NTT

Computing Infrastructure for the AI Era

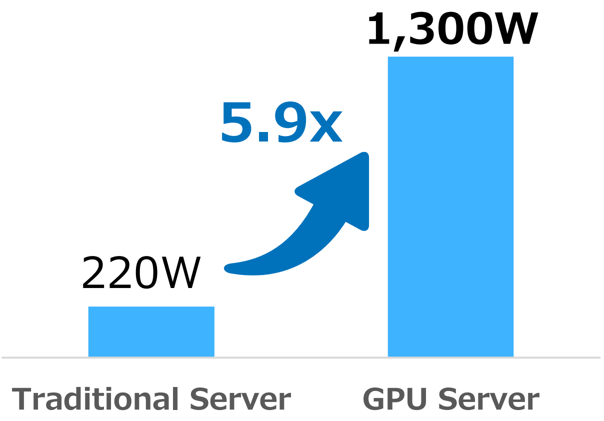
Sachiko Oonishi
Executive Member of the Board, Executive Vice President, CCXO and Co-CAIO
Head of Research and Development Market Strategy
NTT, Inc.

Oct. 6th, 2025

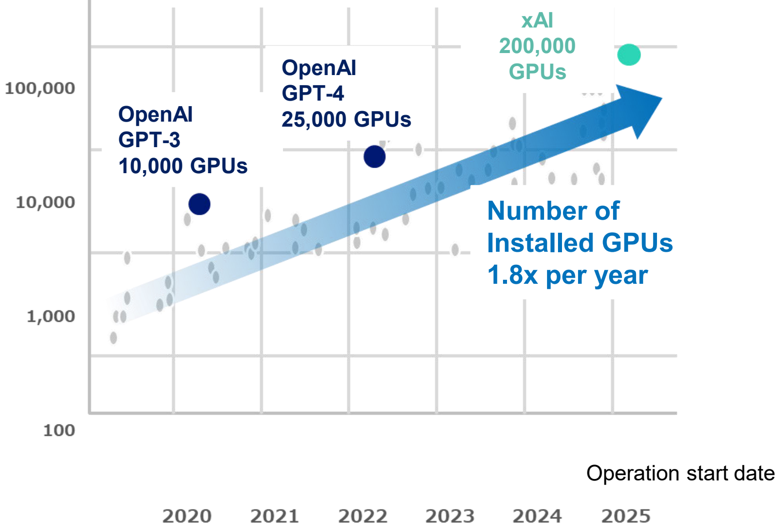
The Rapid Surge in Power Consumption Due to the Wider Adoption of AI

- GPU servers consume 5.9 times the power of traditional servers.
- GPU installations are growing 1.8x annually, driving continuous power consumption increases.

Power consumption of a server (*1)



Scaling of AI computers(*2)

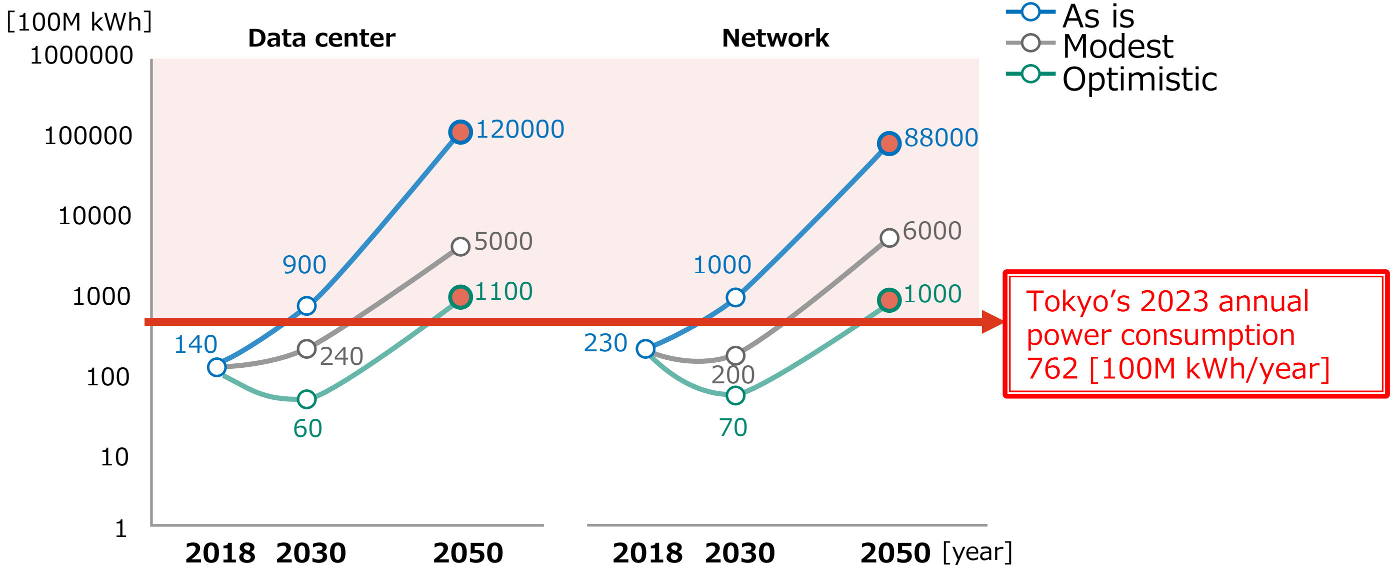


Sources:
*1 Created by NTT from EY Japan “Requirements of data centers of GPU servers”
*2 Epoch AI

The Rapid Adoption of AI Is Fueling a Power Crisis

- At this rate, data center power will surpass Tokyo's 2023 annual consumption.

