

AHEAD

AI INFRASTRUCTURE READINESS:

Adapting Data Centers for the Workloads *of the Future*



The AI boom is rippling through digital infrastructure,

and many data centers are unprepared. AI has already become a larger percentage of data center workloads, with training and fine-tuning large language models, inferencing, and high-density workloads taking up nearly 20% of capacity. However, at least two times the amount of new data center capacity is needed than exists currently.

The size and complexity of AI workloads is also increasing dramatically, as breakthroughs related to generative AI and large language models (LLMs) require even more computing resources than traditional AI. Newer generative AI models are pre-trained on enormous amounts of data — 45 terabytes in the case of [OpenAI's GPT-3 model](#) — which requires incredibly powerful hardware and supporting infrastructure.

Organizations will need to optimize their current infrastructure but also prepare for the future demands of AI as it continues to evolve. In this guide, we'll explore the impact that AI workloads are having on traditional data centers, and some strategies to modernize infrastructure for AI.



How AI Workloads Are Impacting *Data Centers*

Data centers are facing unprecedented challenges with AI workloads requiring specialized hardware that have greater computing power, but also higher energy consumption and heat generation.

HIGHER POWER CONSUMPTION

Today's cutting-edge data center components have much larger power draws than in the past, and it only continues to grow. In fact, [Goldman Sachs](#) predicts that power demand will grow 160% by the end of the decade as power efficiencies decelerate and AI demand continues to build.

Although performance per watt has increased, newer CPUs and GPUs still have much greater energy demands than previous generations. This means data centers will need to find a way to cost effectively support these power-hungry hardware components.

INCREASED THERMAL HEAT GENERATION

The latest components also have higher thermal design powers (TDPs) — or maximum potential heat generated — which is difficult to dissipate. Traditional air-cooling strategies, once effective for lower power draws of up to 20kW, are now struggling to cope with the heat generated by the latest computing hardware exceeding 30kW and more.

AI and machine learning workloads also primarily run on GPUs, which operate optimally at much higher temperatures, typically exceeding 80°C (176°F). This necessitates innovative cooling approaches that can effectively manage the heightened thermal demands of cutting-edge components running AI workloads.

GREATER HARDWARE DENSITY

Finally, the shift to high-density components and data centers is just beginning and will be transitioning for several years. This densification of IT hardware further exacerbates heat dissipation and energy challenges, with more powerful computing capabilities being packed into smaller and smaller enclosures. In fact, a cutting-edge server now contains multiple high-end CPUs and GPUs that leverage increased die densities and other hardware optimizations.

As a result, many racks are only half populated with higher density servers before they reach the limit of data center power supplies and cooling capabilities. This means data centers will need to rethink their designs to adapt to accommodate greater hardware density.

Strategies for Adapting Infrastructure to AI Workloads

Here are some ways data centers can optimize their current infrastructure and prepare for the future.



Implement heat dissipation and cooling options

As mentioned before, traditional air cooling is no longer sufficient for most modern workloads. Air cooling is less efficient because it relies on air conditioning and fans that cover entire rooms or rows rather than specific heat sources. Liquid cooling is a promising alternative to traditional air cooling because it provides targeted cooling directly to data center components, improving efficiency and reducing overall energy consumption.

Although liquid cooling has been used in the consumer computing space for a while now, especially for high end gaming computers, it's now starting to become more popular for enterprise-grade deployments. Liquid cooling is usually more efficient than air cooling and might even be a requirement for future generations of CPUs and GPUs.

