The Potential of IOWN and “GPU over APN” in Shaping Future AI Infrastructure

MWC26

NTT docomo Business

CONTENT

01 Overview of IOWN

02 AI infrastructure and distributed data centers

03 Latest Case Studies : GPU over APN
AI Training / AI Inference / Data Transfer

docomo Business

01 Overview of IOWN

docomo Business

docomo Business

IOWN

What Is It?

And Why Do We Need It?

What is IOWN?

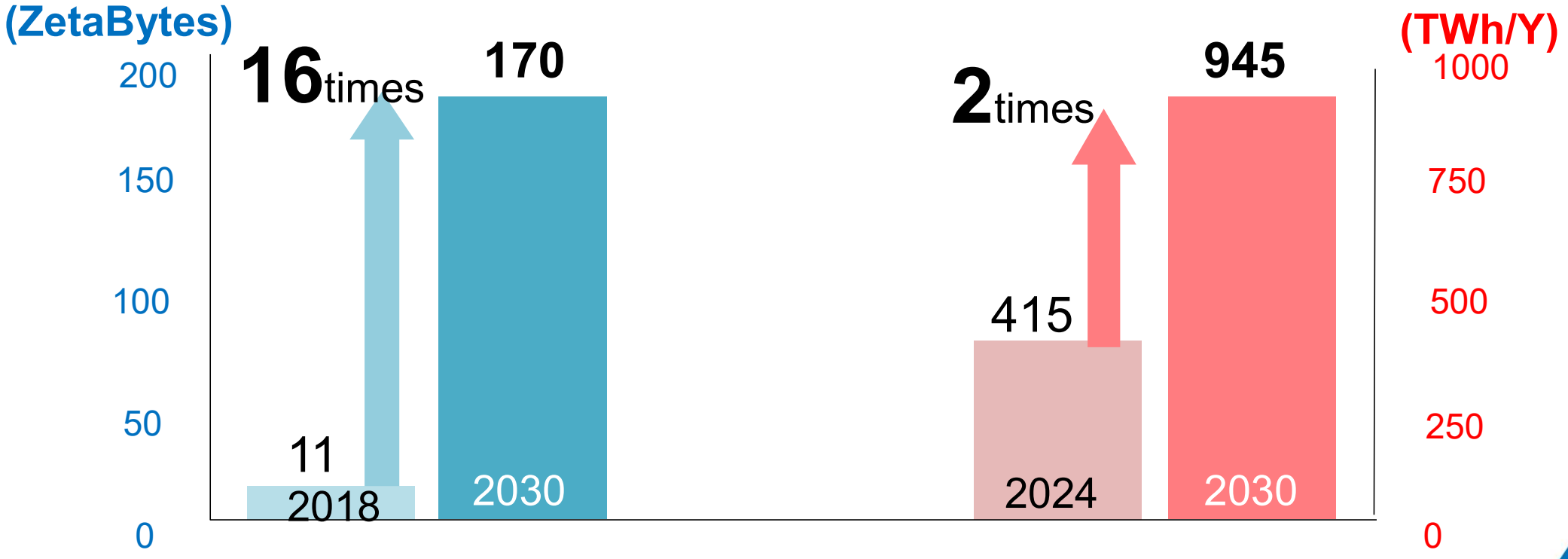
docomo Business

Innovative Optical and Wireless Network

Large Capacity, High Quality	Low Latency	Low power consumption
Transmission capacity **125x**	End-to-end latency **1/200**	Power efficiency **100x**

Why IOWN?

Problem: power consumption on Global Datacenter



Data volume

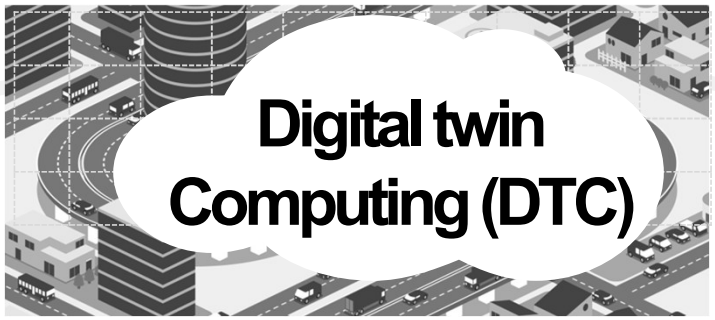
※Center for Low Carbon Society Strategy, Japan Science and Technology Agency