Forecast of data center power consumption



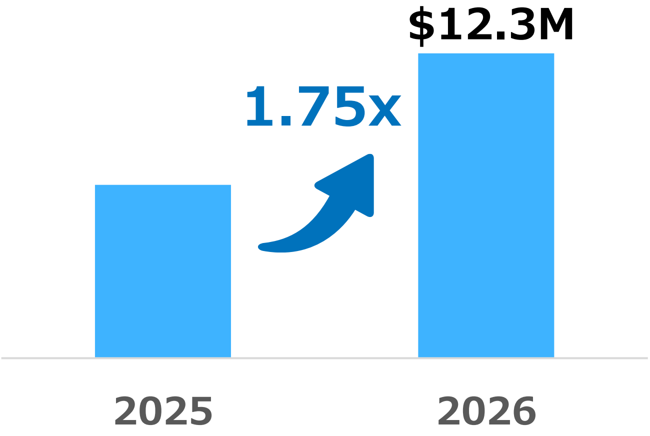
Tokyo’s 2023 annual power consumption 762 [100M kWh/year]

Source: Subcommittee on Basic Policy, Advisory Committee for Natural Resources and Energy (56th Meeting)
https://www.enecho.meti.go.jp/committee/council/basic_policy_subcommittee/2024/056/056_005.pdf

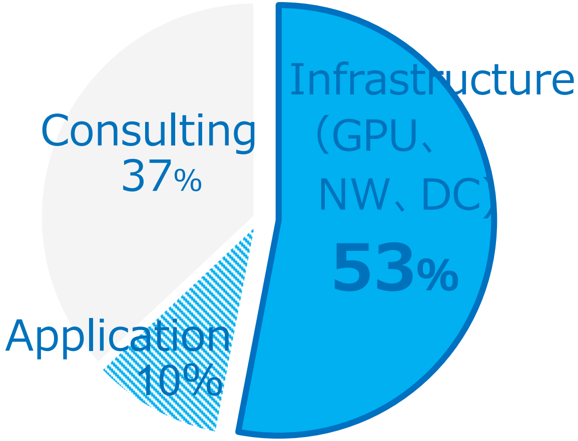
As AI Use Expands, User Costs Are Also Surging

- As AI adoption grows, surging utilization costs are fueling concerns about declining ROI.
- The infrastructure domain will account for over half the AI market.

Enterprise AI adoption costs (＊1)
(Average Annual LLM Budget per Company)



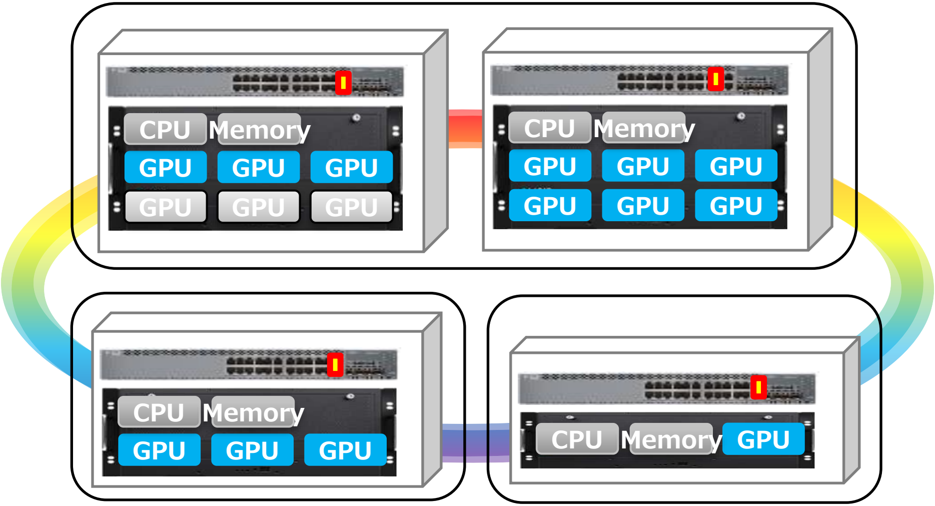
Forecast of
AI market segmentation, 2027(＊2)



Sources:
＊1 Andreessen Horowitz "How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025"
＊2 Created by NTT based on multiple reports from various sources, including Fuji Chimera Research Institute, Yano Research Institute, and Deloitte Touche Tohmatsu MIC Research Institute.

Requirements for Infrastructure in the AI Era

- Efficient operation and reduced energy consumption are essential for AI computing infrastructure.
- IOWN is a crucial component for balancing AI benefits with efficient operation and reduced power consumption.