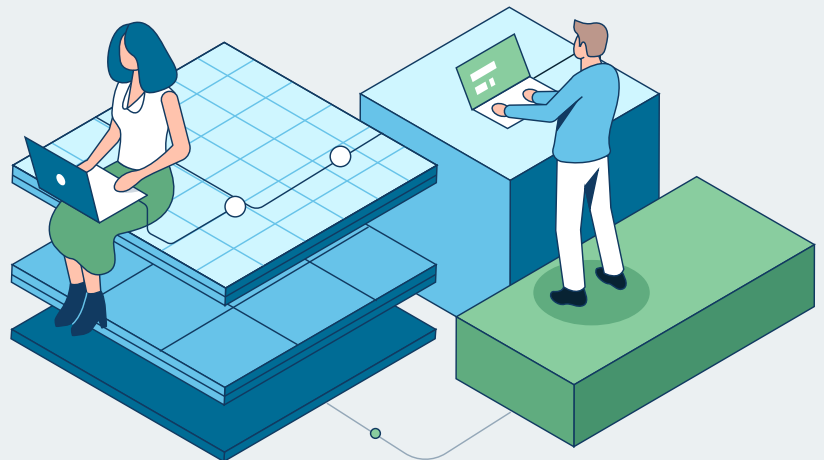
While retrofitting an entire data center with comprehensive liquid cooling infrastructure can be cost-prohibitive, bolt-on liquid cooling solutions offer a more practical and flexible alternative. Hybrid approaches combining air and liquid cooling are especially attractive, allowing targeted cooling for high-density components without extensive facility upgrades. For instance, direct-to-chip liquid-to-air cooling systems can be implemented at the rack or server level using closed-loop designs that dissipate heat via rack or row-based liquid to air CDUs, eliminating the need for facility water, pumps, and chillers. Similarly, rear-door heat exchangers provide rack-level liquid cooling by replacing standard rack doors with self-contained systems that absorb and transfer heat efficiently.

Consider servers with ARM® processors

Although most large-scale workloads currently run on X86 processors from AMD and Intel®, ARM® processors such as NVIDIA® Grace Hopper™ and the new Grace Blackwell™ are growing in popularity. [ARM®-based hardware](#) often has significantly reduced energy consumption without sacrificing performance, and it is also usually smaller and produces less heat, which reduces the electricity requirements for cooling compared to larger X86 alternatives.

The lower power draw from the hardware itself and its supporting infrastructure means ARM®-based servers that are constantly running within data centers can see enormous energy and cost savings over the long term. As the environmental impact of AI workloads gains more attention, ARM® processors could be a more sustainable alternative as well.

ARM® processors have been widely used in the majority of smartphones for a long time because they can achieve a greater performance per watt than alternative processors. Since Grace Hopper™ and Grace Blackwell™ superchips bring this efficiency to broader enterprise AI use cases, many leading companies have been able to accelerate performance with lower energy consumption.



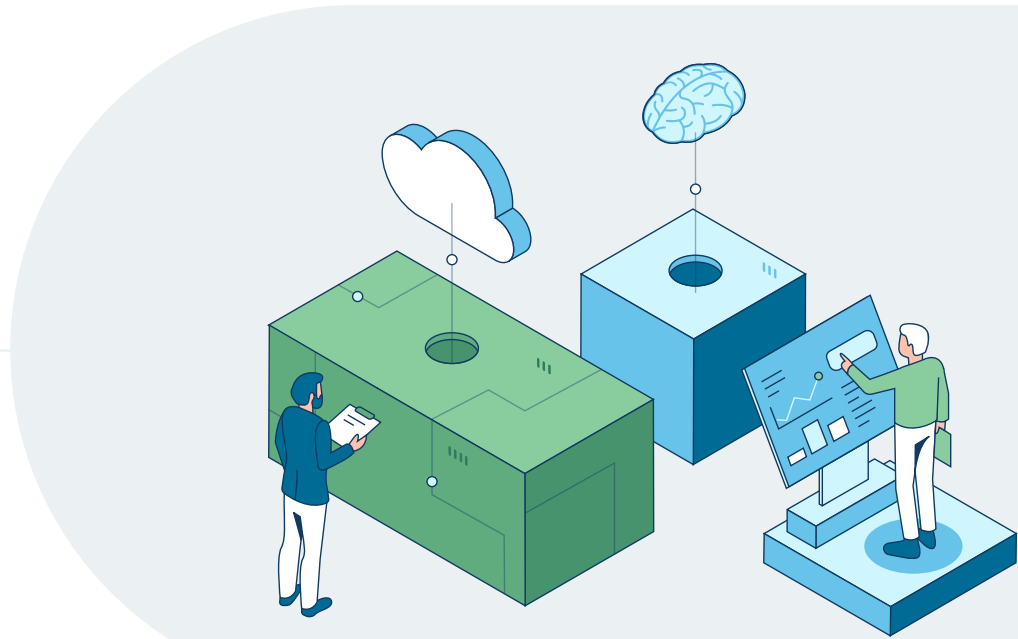
Shift workloads to distributed infrastructure