Power consumption

※Source: IEA “Energy and AI” (published April 2025) Figure 2.11

Key Technologies for IOWN

docomo Business



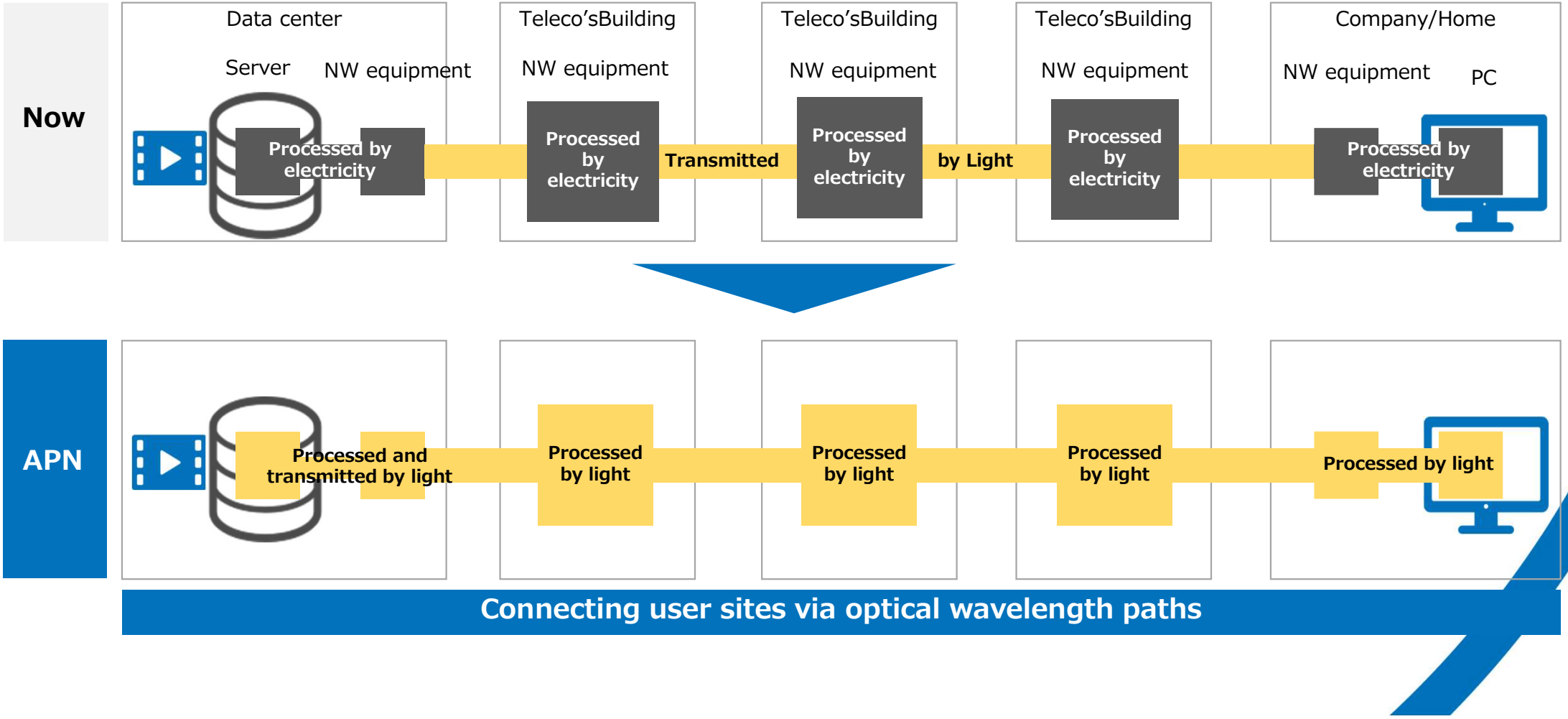
Cognitive
Foundation (CF)

IOWN Optical Computing

Photonics-
Electronics
Convergence
technology
(PEC)

