Creating Resilience

Generative AI and Data Center Decentralization

1 New demand driver: Generative AI

REPORT

CBRE RESEARCH
JULY 2025

Contents

This report is the first in a three-part series analyzing the link between the proliferation of generative AI and the geographical diversification of data center locations throughout Japan (this issue is in bold in the table of contents).

 # 1 New demand driver: Generative AI

1.1. Generative AI as the new key driver of data center demand
1.2. Primary uses and use cases of generative AI
2.1. Characteristics of AI data centers
2.2. AI data centers' tolerance for communication latency

2 Electrical supply issues in Greater Tokyo

3.1. The proliferation of AI data centers
3.2. The tightening data center market in Greater Tokyo
3.3. Changes in data center rack prices
4.1. Electrical supply issues in Greater Tokyo in comparison to regional areas
4.2. "Watt-bit collaboration" and data center decentralization

3 Promising Areas for Decentralization

5. Promising areas for data center decentralization
(1) Renewable energy generation potential
(2) Electricity costs
(3) Concentration of factories and other production bases
(4) Status of infrastructure
(5) Earthquake risk
6. Technological innovation and the potential for further acceleration of data center decentralization

Summary

- The use of generative AI continues to expand across a wide variety of businesses and industries in Japan. One estimate suggests that, as generative AI becomes more widely used, it has the potential to add productive capacity to the Japanese economy equivalent to 27% of GDP . However, generative AI requires vast amounts of electricity. Power consumption by data centers is projected to reach 15x its current level by 2034, at which point it will account for over 5% of the nation's entire electricity demand.
- Building shells and facilities of AI data centers require high-level specifications that exceed those needed by conventional data centers. At the same time, AI data centers tend to have higher tolerance for communication latency, opening up the possibility of new development in regional areas that have not previously hosted data centers.

01

Diffusion of Generative AI and Major Use Cases

1.1. Generative AI as the new key driver of data center demand

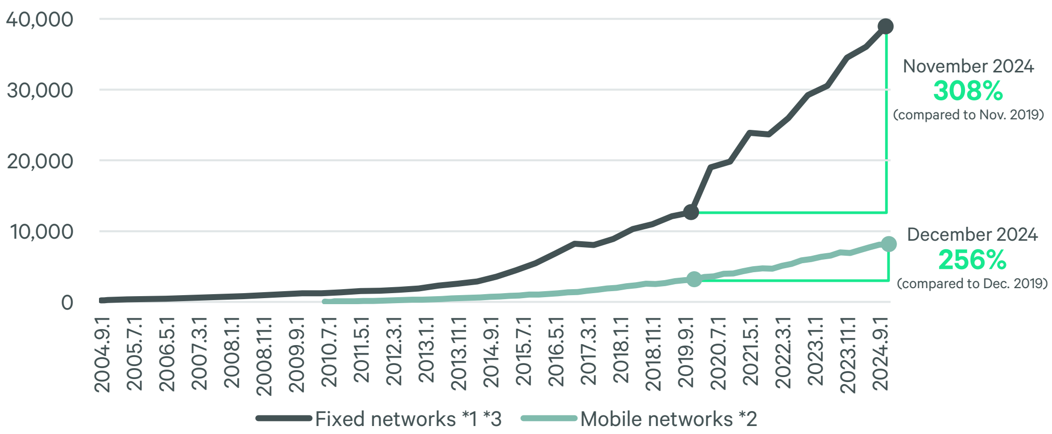
The generative AI boom was triggered in November 2022 with the release of ChatGPT by U.S. developer OpenAI. By February 2025, ChatGPT's weekly user numbers had topped the 400 million mark.*1 Amid high expectations for the contributions generative AI can make to productivity levels across a broad range of businesses and industries, to technology transfer, and as a means of addressing labor shortages, this new technology has already become widely used in Japan, for both personal and business purposes. At the same time, related fields such as cloud services for generative AI and corporate digital transformation are also experiencing rapid growth. As a result of these trends, domestic data traffic (download volume) has spiked sharply in recent years, tripling between November 2019 and November 2024 on fixed communication networks, and growing by 2.5x on mobile communication networks between December 2019 and December 2024 (Figure 1-1). To process and store this sudden increase in data volume, additional computing resources, including high-spec servers and applications, are required. Demand is on the rise for data centers capable of operating such resources, and further improvement in the specifications of building shells and facilities is becoming increasingly vital.