Rather than only focusing on increasing centralized data center capacity, organizations can consider a hybrid approach that shifts workloads to the cloud or distributed infrastructure. In fact, a hybrid approach with near edge or far edge infrastructure can free up data center capacity for AI inferencing workloads that require real-time decision, data localization, and enhanced privacy.

As the demand for real-time analytics grows, there will also be a need for optimized and efficient edge computing solutions to support low-latency inferencing. That's why [Gartner](#) predicts that 75% of enterprise-generated data will be created and processed outside a traditional centralized data center or the cloud by 2025. For example, [recent hardware and software innovations](#) are driving adoption for both near and far edge computing.

Near edge infrastructure is generally a more traditional data center located closer to end users, which can add capacity for AI workloads and alleviate some of the strain on traditional data centers. Far edge infrastructure are devices or micro data centers deployed in the physical location where data is generated and often in very remote environments in the case of [industrial edge](#).





Adopt AIOps for infrastructure management

The growth of distributed systems, containerized workloads, hybrid computing, and other approaches for deploying advanced AI workloads requires more flexibility and adaptability to manage than legacy applications and workloads. This means traditional IT operations processes are no longer enough to maintain data center performance and resilience.

For these reasons, many organizations are starting to integrate AI and machine learning technologies into data center operations to help ease the burden of advanced workloads. This newer approach called AIOps enables organizations to analyze data center workloads and adjust resource utilization in real time to optimize their infrastructure. AIOps can increase efficiency, and in turn, reduce infrastructure costs by minimizing waste and helping organizations maximize their existing data center resources.

However, organizations that begin adopting AIOps practices will likely need to modernize their data center infrastructure. AIOps requires sharing data from different components, but some legacy hardware and software systems are not designed for interoperability. This means organizations that want to integrate AIOps into their data center operations may need to consider new systems that provide open APIs and SDKs. By proactively upgrading data center components for AIOps, however, organizations will be able to better handle current AI workloads and reduce infrastructure costs in the long run.

Data Center *Modernization* with AHEAD

AHEAD is a leading provider of enterprise solutions that accelerate the impact of technology investments. We can help you [modernize your data center infrastructure](#) and adopt AIOps to handle the demands of AI workloads. Our diverse team of hardware and AI experts can design and implement infrastructure solutions that are more reliable, resilient, and secure.

More specifically, our engineers can customize enterprise-grade solutions for AI using servers with ARM® processors or other more energy efficient computing options to reduce the costs and thermal generation of data center infrastructure. We can also help you evaluate your cooling requirements and select the right liquid-cooled and air-cooled hardware to easily and cost-effectively dissipate heat.

In addition, [AHEAD Foundry™](#) is our integration facility with comprehensive services for building and configuring hardware and networking infrastructure at scale rather than on-site. This means we can help you design and deploy large-scale edge solutions, far edge systems, or modernize your existing data centers using a plug and play approach.

[Contact AHEAD](#) to learn more about our data center modernization services and take the first step toward preparing your infrastructure for an AI-centric future.



Combining cloud-native capabilities in software and data engineering with an unparalleled track record of modernizing infrastructure, we're uniquely positioned to help accelerate the promise of digital transformation.

[Contact Us](#)

National Hubs

CHICAGO

444 W. Lake Street
Suite 3000
Chicago, IL 60606

NEW YORK

500 Fifth Avenue
Suite 1500
New York, NY 10110

ATLANTA

1117 Perimeter Center
W406
Atlanta, GA 30338

SAN FRANCISCO

2000 Crow Canyon Place
Suite 250
San Ramon, CA 94583