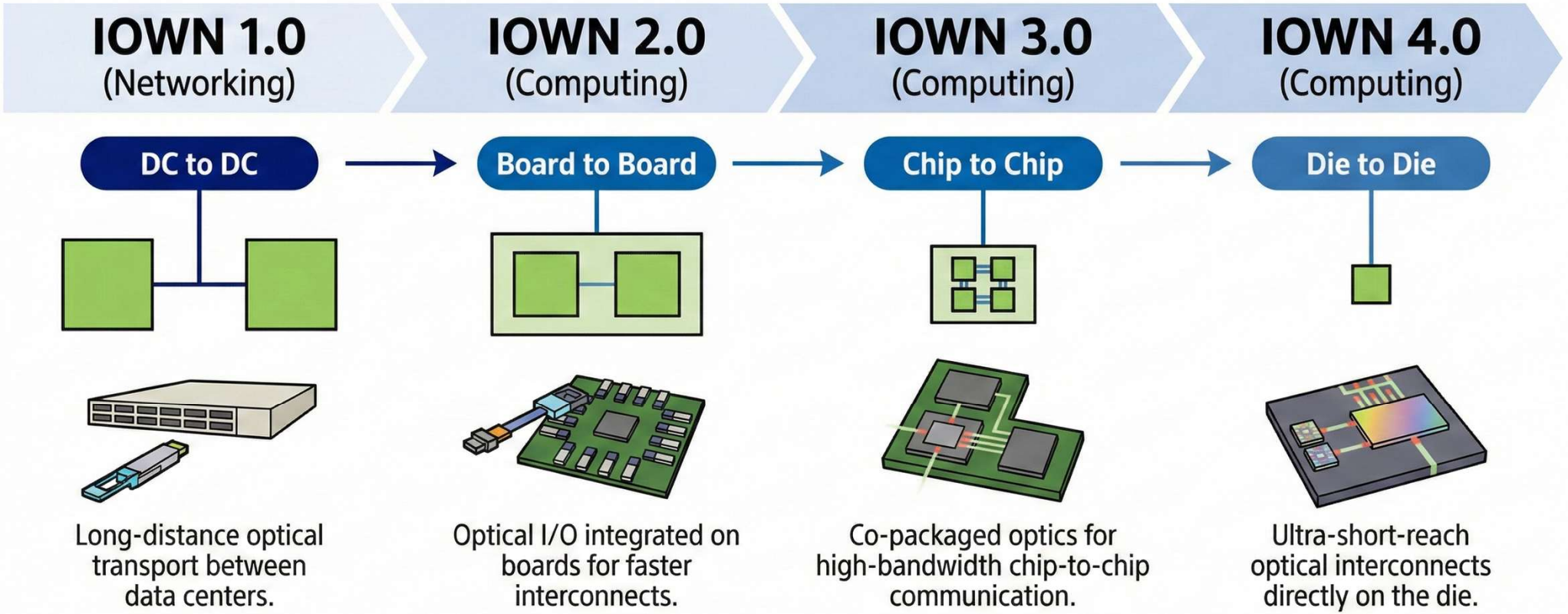
All Photonics Network (APN)

Comparison of Conventional Networks and IOWN APN



IOWN RoadMap

docomo Business



Progression towards smaller, faster, and more integrated optical-electrical convergence.

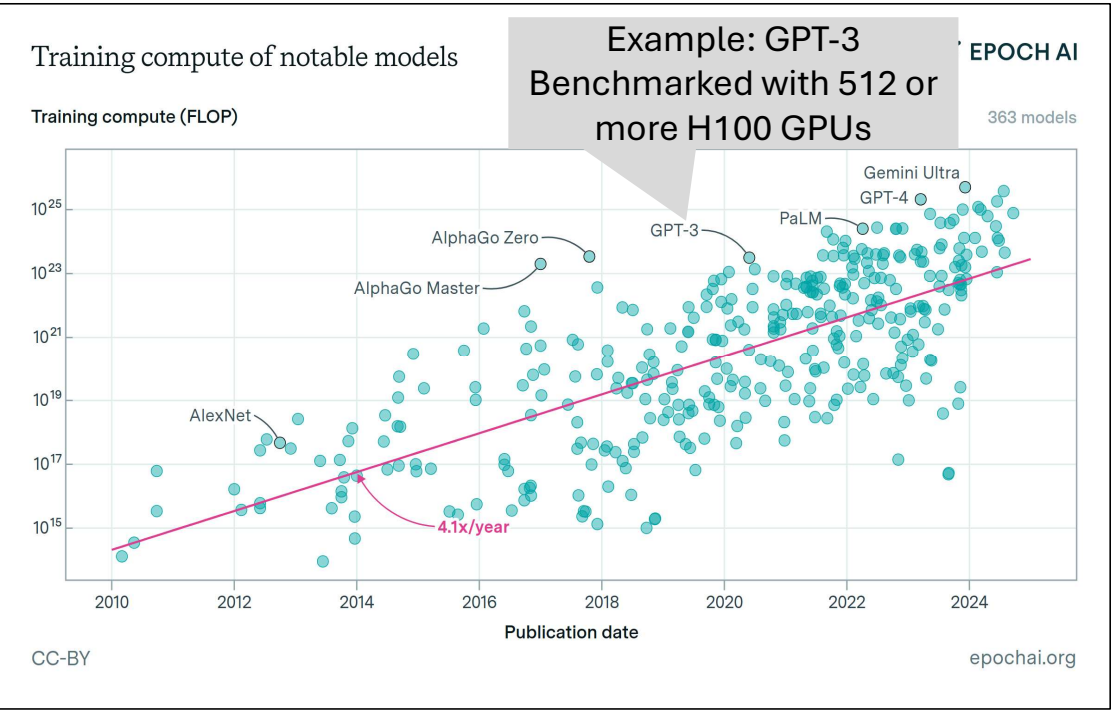
02 AI infrastructure and distributed data centers

