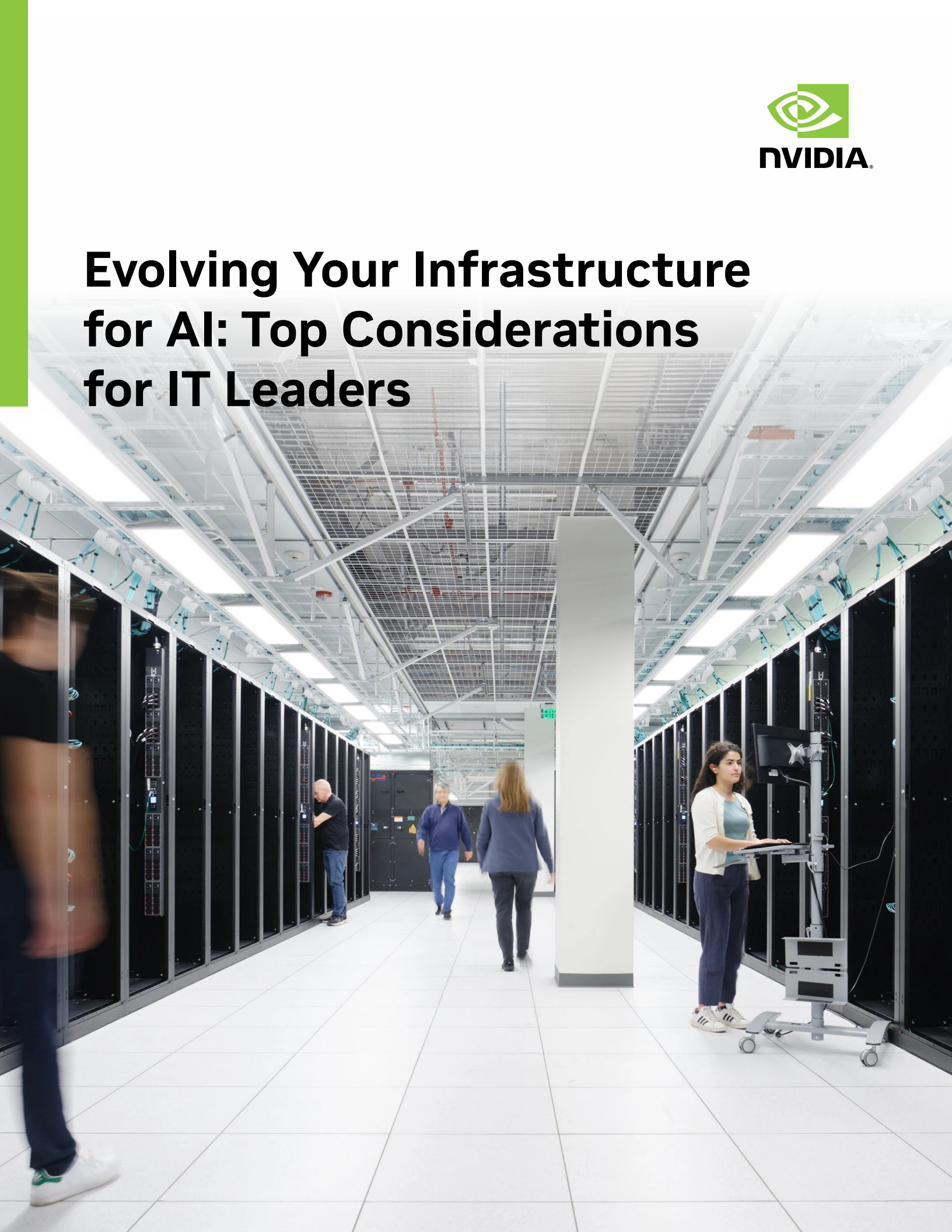
Evolving Your Infrastructure for AI: Top Considerations for IT Leaders



Constructing a scalable infrastructure to support artificial intelligence includes a range of important considerations, from strategic investment decisions to assembling the right team. By addressing these fundamental factors, IT organizations can lay a foundation that maximizes the potential of their AI initiatives.

1. Consider the entire stack and strategize holistically

Fast computation undoubtedly plays a significant role in AI infrastructure, however, any weak link within the overall solution can hinder productivity. IT leaders are increasingly recognizing the importance of a robust end-to-end stack, encompassing storage, networking, and software. This ensures optimal resource allocation and prevents projects and plans that could impact scalability and prolong time-to-solution.

> Enterprises can centralize AI infrastructure and provide immediate, on-demand access for teams by adopting an **AI Center of Excellence**. This enables a holistic approach to AI development and deployment with the ability to reduce risk, scale faster, and deliver better return on investments.

As you think ahead, it's also important to stay up-to-date on the latest advancements in generative AI and its limitless potential. Take advantage of opportunities for your teams to explore and engage with **AI models**, and identify solutions that best suit your organization's needs.

