kyndryl

Artificial Intelligence

Infrastructure & Power Challenges

Introduction

As the world's largest managed infrastructure services provider, Kyndryl has a unique perspective on emerging technologies and the global forces shaping how they are deployed.

The need for a steady hand and experienced advisor is more important than ever as enterprise technology leaders and governments seek to navigate a challenging IT landscape. We are experiencing great technological progress, marked by the rapid rise and democratization of artificial intelligence tools and solutions, all promising to deliver business and societal value at scale.

This proliferation has led to an increase in demand for computational power and storage capacity, which strains traditional data centers. AI algorithms require vast amounts of data to train and operate effectively, and that necessitates large-scale infrastructure to process and store this information efficiently.

At Kyndryl, our more than 30 years of IT infrastructure expertise — along with our partnerships with hyperscalers, data center operators, and our customers — enables us to provide unique insight into the challenges enterprises face with regard to data center modernization. That includes aspects of compute infrastructure, data, and workloads that will be needed to meet future AI needs.

Impediments to AI Acceleration

There are more than 2,700 data centers the United States alone, carrying out processing workloads that drive the U.S. economy forward. Virtually every industry is confronted with three primary challenges to the expansion of AI and the myriad opportunities it promises to unlock. These must be addressed to create the environment for expansion.

Computational Demands: Generative AI applications require significant computational power for training complex machine learning models, which can exceed the capabilities of traditional data center hardware.

Cooling Requirements: If the expansion of AI relies on the installation of more Graphics Processing Units (GPUs) for processing intense workloads, the insides of data centers will need to be reimagined to accommodate denser racks, each filled with GPUs that require vast amounts

of power. Greater rack density and more power means more heat will be produced, such that traditional air-based cooling methods won't suffice. Liquid cooling appears to be the most prominent answer to this challenge, but this space is still in need of greater innovation to meet a rapidly increasing demand.

Energy Efficiency: In some regions, traditional data centers already strain power grids. AI processes are even more energy intensive. This is prompting important questions around how and where to build data centers — and also how to power them sustainably while keeping operational costs in check.

Industry Trends in AI Infrastructure

Because Kyndryl manages the mission-critical IT infrastructures of our customers, our experts have unparalleled insight into what direction industries are moving. And because we work closely with hyperscalers and data center operators, we have insight into how data processing infrastructure is being reshaped to accommodate emerging technologies. Trends we are watching include:

AI is Driving New Demands for Data Centers

The rapid adoption of generative AI is driving the upward trajectory of rack power density, or the amount of computing equipment installed and operated within a single server rack. Average rack density[1] has been slowly climbing over the past few years and will see significant jumps in the coming years. Few enterprises will build their own AI-ready data centers due to high costs and risks. Instead, they will purchase these services, providing a business opportunity for hyperscalers, who in turn will have more demand for new data centers.[2]

Energy Projections are Shaping AI Infrastructure Business Decisions

As enterprises integrate more advanced AI algorithms and machine learning models into their operations, the need for scalable and efficient data center infrastructure becomes paramount. This shift is driving innovation and new approaches to energy-efficient hardware to handle the computational demands while minimizing power consumption. This also requires a rethink around cooling technologies, and optimizing data center layouts to manage the heat generated by high-density computing systems. These solutions would demonstrate that power density and power efficiency need not remain at odds with each other. Critical to this effort, however, will be continued communication between industry, utility companies, and policymakers to ensure regulation doesn't impede innovation and that industry is leveraging tools aimed at removing bottlenecks (e.g. Inflation Reduction Act (IRA) loan mechanisms and grants).

The Underlying Grid Data Centers Rely Upon Needs a Rethink

The cost and availability of power remains a chief concern when thinking about future data centers, which tend to be clustered in places with established networks and access to a plentiful energy supply. While nascent conversations around the development and use of onsite power sources, such as small modular nuclear reactors (SMRs), are taking place, their widespread use will not be in the near-term. To be sure, data centers don't need to upgrade in order to be powered. The underlying energy system that transmits electricity to data centers does. At a high level, new energy sources — be they natural gas plants, wind, solar, nuclear, or hydrogen — need transmission lines connecting them to the backbone of the power grid. Currently, that backbone needs investment to increase its capacity and resilience. It needs to be bigger and more capable to accommodate more connections. Finally, data centers need upgraded ligature connecting them to the grid, allowing more power to flow to them. Further, by leveraging the next generation of battery energy storage systems (BESS), stakeholders can better manage increasing demand and capacity requirements.

Location, Location, Location

Cost typically plays the biggest role in determining where new data centers are located. Regions that can offer attractive tax incentives and low energy costs therefore make prime spots for new builds. Also important is the ability to easily connect to power, proximity to users, early community relationship management, and skilled labor availability. The demands are higher with AI, prompting hyperscalers to explore building data operations abroad. But opportunities still exist to invest resources domestically. One trend gaining ground is the

[1] Rack density is calculated by the rack's power requirements (in watts) divided by the available space (measured in rack units).

[2] Currently, large hyperscaler facilities have an estimated average density of 36kW per rack. To meet demands, that number is estimated to grow to 50kW by 2027. Many AI cluster requirements are projected to hit 80-100kW per rack.

potential for industry leaders to retrofit existing, unused data centers to respond to the demands of AI infrastructure while investing in local communities.[3]

Additional Energy and Efficiency Considerations

While there is much work to be done toward generating electricity and upgrading the grid that delivers it to data centers, there is onsite work to be done that can yield worthwhile results. Because Kyndryl has unique expertise in increasing the efficiency of data centers, it is clear there is wasted capacity to be reclaimed for growth.