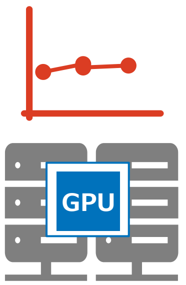
Efficient Infrastructure Operations

Reducing Power Consumption of Infrastructure Equipment

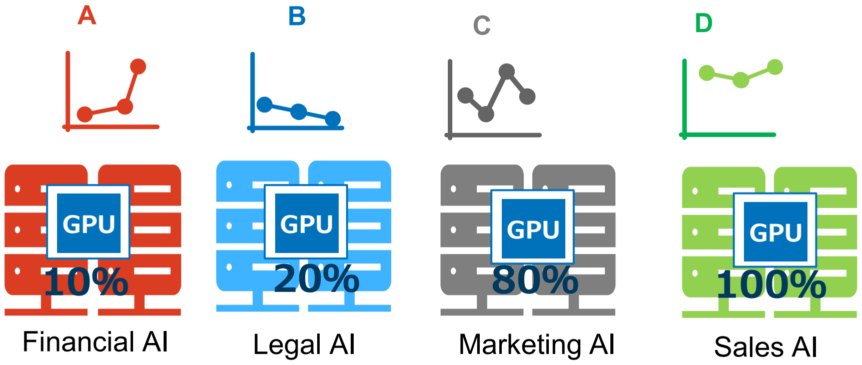
Wider AI Adoption is Reshaping Compute Utilization

- As AI use expands across business processes, fluctuating compute utilization drives up power and costs.

Previously: Limited AI applications led to high GPU utilization for specific tasks



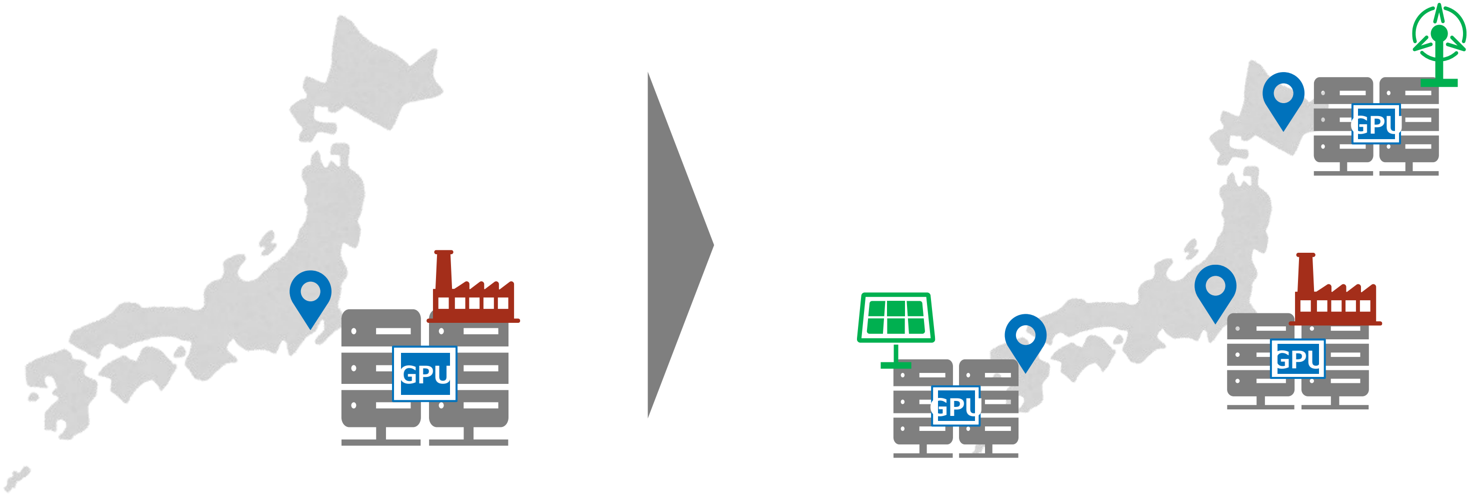
The expansion of AI and task diversity is causing fluctuating GPU utilization



✓ **Only 50% of the total resources are being utilized.**
⇒ Optimizing scattered resources via GPUaaS can halve hardware investment.