As the use of generative AI becomes more widespread, data traffic volumes are expected to grow exponentially. The promise of other technological advances, such as autonomous driving, has led to one estimate suggesting that domestic data traffic in 2040 will reach 348 times the data traffic volume of 2020.*2 Furthermore, generative AI alone is projected to contribute an additional JPY 148.7 trillion in added productivity to the Japanese economy through supplementing human labor, equivalent to 27% of Japan's 2022 GDP. It is estimated that the manufacturing sector has the greatest latent potential for growth in this respect (Figure 1-2).

*1 "OpenAI's weekly active users surpass 400 million", Reuters, February 21, 2025

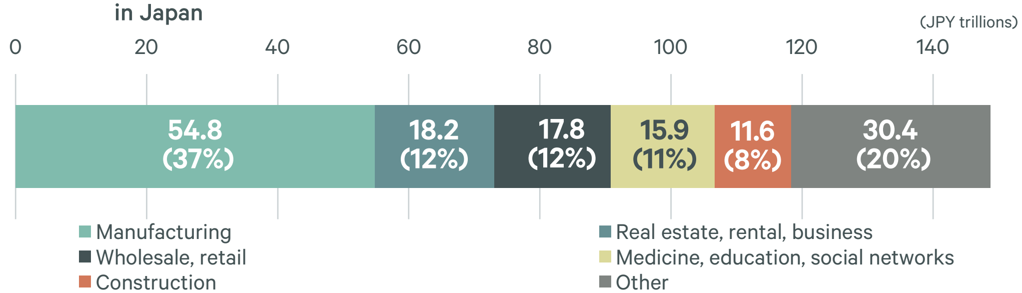
*2 "A Proposal for the Establishment of a New Communication Infrastructure to Support the Data Explosion - Acceleration Toward a Data Usage Society Through Generative AI", Mitsubishi Research Institute, September 2023 (reference date: June 5, 2025)

Figure 1-1: Domestic Download Traffic for Fixed and Mobile Communication Networks



Note: Services for individuals (FTTH, DSL, CATV, FWA) (Includes some B to B services)
*1 Included some mobile communication network traffic with cellphone carriers until May 2011
*2 Taken from "Current state of mobile communication network traffic," Ministry of Internal Affairs and Communications (calculations made in March, June, September and December)
*3 Inconsistencies caused by the increase in cooperating ISPs from five to nine companies in May 2017, and the consequent changes in recorded and estimated data
Source: Created by CBRE based on "Japanese internet traffic data" and "Current state of mobile communication network traffic", Ministry of Internal Affairs and Communications, June 2025. https://www.soumu.go.jp/johotsusintokei/field/tsuushin06.html

Figure 1-2: Potential for Increased Productivity as a Result of the Implementation of Generative AI in Japan



Source: Created by CBRE based on "The Economic Impact of Generative AI: The Future of Work in Japan", Access Partnership, June 2023.

1.2. Primary Uses and Use Cases of Generative AI

In the past, the primary functions of AI have been limited to routine data analysis and pattern recognition. Generative AI differs in that, in addition to these analytical functions, it can generate new content. The primary uses of generative AI can be broadly divided into two categories: development (training) and implementation (inference). Training refers to the process by which generative AI detects patterns and relationships within the vast amounts of data it is fed and forms the basis upon which the inference models of generative AI are built and improved. Inference refers to the generation of outputs in the form of text, images, audio, or programming code, based on user prompts. Major use cases of both the training and inference elements of generative AI are listed by business sector in Figure 1-3. Compared to previous AI systems, generative AI is expected to enable significant improvements in both individual and corporate productivity.

