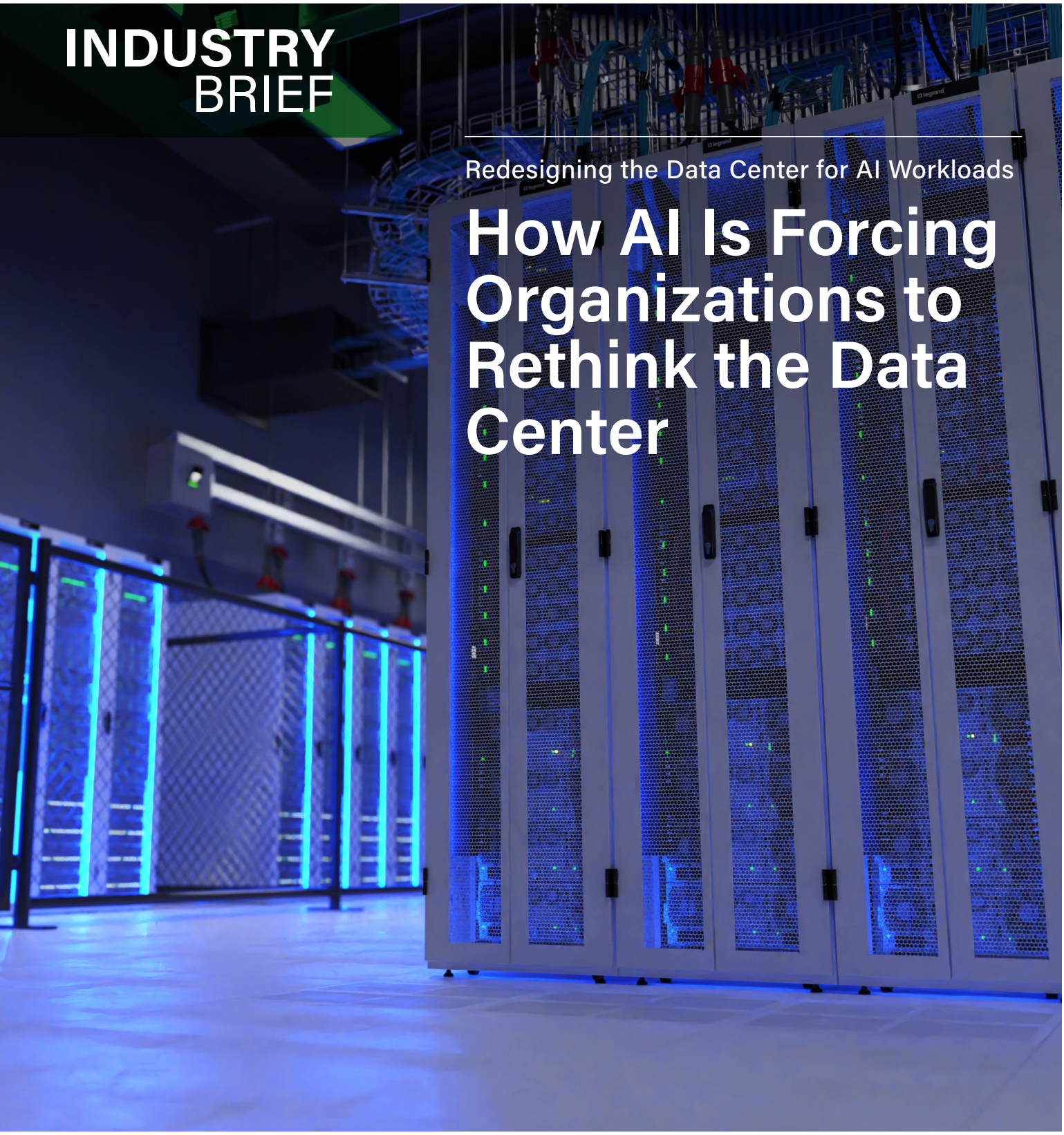
Raritan | Server Technology

INDUSTRY BRIEF



Redesigning the Data Center for AI Workloads

How AI Is Forcing Organizations to Rethink the Data Center

#LegrandImprovingLives

legrand®

EXECUTIVE SUMMARY

The explosive advent of the AI universe is revolutionizing the data center industry in dozens of ways. From the smallest edge data closets to the largest hyperscale data centers from tech giants like Google and Meta, the power consumption, heat generation, and compute density of hardware powering AI applications and their workloads demand a radically different approach to data center design, construction, and operation.