2. Decide on a cloud-first, on-premises, and/or hybrid approach

A cloud-first approach offers quick access to the powerful compute resources required by AI. This equips organizations with both the scalability and flexibility they need to train and deploy AI models, regardless of project size and complexity.

Additionally, the pay-as-you-go model of cloud services eliminates the need for upfront investments for on-premises infrastructure. Enterprises can optimize expenses by paying only for the resources they use, avoiding over-provisioning, and reducing the financial risks associated with underutilized infrastructure.

Moreover, the cloud's extensive ecosystem of AI tools and services empowers organizations to accelerate innovation and bring AI-driven solutions to market faster, without starting from scratch.

In situations where organizations deal with sensitive data or operate in highly regulated industries, data security and compliance become paramount. In such cases, an on-premises infrastructure is crucial as it allows data to remain in-house and tightly secured.

Although on-premises infrastructure may involve higher initial expenses, it provides long-term advantages in terms of reduced operational costs. This cost-effectiveness over time makes it an appealing choice for organizations seeking to retain control over their data while maintaining predictable costs.

Adopting a hybrid approach for AI infrastructure enables enterprises to leverage the scalability and flexibility of the cloud, which allows for better resource allocation and cost management in the short term for supporting pilots, while reducing long-term costs on-premises for established models that are ready to scale. Moreover, this approach also allows non-sensitive workloads to be offloaded to the cloud, while ensuring the protection of sensitive data to meet compliance requirements within a self-hosted deployment model.

3. Invest in an accelerated AI infrastructure

AI requires a departure from traditional corporate IT infrastructure, as it calls for specialized hardware, software, and AI algorithms that heavily rely on parallel processing and the power of accelerated computing. Conventional, non-accelerated data centers cannot effectively handle the demands of AI workloads, which often involve processing and analyzing vast amounts of data that can be accessed quickly.

Modern AI infrastructure requires high-capacity, high-performance storage solutions capable of efficiently storing and retrieving large volumes of data. Consequently, it becomes imperative to build a dedicated infrastructure specifically tailored for AI, rather than trying to repurpose existing infrastructure. An accelerated infrastructure and AI software for optimization are necessary across the AI pipeline—from data prep, to training, customization, and deployment.

4. Foster the growth of your team and cultivate AI proficiency

With global technical staff and IT skill shortages, building a dedicated team with expertise in AI infrastructure can be a challenge. According to IDC,* 52% of global organizations are already experiencing the negative impacts of technical skill shortages.

There are many **AI training and certification programs** available to help your team develop key skills and gain hands-on experience. Some solutions offer deployment and management services, enabling organizations to focus on business objectives, rather than building and managing infrastructure. If given the opportunity to hire more staff, IT leaders should consider those with expertise in managing infrastructure and cloud platforms, particularly those with a strong understanding of cloud technologies. Those skilled in DevOps practices and automation tools should also be considered. These individuals can streamline the deployment, monitoring, and maintenance of AI projects, providing smooth operations and minimizing downtime.

5. Weigh budget considerations with long-term AI goals

Investing in infrastructure that'll work with unknown, future workloads is a crucial part of a long-term AI strategy. And with accelerated computing—which uses parallel processing on GPUs—demanding applications are sped up while increasing energy efficiency and cost savings in the long run.

Cloud-based solutions offer a cost-effective way to start AI initiatives by reducing acquisition costs and shifting capital expenditures (CapEx) to operational expenditures (OpEx). Yet, while cloud solutions may have lower initial costs, long-term expenses can add up. IT leaders should evaluate the total cost of ownership (TCO) over time and consider factors such as data storage, compute resources, and ongoing maintenance.

In general, it's important to consider return on investment (ROI) as a key metric rather than the initial TCO. Building AI infrastructure requires dedicated resources, careful planning, and consideration of cloud and on-premises solutions. By leveraging the right blend of technology and strategy, organizations can navigate the challenges associated with building AI infrastructure and drive successful outcomes.

* Source: IDC, Skills Forward: a 2023 IT Skills Shortage Survival Guide, DR2023_LL2_AL_GS, March 2023

Learn how NVIDIA can help future-proof your infrastructure and ensure that you're AI-ready.

Get Started with NVIDIA Leading Enterprise Solutions

Generative AI

Transformative innovation for organizations worldwide.

AI Inference

Faster, more accurate AI model deployment—from anywhere.

AI-Powered Cybersecurity

Zero-trust, real-time threat detection at scale.

NVIDIA Training

Online courses and instructor-led workshops for your teams.

Data Analytics

Accelerated analytics solutions from desktop to data center.

Request a Consultation

To learn more about how NVIDIA can help you address your business challenges, visit: **nvidia.com/executives**

Talk to an expert about your business goals.

NVIDIA