Figure 1-3: Major Use Cases of Generative AI by Business Sector

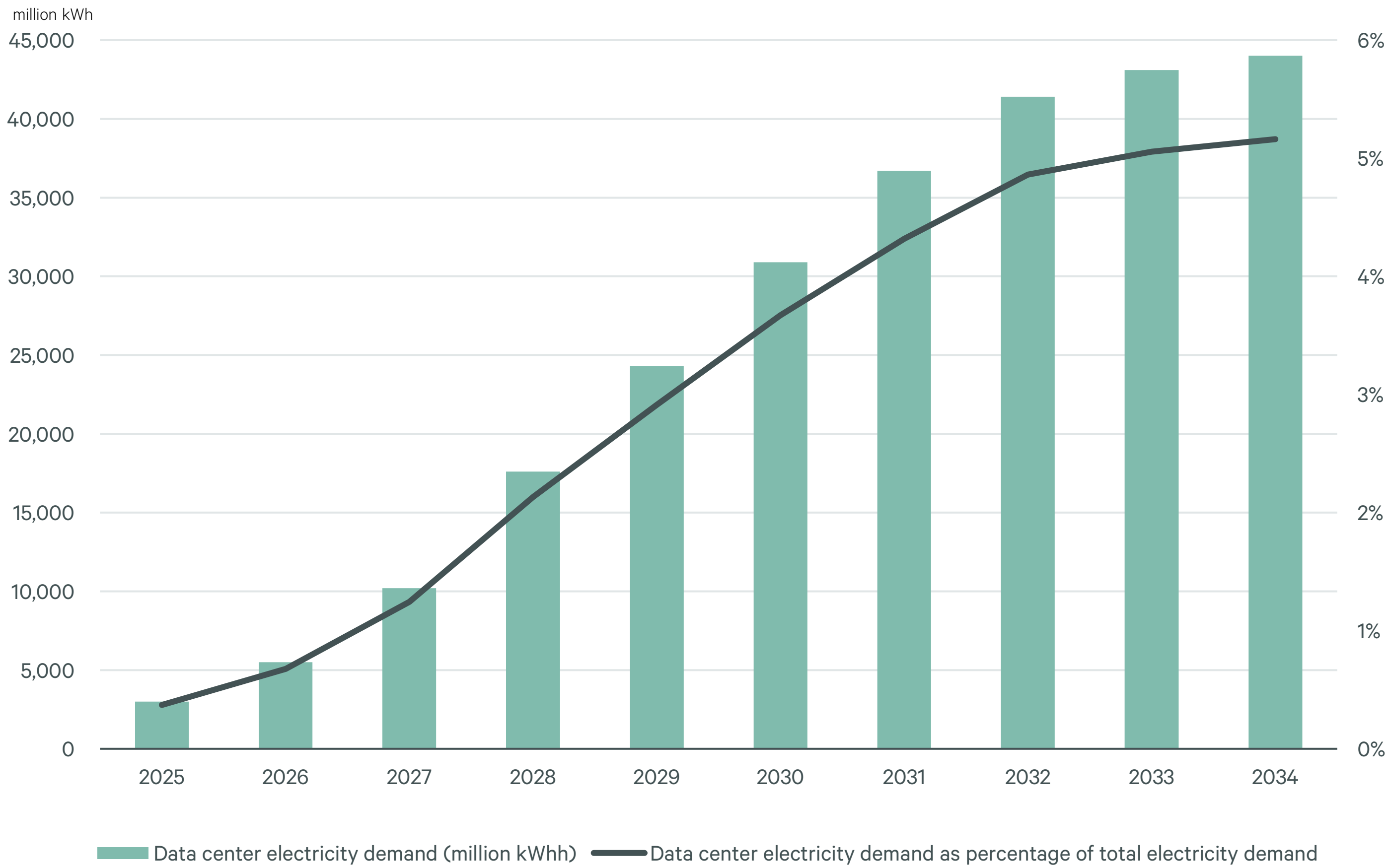
Industry type	Training	Inference
Manufacturing	• Training for autonomous driving • Development of quality control and inspection systems	• Control of vehicles using autonomous driving systems • Task automation in smart factories • Quality control, anomaly detection, and predictive maintenance
Finance and insurance	• Development of chatbots for investment advice • Development of unauthorized use detection systems	• Customer relations using call centers • Unauthorized use detection and risk management • Creation of documents, including drafts for permission and contracts
Medical and healthcare	• Chemical compound analysis for the development of new drugs	• Diagnostic support (medical imaging diagnosis, etc) • Individualized medicine (creation of treatment plans tailored to each patient, etc) • Automated creation of medical documentation based on electronic medical records
Retail	• Development of platforms for the efficient use of product information	• Optimization of ordering plans by predicting demand based on past purchase records • Individualized marketing tailored to each client
Media and advertising	—	• Support for the writing of articles and scripts • Creation of visual content (images and video) • Translation support
Public sector	—	• Broadcasting of emergency information, and use of AI chatbots to handle queries from citizens • Analysis of public data
Cross-industry	• Development of large language models for each industry	• Improvements in operational efficiency for managerial departments such as system development, general affairs, legal, and HR

Source: Created by CBRE based on publicly available materials, June 2025.

1.2. Primary Uses and Use Cases of Generative AI

However, as the use of generative AI necessitates ultra-high-speed computational processing power, it requires servers equipped with semiconductor chips called GPUs, which excel at handling parallel processing. To operate these GPU servers, electrical power equivalent to over 10x that of conventional servers is reportedly required. Data centers tailored to the needs of generative AI (hereafter referred to as “AI data centers”) are those that can stably operate these GPU servers. Strong demand for these data centers from cloud operators and other operating companies has led to an increase in the number of new data center developments. Amid this environment, data center power demand is projected to increase significantly. Electricity consumption by data centers in Japan is estimated to reach approximately 15x its 2025 levels by 2034, expanding during that time frame from just 0.4% to over 5% of the nation’s entire electricity demand (Figure 1-4).