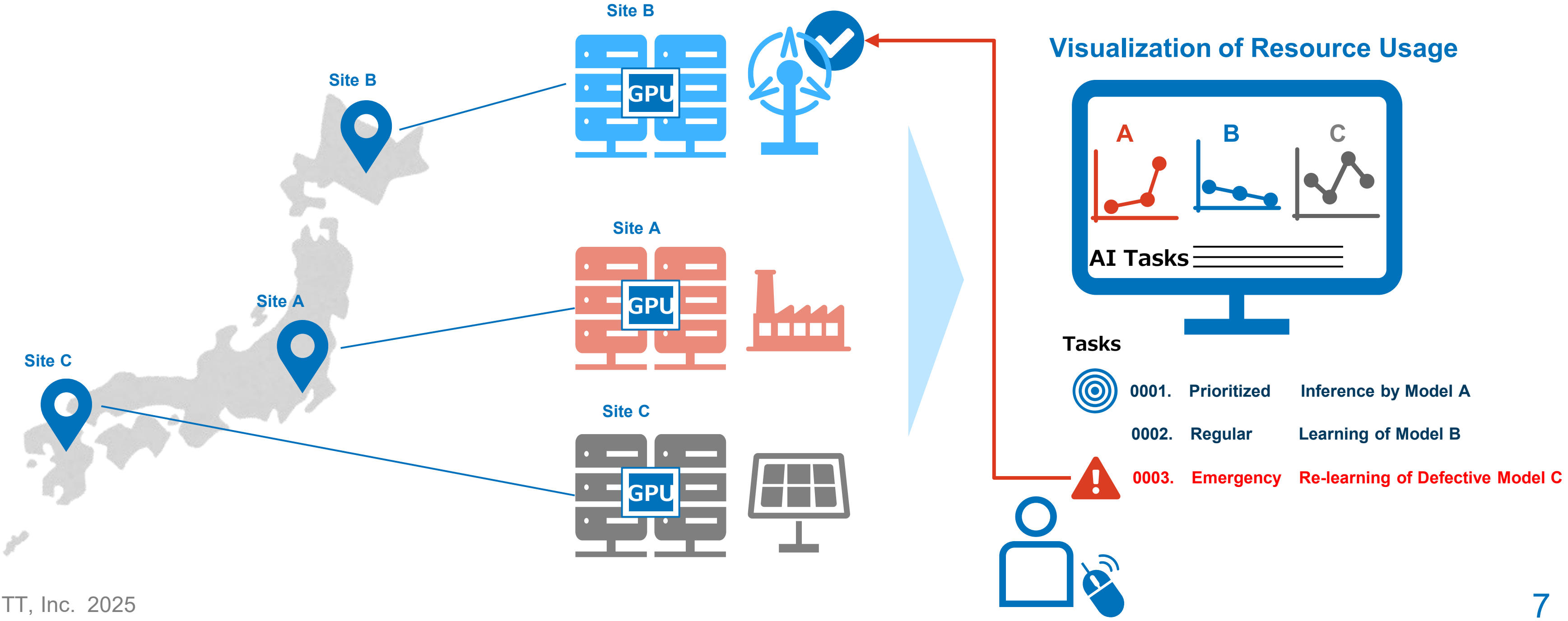
The Reality of Compute Infrastructure in the AI Era

- Growing infrastructure demand, led by AI, is maxing out power supply in the Tokyo Metropolitan Area.
- Data centers and GPUs are increasingly being distributed to regions with power surpluses.



Infrastructure Optimization Initiatives at Leading Companies

- Leading companies are starting to efficiently operate infrastructure by visualizing resource status (e.g., GPU usage and power consumption) and prioritizing AI workloads in resource-rich locations.

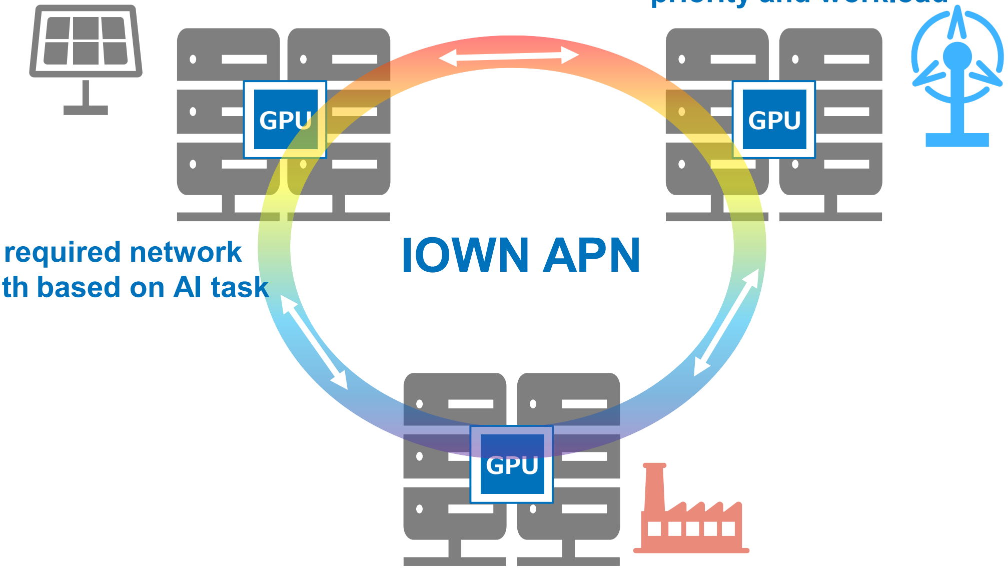


Tasks

0001.	Prioritized	Inference by Model A
0002.	Regular	Learning of Model B
0003.	Emergency	Re-learning of Defective Model C

Infrastructure Requirements in the AI Era:
1. Efficient Infrastructure Operations

- High-speed APN links to connect multi-tiered GPUs and distributed GPUs.
- Flexible allocation of compute resources, such as GPUs, based on AI application and utilization.
- Dynamically assign AI tasks to sites with sufficient power, based on application and workload, ensuring stable and efficient operation.



Allocate GPUs based on AI task priority and workload

Allocate required network bandwidth based on AI task priority

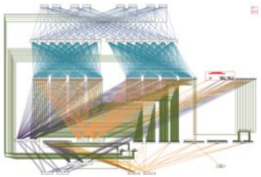
As AI Utilization and Data Volume Increase, Electrical Processing Is Reaching Its Limit

- Increased AI processing → More GPUs → Greater intra-computer communication
- Increased power consumption and heat generation present a limit for electrical communication → Transition to optical communication.