NTT docomo Business

Backgrounds

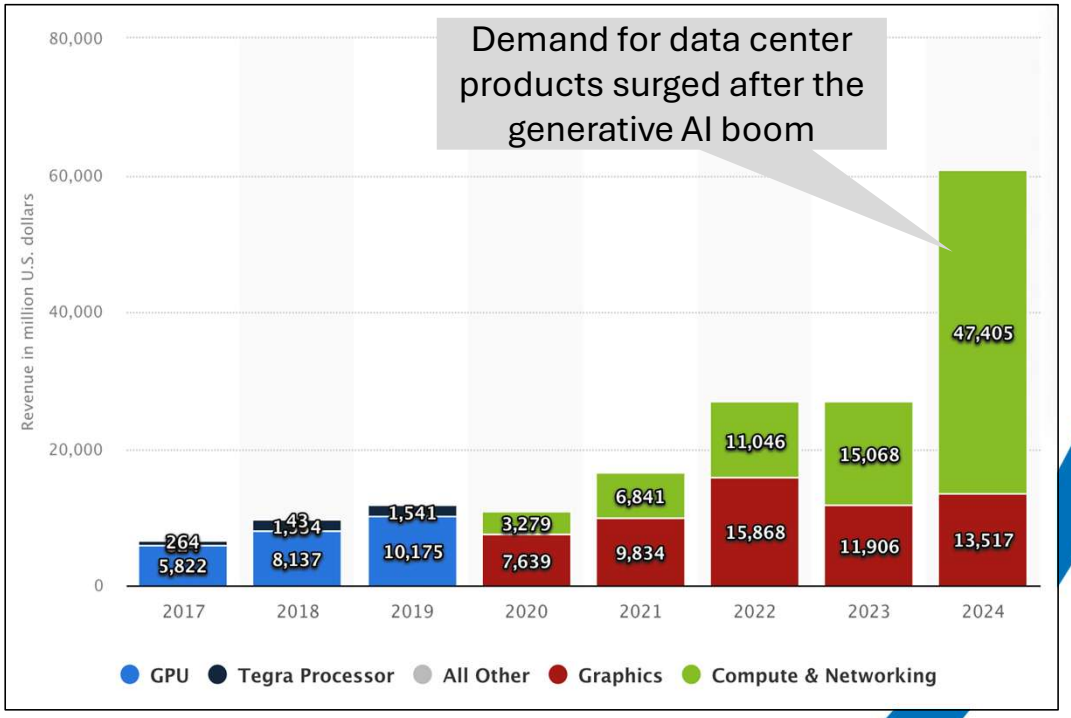
More and more GPUs as compute resources are needed, in areas such as generative AI, data science, and media processing.

Computational complexity of AI model training



https://epochai.org/data/notable-ai-models

Sales of NVIDIA Corporation of the GPU industry



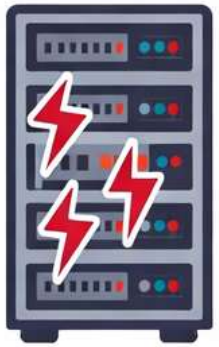
https://www.statista.com/statistics/988034/nvidia-revenue-by-segment/