Figure 1-4: Projected Data Center Electricity Demand and Percentage of Total Electricity Consumption up to 2034



Source: Created by CBRE based on “Estimated Nationwide and Supply District-Specific Demand (FY 2025)”, Organization for Cross-regional Coordination of Transmission Operators, Japan January 22, 2025.

02

AI Data Center Requirements

2.1. Characteristics of AI Data Centers

As AI data centers require massive amounts of electrical power, the shells housing them also require high-level specifications, including reinforced floors and high ceilings, to allow them to accommodate the power, air conditioning, and IT equipment necessary. Figure 2-1 illustrates the standard specifications required by the three basic types of data centers: the conventional "enterprise" type, which has operating companies as its end users; "hyperscale" data centers, used by major cloud operators; and AI data centers, designed to house GPU servers.

(1) Electrical facilities (power density, IT power receiving capacity)

The electrical power capacity for each rack, housing servers and other IT equipment, is higher for AI data centers than for the other two types of data centers. The training process requires particularly large quantities of electricity, as AI systems consume vast amounts of power in the process of absorbing the colossal volume of data from which they learn. As the inference process, by comparison, involves outputting what has been learned by the model created by the training process, it requires considerably less electricity to operate. Even then, the power requirements are at a level well above those of hyperscale data centers, which were previously considered the gold standard in terms of specifications. In terms of IT power receiving capacity, a measurement of the electricity required by their IT equipment, AI data centers again lead the way. It should be noted, however, that the scale varies significantly by purpose, with some cases requiring only a few megawatts.

Figure 2-1: Standard Specifications of the Three Primary Types of Data Centers

	Enterprise-type data centers	Hyperscale data centers	AI data centers	
			Reasoning	Learning
Power capacity (kVA per rack)	6kVA (ave)	8kVA (ave)	20kVA+	40kVA+
IT power receiving capacity	5MW+	15MW+	A few to several hundred MW	
No. of racks*	100+	2,000+	Tens to thousands	
Total floor area	1,000 tsubo+	5,000 tsubo+	1,000 tsubo+ (varies significantly based on power capacity)	
Primary cooling facilities	Air cooling	Air cooling	Air cooling/liquid cooling	Liquid cooling
Ceiling height	5.0m+	6.0m+	6.0 to 7.0m+	
Floor load (/m^2)	1.0t+	1.5t+	1.5t+	2.0t+
Location	City center (within 15 km of IX*)	Suburban area (30 to 50 km of IX)	City center (within 15 km of IX*) or near end users	Anywhere from city center to suburban area

*1 IX is the abbreviation for Internet Exchange Point, a physical location through which Internet infrastructure companies, such as Internet service providers (ISPs) and network operators, such as data centers, can exchange data with each other.
*2 Racks refer to standardized shelves upon which servers and other computer equipment are fixed for storage. The most widely used racks in the industry today are the 19 inch (48 centimeter)-wide racks developed by America's Electronic Industries Alliance.
Source: CBRE, June 2025.

2.1. Characteristics of AI Data Centers

(2) Cooling facilities

As GPU servers consume significantly more electricity than standard servers, they also generate significantly more heat. As a result, the traditional thermal management techniques used by conventional data centers are insufficient for the cooling needs of AI data centers, forcing AI data centers to adopt liquid cooling systems. There are three main types of liquid cooling: rear-door heat exchangers, direct-to-chip cold plates, and immersion cooling. The cooling capacity and implementation difficulties of each are outlined in Figure 2-2. While liquid cooling systems incur higher initial costs than air cooling systems, their superior efficiency can improve the power usage effectiveness (PUE[*3]) of the data centers that use them.

(3) Reinforced floors, high ceilings

GPU servers weigh at least three times as much as standard servers, while liquid cooling systems are also heavier than their air cooling counterparts. For this reason, the floor load is higher for AI data centers than for the other two types of data centers. Furthermore, delivering the power required to run GPU servers is challenging via standard power cables. For this reason, AI data centers tend to use ceiling-installed belt-shaped power distribution conduits known as bus ducts, which are more suited to carrying the high-ampacity power required. While bus ducts are already commonly used in hyperscale data centers, AI data centers, which must carry even higher amounts of electrical current, tend to use either wider or multiple bus ducts. To accommodate these systems, AI data centers often require ceiling heights that exceed those demanded by hyperscale data centers.

*3 PUE is a measure of energy efficiency (with a figure closer to 1.0 representing greater efficiency), defined as PUE = [Total Facility Energy] / [IT Equipment Energy]. The average data center is reported to operate with an average PUE of around 1.5.