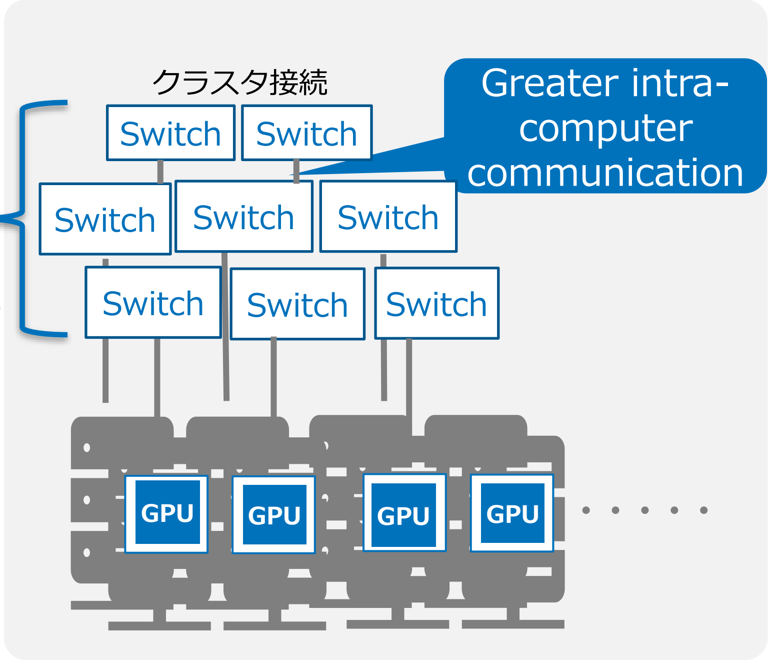
Increased GPU processing drives higher data transfer volumes.

NTT DATA’s GPUaaS

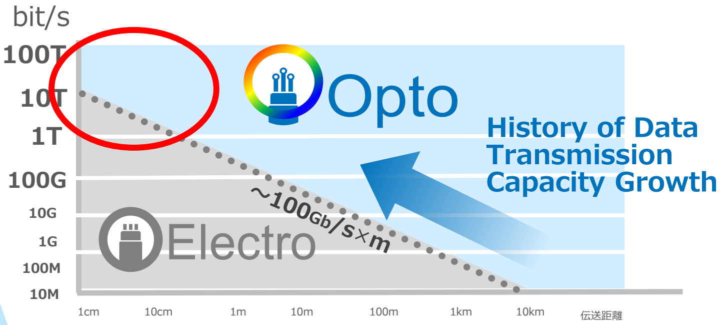
The computing infrastructure for developing NTT's LLM, "tsuzumi"



High-volume intra-computer communication



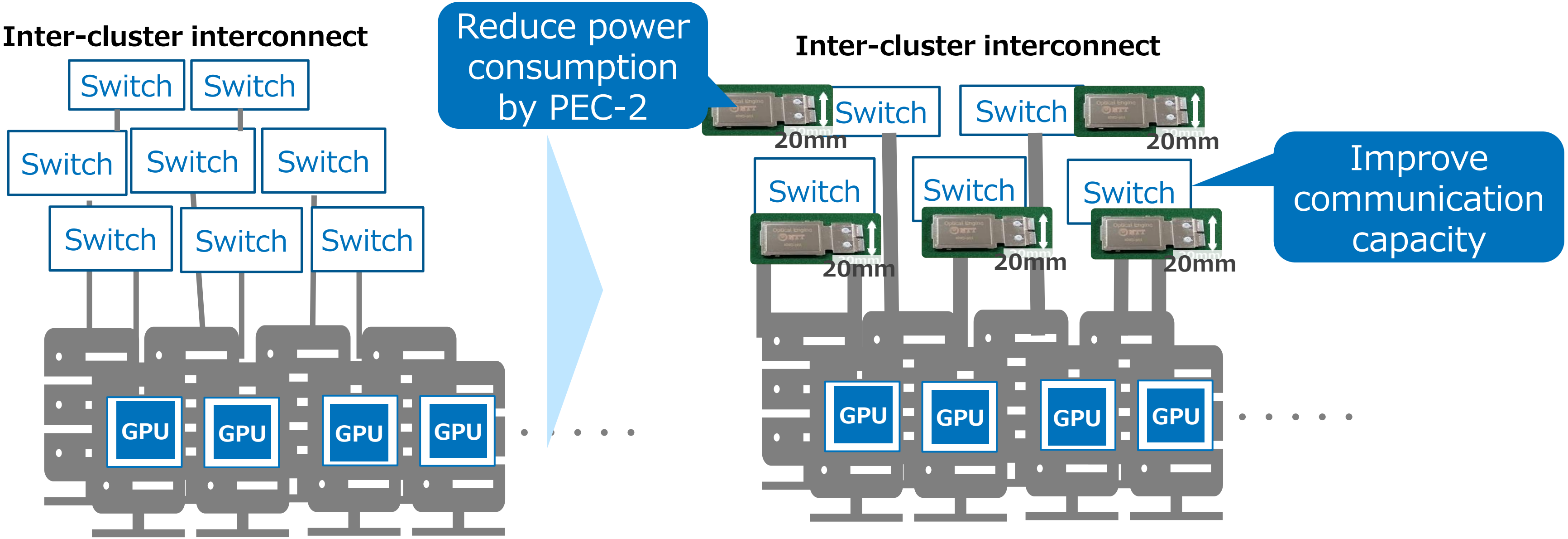
Greater intra-computer communication



IOWN’s Photonics-Electronics Convergence devices can suppress heat generation and power consumption even with high-capacity communication.