There's no consensus, however, about the emerging needs of next-generation data centers. Dozens of vendors, from startups to Fortune 500 companies, all tout their perspectives. Though many views on the needs of data centers for accelerated computing overlap, they're often different and sometimes in conflict. It's important to take in diverse voices when attempting to make AI infrastructure design and procurement decisions with implications that will last a decade.

At Legrand, we're not just a supplier for and collaborator with the world's largest enterprise data centers, colocation companies, and hyperscalers. We are data center insiders deeply embedded in the cutting edge of data center design. If you have a role in designing and building data centers supporting AI infrastructure, some of this paper's elements will be familiar, and others will be unfamiliar, but they're all proven in deployments worldwide and represent a new frontier in data center design. We're here to reveal these cutting-edge approaches, opening a world of possibilities for designing and building data centers for AI, machine learning, neural networks, deep learning, and generative AI applications.

EXPLORING THE NEW LANDSCAPE

The introduction of next-generation generative AI platforms, such as the groundbreaking release of ChatGPT in 2022, has been a significant catalyst for change in the data center industry. It introduced a new class of IT workloads—generative AI with high computational demands—leading to countless new companies and explosive growth in a market that didn't exist a few years ago, transforming the data center space at an accelerated rate.

Investment in AI is a movement surging at an unprecedented pace across industries.

- According to Goldman Sachs Economic Research, global AI investment could reach $200 billion USD by 2025.
- Market leaders like Meta, Amazon, Microsoft, and Google have announced tens of billions of dollars in new AI data center investments.
- Colocation companies and real estate developers are also jumping on the AI bandwagon, underscoring the widespread adoption of AI and the urgent need for AI data centers.
- Nearly 60,000 new AI startups have been launched from 2020 to 2023.

While existing data center capacity can support some generative AI applications, the most intensive workloads demand IT hardware that rivals high performance supercomputers. Compute capacity, specialized silicon, specialized networking, and the highest-performance storage are combined across hundreds or thousands of servers to deliver the performance needed to stay competitive in the rapidly evolving marketplace, where less efficient, less capable solutions are quickly left behind.

Because of the shift toward AI, the IT infrastructure needed to support these applications is increasingly hot, dense, and power-hungry. Deployed at a massive scale, the shift toward accelerated compute infrastructure brings both possibilities and challenges for organizations aiming to stay on the leading edge of innovation. To cope, data center providers are changing everything: designs, construction methods, cooling technologies, monitoring, and operations to meet these evolving needs. Legrand sees this in the rate of data center construction; Dell'Oro Group has predicted data center CapEx will surge to a 24% compound annual growth rate (CAGR) by 2028 due to "surging demand in AI-related data center infrastructure."

Legrand works with companies of all sizes who have deployed or are about to deploy AI infrastructure to rethink what's possible in the data center. By supporting their distinctive approaches to AI with a mix of off-the-shelf and custom technologies, Legrand can help realize benefits and advantages while mitigating the complexities and risks of AI infrastructure. Starting with the rack, we can assist organizations in optimizing existing data center white and gray space while contributing ideas to next-generation data center designs.

The data center industry is experiencing an exciting transformation. This transformation presents a unique opportunity for data center professionals to collaborate and shape the future of this critical infrastructure in the age of generative AI. In this brief, we will share our unique, insider perspectives on how the data center industry is changing and how we've built new, distinctive products to help organizations move forward faster into the world of AI.

DESIGN SHIFTS FOR AI INFRASTRUCTURE

As Legrand collaborates with leading organizations worldwide, we see shift after shift; old design assumptions and workflows are disappearing and being replaced by new ideas and technologies. Some of the trends we're seeing include:

Rising Infrastructure Costs

Before the advent of generative AI, the data center industry focused on efficiency at every stage, from design to deployment. Now, in today's AI world, while efficiency is still top priority, there is a focus on uptime and functionality for racks full of servers.

Suppose a company wants to power a leading-edge generative AI implementation. In that case, one rack fully populated with NVIDIA® DGX H100 AI servers, for example, can cost hundreds of thousands of dollars and may generate millions of dollars in revenue. Racks of these servers are integral to delivering AI-driven services and core assets for driving business revenues—if they can remain fully functional with minimal downtime. The old economies of the data center are evaporating in the face of these costs. Every element of the data center, from cabling to racks to power and cooling, must be the best available because if something simple fails, millions of dollars of sellable AI capability—and their ability to meet SLAs—is compromised. The costs of building a new data center can now reach $6 to $8 million per MW and continue to rise.