Typically, data center energy use is measured by PUE, the ratio of total facility energy to IT equipment energy used. In Kyndryl's experience, taking steps to improve the PUE of a data center can make available additional power — often up to 30%. At scale — and in concert with efforts to upgrade the underlying grid — this can make a sizable difference.

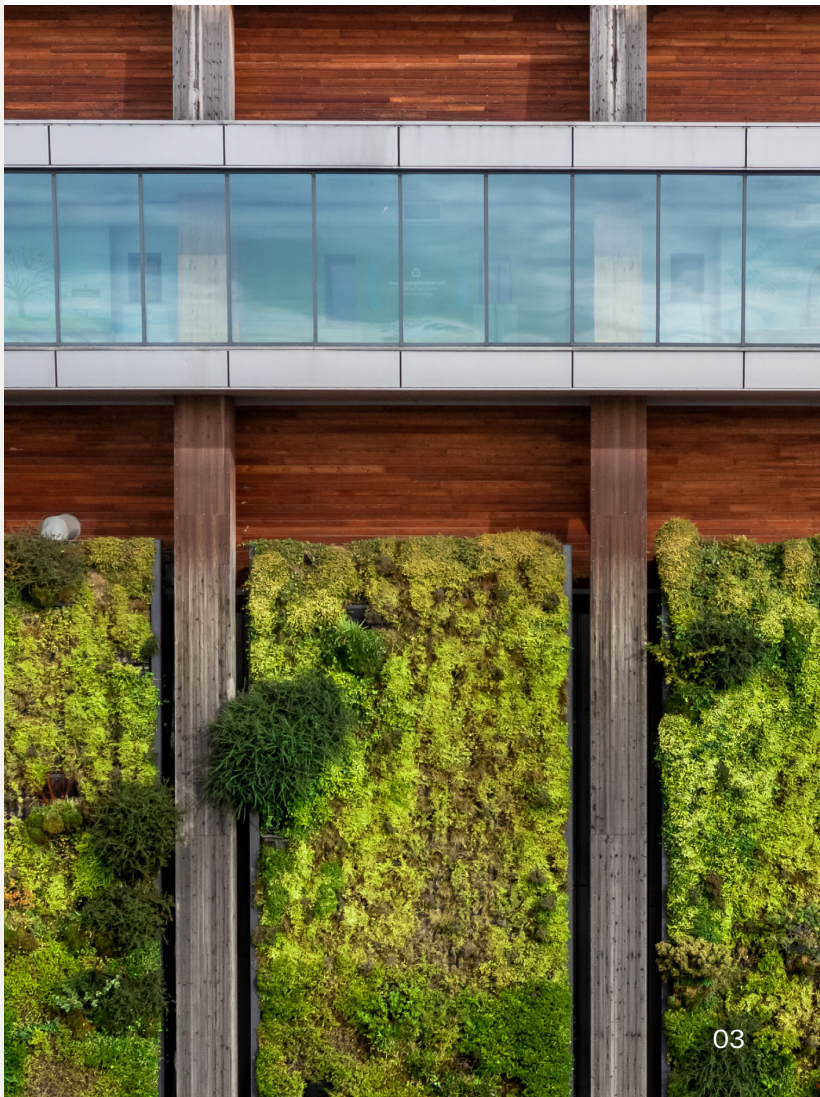
Data centers currently account for about 1% of the world's electricity, and AI currently contributes to a small fraction of this. Even with significant growth, AI's energy demands are unlikely to overshadow other electrification drivers, such as electric vehicles and industrial processes. Moreover, advancements in GPU technology will significantly improve energy efficiency. These innovations, coupled with improvements in hardware architecture, software, and algorithms, promise continued gains in efficiency, mitigating the potential energy impact of AI.

The new AI models are power-hungry, and they will add to the electricity consumption by data centers.[5] However, no consensus has emerged on the size of this increase and whether some estimates around electricity demands are overblown.[6] Historical forecasts[7] of exponential IT energy increases, such as those related to streaming services, have often proven inaccurate due to efficiency improvements.

It is crucial for industry leaders to share data and efficiency considerations with government stakeholders shaping policy around AI. This coordination is necessary to drive cutting-edge advancements while setting global standards around security and sustainability.

Conclusion

Kyndryl is committed to operating at the heart of progress, where innovation, environmental stewardship, and social impact converge. We have committed to be Net Zero in our greenhouse gas emissions by 2040. And we proudly work with enterprises to help them achieve their environmental, social, and governance goals. By addressing the challenges and leveraging the opportunities presented by AI and data center modernization, we aim to lead the way in sustainable and efficient technology solutions.



[3] In May 2024, Microsoft, in conjunction with the White House, *announced* a $3.3 billion investment to build an artificial intelligence data center near Racine County, Wisconsin. The location for data center is on the same 1,000-acre spot as a previously proposed project by Foxconn, an electronics company, that never came to fruition. As part of that unsuccessful project, the state of Wisconsin had already invested hundreds of millions of dollars expanding Interstate 94 and nearby highways, and also bringing utilities to the site. Local utility company, We Energies, *plans* to spend more than $350 million on a distribution project for Microsoft.

[4] Nvidia's latest GPU, for example, uses 73% less electricity to train AI models compared to its predecessor.

[5] According to recent studies, AI-powered Google searches could require more than ten times the electricity of standard searches, with projections ranging from 7 to 9 watt-hours per search request. Studies also show Google's new AI models demonstrate substantial efficiency improvements over previous iterations, reducing energy consumption significantly while maintaining or improving performance.

[6] Schneider Electric's 2023 *white paper* estimated that AI workloads accounted for 8% of the estimated 54 GW of electricity used by data centers last year. This share is expected to rise to 15–20% by 2028 when data center demand is forecast to exceed 90 GW.

[7] *The growing energy footprint of artificial intelligence*, Alex de Vries, October 2023

kyndryl