Infrastructure Requirements in the AI Era:
2. Reducing Power Consumption

- By implementing Photonics-Electronics Convergence devices (PEC-2) in network switches, we can achieve both increased communication capacity and reduced power consumption.



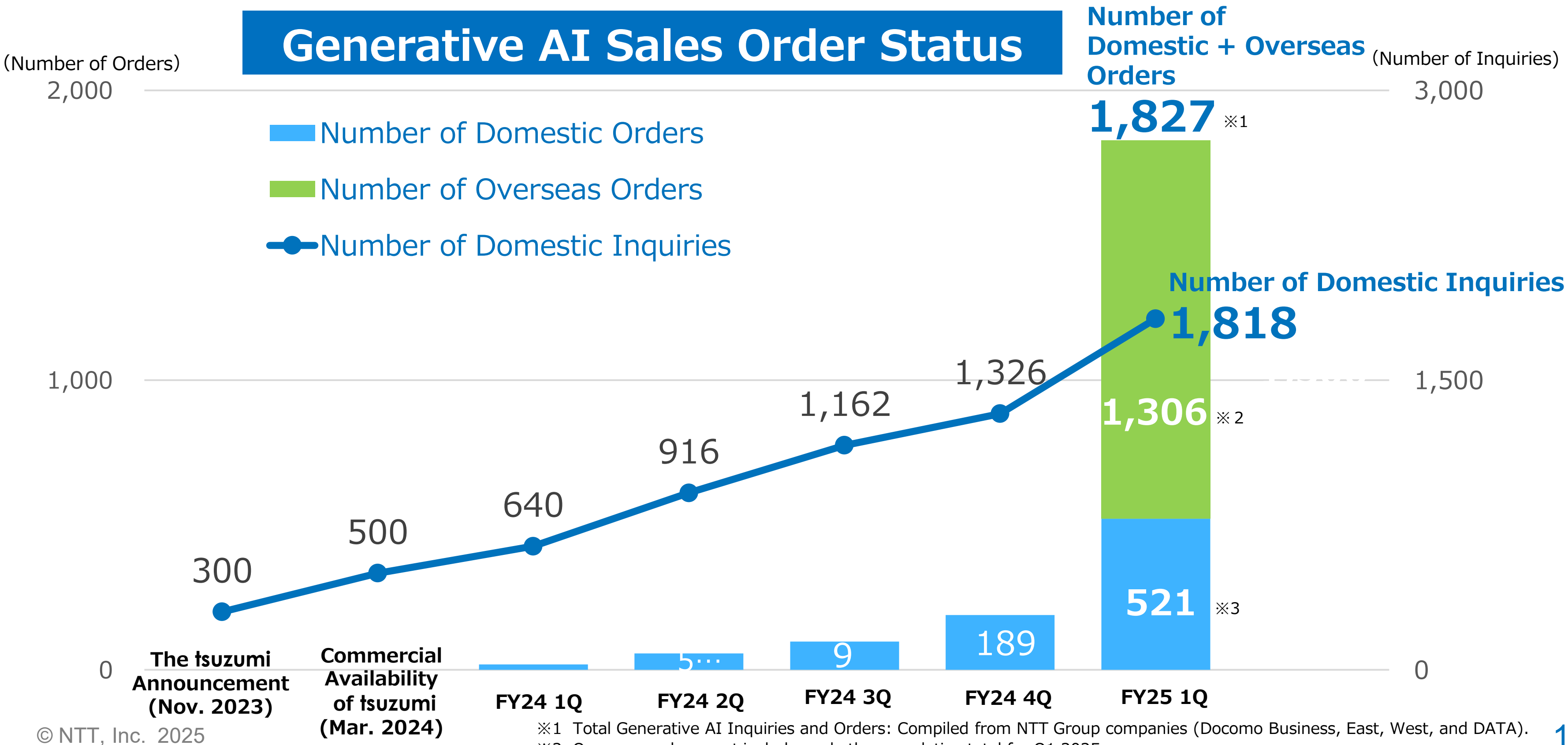
Inter-cluster interconnect

Reduce power consumption by PEC-2

Inter-cluster interconnect

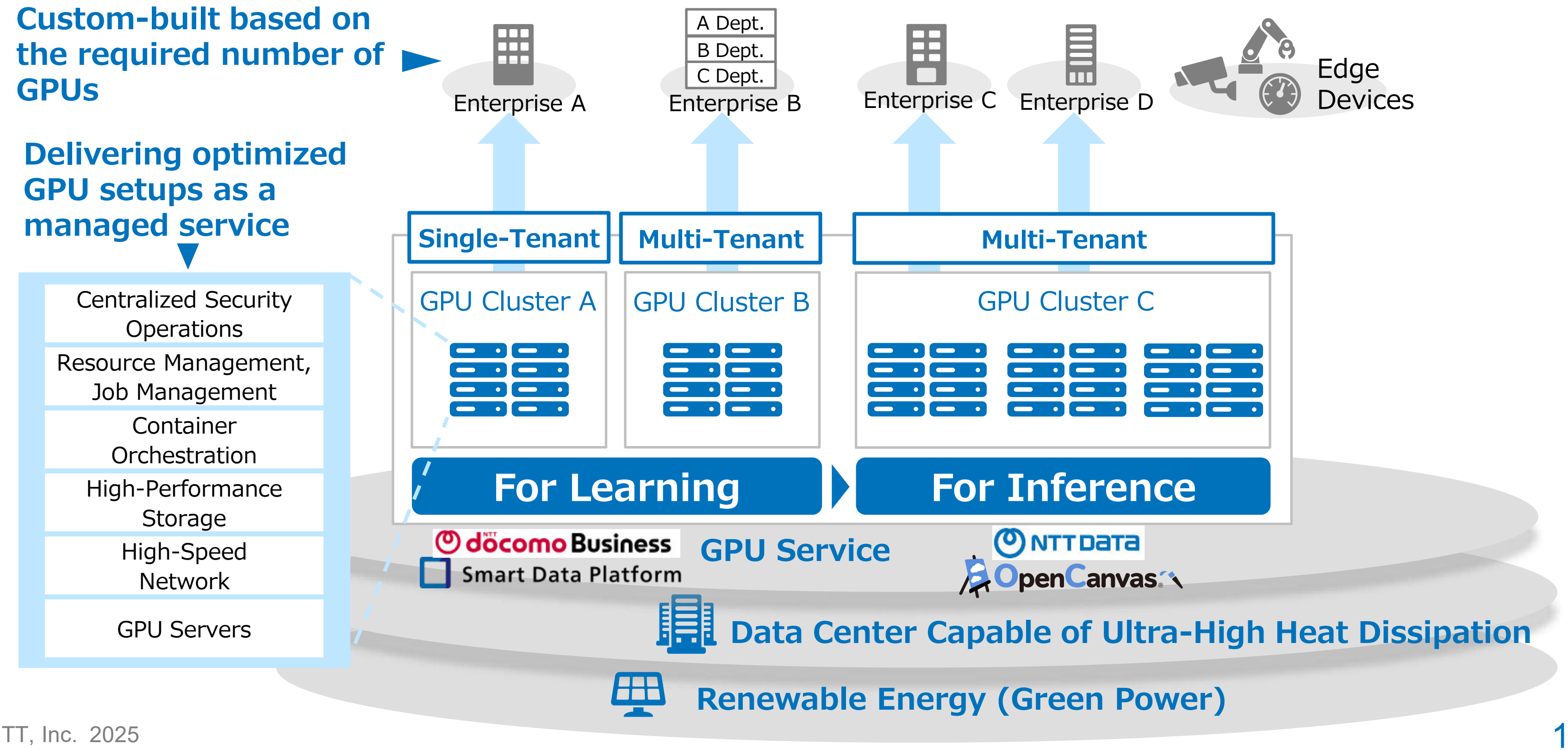
Improve communication capacity

AI Adoption Is Progressing Steadily



※1 Total Generative AI Inquiries and Orders: Compiled from NTT Group companies (Docomo Business, East, West, and DATA).
※2 Overseas order count includes only the cumulative total for Q1 2025.
※3 Including pre-approved projects.