Data Center Size

Three factors are driving the explosion in data center size:

- Most AI training clusters have thousands of servers, and data center providers are doing everything possible to maximize how many servers they can support within a given footprint.
- Because AI infrastructure designs exceeding 200kW per rack are so hot and dense, new data centers require enormous amounts of real estate to accommodate power and cooling equipment.
- There's a worldwide shortage of data center capacity, and providers want to meet the need. Vacancy levels are at a record low.

Not that long ago, 50 MW data centers were enormous. Today, one prominent real estate company reports that its standard transaction size is 300 MW, and it is getting inquiries about building data centers up to 1GW. Much of this size is due to just how demanding these servers are and how much pressure they put on electrical supply, distribution, and heat management.

Power and Cooling

One statistic illustrates how much power and cooling must change to support AI infrastructure. One NVIDIA DGX AI server is 10kW—one server. AI infrastructure power consumption is commonly over 60kW per rack, and some Legrand customers are deploying more than 100kW per rack. One colocation company even reports supporting up to 300kW per rack. These densities have many implications for design, including:

- Power: Today, the normal power design for an AI data hall is multiple drops of 400 Volts 60 Amp power in the US, and 400 Volts 63 Amp power in Europe. Having said that, Legrand has received requests from leading organizations in the AI, high performance computing, and cryptocurrency spaces to provide more than 125 Amps to the rack. Not all facility sites have the power access or power quality to reach these numbers, so power is becoming more of a factor in site selection. Legrand even consults with organizations with power challenges to ensure their power supply can support tomorrow's demanding AI hardware.
- Power Distribution: The intense power demands at the rack require additional power distribution. It is becoming normal to see three, four, or even six rack PDUs in each rack, which has implications for maintenance and serviceability. Overhead power distribution is also the new normal because AI infrastructure requires a lot of power and cabling, including new requirements for east-west cabling that cannot be supported with underfloor legacy wiring conduits.
- Cooling: Air cooling alone is insufficient to cool these racks. Cabinets must guarantee unimpeded airflow and optimal containment and facilitate close-coupled cooling. Many data centers augment air cooling with liquid cooling through in-row cooling/rear door heat exchangers or direct-to-chip.

These changes will force providers to reconsider how they power their data centers. Data center power consumption in the US is expected to rise from 200TWh in 2022 to 260TWh in 2026, accounting for six percent of all power use nationwide. Data-center-owned natural gas energy production, fuel cells, on-premises renewables, and even nuclear power are being evaluated and tested by market leaders.

THE NEW AI RACK

In legacy data centers, racks were furniture, a commodity, or an afterthought. Today's racks and cabinets for AI infrastructure are very well thought out and engineered and must fit the demands of AI infrastructure. Some critical considerations include:

- **Weight**: A single NVIDIA DGX H100 server, for example, weighs nearly 300 lbs. Today's AI racks must support new weight distribution from components like rear door heat exchangers, in addition to the total weight, fully populated, of more than 3,000 lbs.
- **Height**: The old 42U racks are fading; organizations want to leverage every vertical rack unit they can. 48U, 51U, even 52U racks are in demand. In legacy data centers, access and egress are challenges. Cabinets are typically 42 , 47 and 52 U heights.
- **Width & Depth**: The old 24-inch (600mm) wide racks were built to fit standard 2 ft x 2 ft (600mm x 600mm) floor tile. With different flooring, we are seeing demand for different racks to maximize the number of chips in each rack unit. Some customers want a purpose-built rack, with one customer requesting a 36-inch (914mm) width. Integrating wider, deeper racks alongside conventional racks can be challenging, as enterprises and colocation providers are in a difficult position due to their reliance on existing hall space.
- **Sensors**: AI environments are so dense and hot, running at the limits of hypothetical performance, and must be operated within strict condition thresholds to maintain efficiency. Many data center providers use sensors to detect hot spots, airflow gaps, humidity problems, water/leaks, and vibrations. By consolidating sensor data through daisy-chained connections to intelligent rack PDUs, organizations can aggregate data to identify emerging problems, locate hot spots, and improve operational efficiency.
- **Floor designs**: Depending on its pounds per square foot (PSF) rating, traditional raised flooring may not support the weight of today's racks. Many providers are putting AI infrastructure in greenfield containers or greenfield prefabricated halls because slab floors are becoming the default to handle hundreds of thousands of pounds of IT equipment. Not all slab floors, however, can handle the weight of a fully populated AI rack. We have seen racks cut trenches or divots into concrete slab floors.
- **Customization**: Though many data center providers are satisfied with stock products, others have distinctive design requirements that must be met. Legrand is experiencing a significant increase in demand for tailored data center products and solutions, particularly in customized containment. This rise is driven by the specific challenges faced by individual data centers, in that each location is different and requires a unique approach.
- **Integration**: Increasingly, specialized rack and stack integrators are building integrated rack-scale systems, cabling and testing them, and then crating and shipping fully populated racks to the data center. This helps reduce the time needed to deploy large installations.