Figure 2-2: Comparison of Air Cooling and the Three Main Types of Liquid Cooling

	Air cooling	Liquid cooling		
		Rear-door	Cold plate	Immersion
Summary	The traditional method in which air conditioning machines pump cold air into the server room, and exhaust air from the servers is returned to the air conditioner for heat removal.	A liquid heat exchanger is attached to the rear of the server rack to remove the heat generated by the server.	Liquid is pumped directly onto the IT equipment's heat-generating components, while heat is absorbed via cold plates.	IT components are cooled by immersing them in a thermally conductive dielectric liquid. Heat is directly absorbed by the liquid, which is then cooled by an external cooler.
Cooling effectiveness				
Difficulty of implementation into existing facilities				

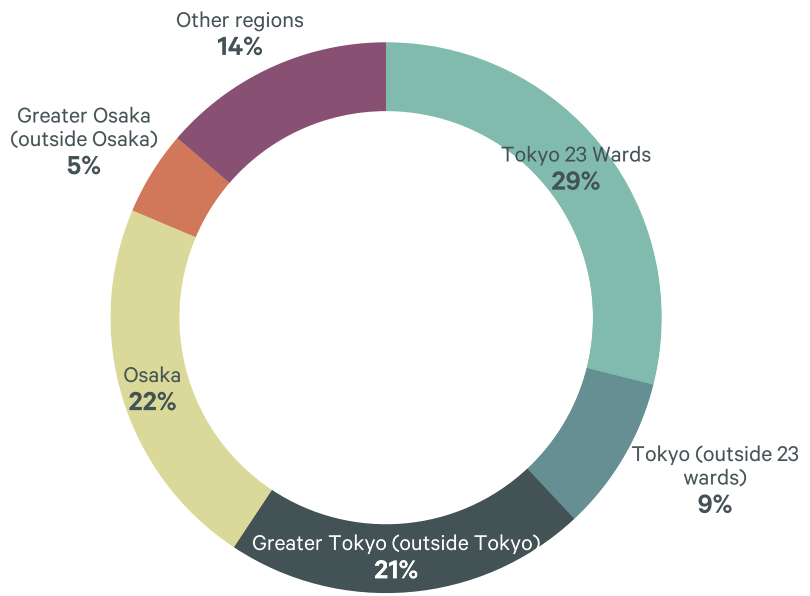
Source: CBRE, June 2025.

2.2. AI Data Centers' Tolerance for Communication Latency

In addition to facility and shell specifications, AI data centers also differ from the other types in terms of locational requirements (see "Location" in Figure 2-1). As of 2024, data centers in Japan are predominantly located in major metropolitan areas, with approximately 60% in the Greater Tokyo area, 30% in the Greater Osaka area, and just 10% in regional cities (Figure 2-3). The major factor behind this trend is that most of the nation's internet exchange points (IXes) are concentrated near Tokyo and Osaka, and proximity to IXes is the optimal means of reducing latency and securing high communication quality for data centers. AI data centers, on the other hand, require only a limited amount of internet-based communication for the training part of their operation; instead, their primary focus is on absorbing vast quantities of data. In this respect, their tolerance for latency is relatively high, meaning that they do not necessarily need to be located in major metropolitan areas.

For the inference part of their operation, frequent outputs are required in response to user instructions. This means that lower latency is a significant advantage, and metropolitan locations are desirable. However, even for this inference operation, so-called "edge data centers", which process requests near their end users and do not need to communicate via the internet, need not be located in metropolitan areas. For example, in the case of a "smart factory" that collates and analyzes production processes to improve productivity, the data center responsible for this analysis and output would be installed on the grounds of the factory, wherever that might be. For these reasons, AI data centers are more amenable than other types of data centers to being located not only in suburban areas outside major metropolises, but even in regional areas with low population and business concentration, which have not hosted data centers in the past.

Figure 2-3: Major Data Center Locations by Region



Source: CBRE, June 2025.

Report Contacts

Research Analysts

Yuji Iwama
Senior Director
Office Team Leader/ Data Center
yuji.iwama@cbre.com

Chinatsu Hani
Senior Director
Head of Research
chinatsu.hani@cbre.com

Contributors

Data Center Solutions

Hirokazu Ono
Associate Director
Head of Data Center Solutions Japan
hirokazu.ono@cbre.com

Takashi Kishida
Associate Director
Data Center Solutions Japan
takashi.kishida@cbre.com

Valuation Advisory & Consulting Services

Yukihiro Okada
Consultant
Consulting Services
yukihiko.okada@cbre.com