NTT's GPU as a Service: Delivering Compute Resources When and Where Needed, Based on AI Utilization



Custom-built based on the required number of GPUs

Enterprise A

A Dept.
B Dept.
C Dept.
Enterprise B

Enterprise C

Enterprise D

Edge Devices

Delivering optimized GPU setups as a managed service

- Centralized Security Operations
- Resource Management, Job Management
- Container Orchestration
- High-Performance Storage
- High-Speed Network
- GPU Servers

Single-Tenant	Multi-Tenant	Multi-Tenant
GPU Cluster A	GPU Cluster B	GPU Cluster C

For Learning → **For Inference**

docomo Business
Smart Data Platform

GPU Service

NTT DATA
OpenCanvas

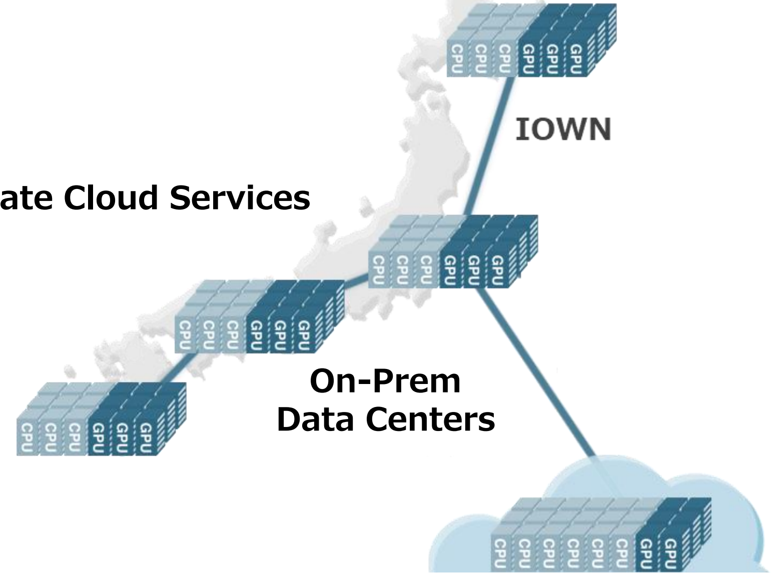
Data Center Capable of Ultra-High Heat Dissipation