MARKET SEGMENT DIFFERENCES

It's also worth noting a few differences occurring by market segment. We're seeing that:

- Colocation companies are wrestling with demands for AI because much of their capacity can only support 12-15kW per rack, so for AI workloads, it's essentially not fit for purpose. They're actively driving greenfield builds for AI and are gradually repositioning themselves as AI-as-a-service providers. However, the economics of those AI data centers, and service provider businesses, radically differ from what they're accustomed to.
- For now, many enterprise customers are sitting on the sidelines, waiting for a market consensus around infrastructure designs and operations management. If they do an AI project, they're likely partnering with a hyperscaler or colocation provider because a limited amount of talent and resource supply shortages compromise their ability to move quickly on their own.
- Hyperscalers are building everything they can build, as quickly as they can build it. They're also doing everything they can to tie up vendor supplies, because taking supply out of the market improves their position in the race toward universal AI.

CONCLUSION

All these changes, taken together, paint a picture of a rapidly changing data center landscape. From power considerations and environmental challenges to new demands on racks, data center providers of all sizes are already exploring the future. To get where the world of AI is taking them, they're leveraging trusted partners who can help them move from legacy approaches to the leading edge of data center design.

Fortunately, Legrand is one of those trusted advisors. We work with the world's most successful data center providers, helping them to architect data centers for AI infrastructure that will support new capabilities for years to come. We've been designing and redesigning products that play a substantial role in data center innovation for decades. You can choose Legrand for:

- Overhead track busway
- Rack power distribution
- Server and network cabinets
- Fiber, cabling, and network
- Rack cooling, including rear door heat exchangers
- Environmental and security sensors

Legrand offers a compelling value proposition for organizations that want the best. To learn more about our capabilities to meet the demanding requirements of today's data center and industrial markets, visit us at www.legrand.us/critical-power-and-infrastructure.

TRUSTED BY THE WORLD'S
LARGEST DATA CENTER OPERATORS

Overhead Track Busway

OOB Management & Services Routers

Fiber Optics

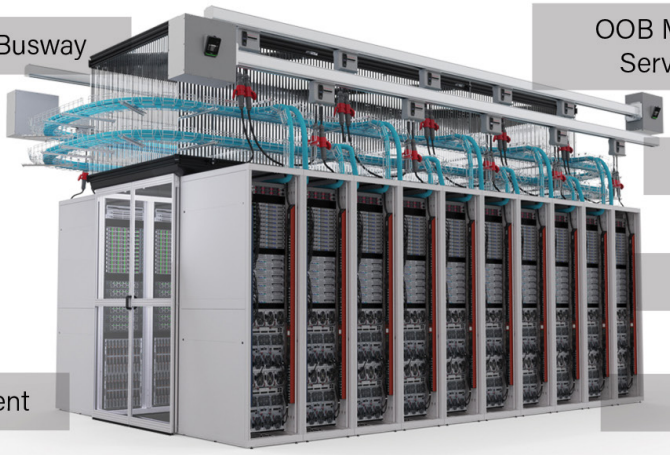
Rear Door Heat Exchange

Remote Access KVM

Intelligent Rack PDUs

Cabinets & Containment

Transceivers



Learn more at
legrand.us/datacenter

Delivering expertly engineered solutions that meet the ever-evolving needs of data centers and building network infrastructure.

Approved Networks | Raritan | Server Technology | Starline | USystems | zpe

To learn more visit

www.legrand.us/critical-power-and-infrastructure

Approved Networks, LLC
800.590.9535
approvednetworks.com

Raritan Americas, Inc.
800.724.8090
raritan.com

Server Technology, Inc
800.835.1515
servertech.com

Starline Holdings, LLC
724.597.7800
starlinepower.com

Legrand
877.295.3472
legrand.us