engineering from Union College and an MBA from University of Rhode Island. Wendy is an ASQ Certified Reliability Engineer.