Renewable Energy (Green Power)

Jointly Developing and Operating a "Mobility AI Platform" with Toyota to Realize a Zero-Traffic-Accident Society

TOYOTA NTT

Distributed Computing Platform

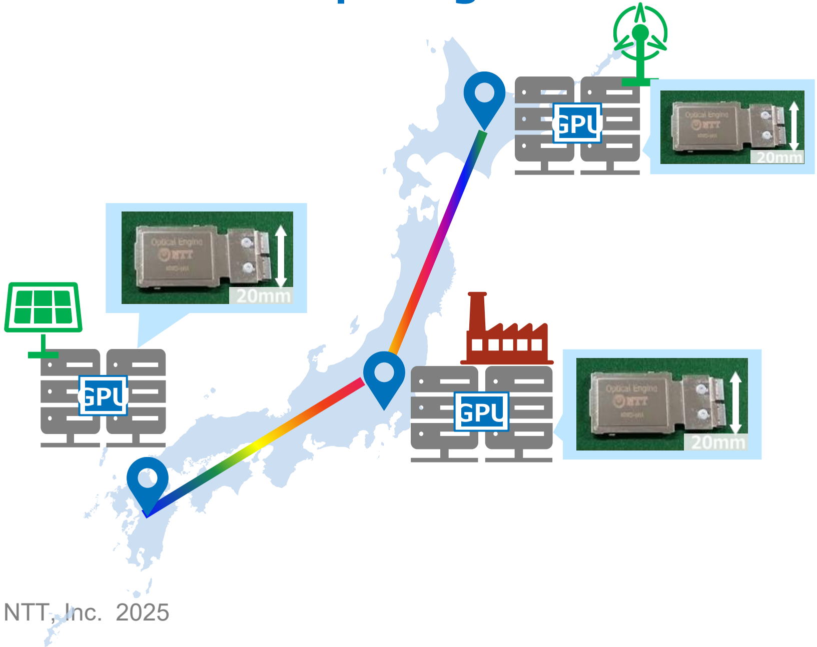


Public Cloud Services

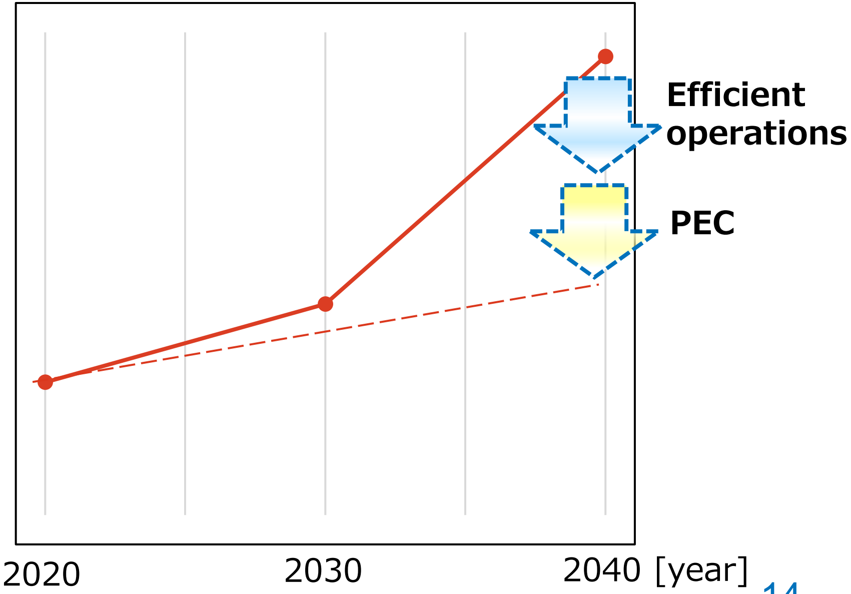
Computing Infrastructure for the AI Era

- Efficiently managing compute resources to boost GPU utilization and reduce overall power consumption.
- Introducing Photonics-Electronics Convergence devices enables further reduced power consumption by running the same infrastructure with optimized GPUs and equipment.

Future Computing Infrastructure



Infrastructure Power Consumption



Innovating a Sustainable Future for People and Planet

NTT

This document is a translation of the Japanese original. The Japanese original is authoritative.

The forward-looking statements and projected figures concerning the future performance of NTT and its subsidiaries and affiliates contained or referred to herein are based on a series of assumptions, projections, estimates, judgments and beliefs of the management of NTT in light of information currently available to it regarding NTT and its subsidiaries and affiliates, the economy and telecommunications industry in Japan and overseas, and other factors. These projections and estimates may be affected by the future business operations of NTT and its subsidiaries and affiliates, the state of the economy in Japan and abroad, possible fluctuations in the securities markets, the pricing of services, the effects of competition, the performance of new products, services and new businesses, changes to laws and regulations affecting the telecommunications industry in Japan and elsewhere, other changes in circumstances that could cause actual results to differ materially from the forecasts contained or referred to herein, as well as other risks included in NTT's most recent Annual Securities Report and in any other materials publicly disclosed by NTT on its website.