GPU Infrastructure Requirements and Challenges

docomo Business

From the perspective of infrastructure constraints

Power density per rack

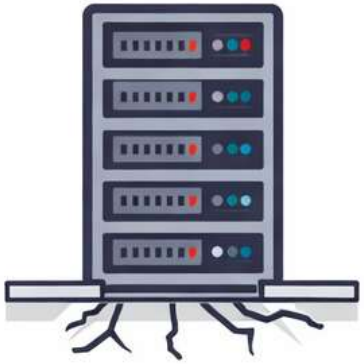


Exceeds power supply capacity

Cooling capacity

Insufficient air cooling

Floor Load Limit



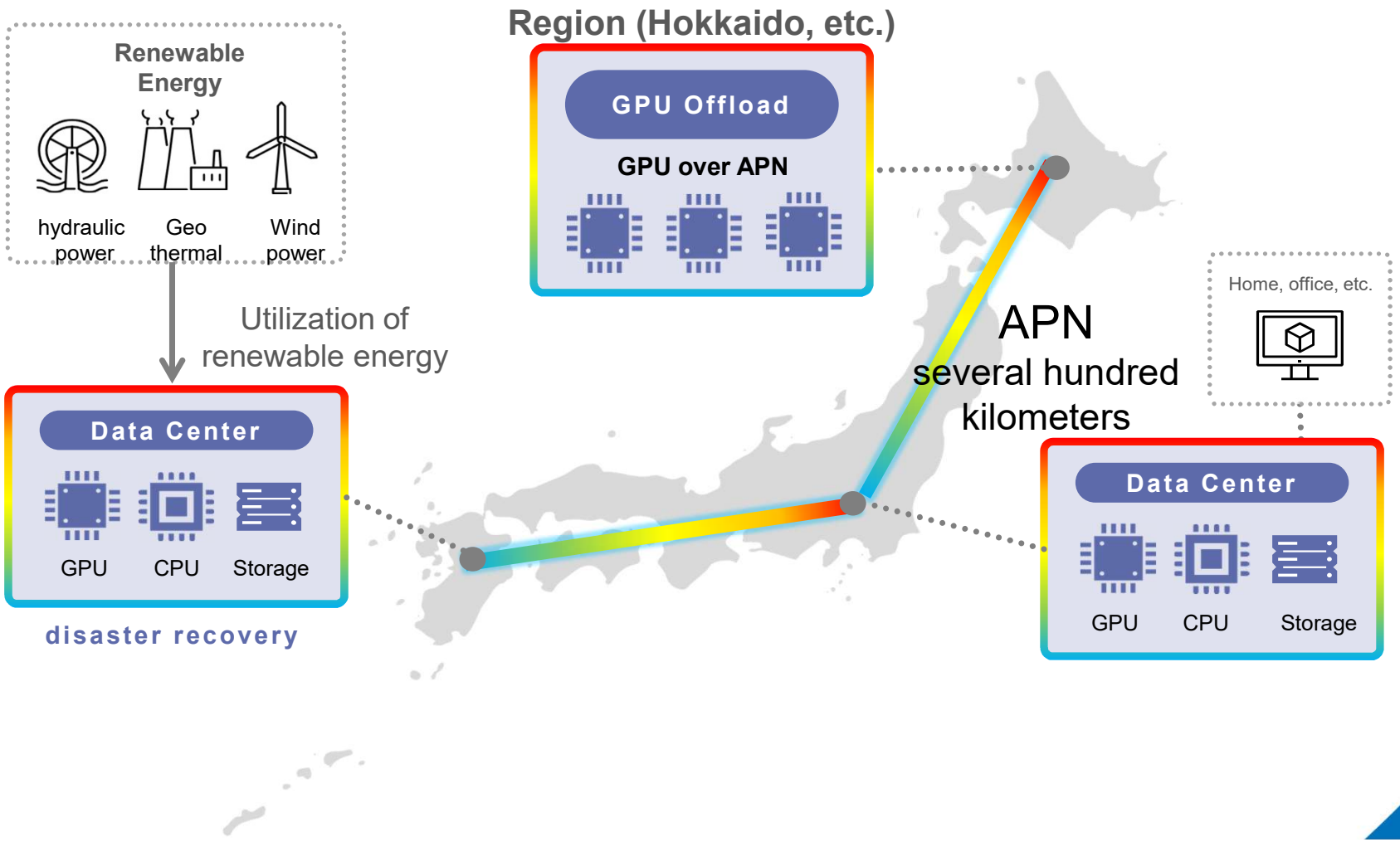
Exceeds floor load limits

-> "Unavoidably" to Distribute

Distributed Data Center x IOWN

docomo Business

Across distributed data centers Connect with IOWN APN



03 Latest Case Studies : GPU over APN

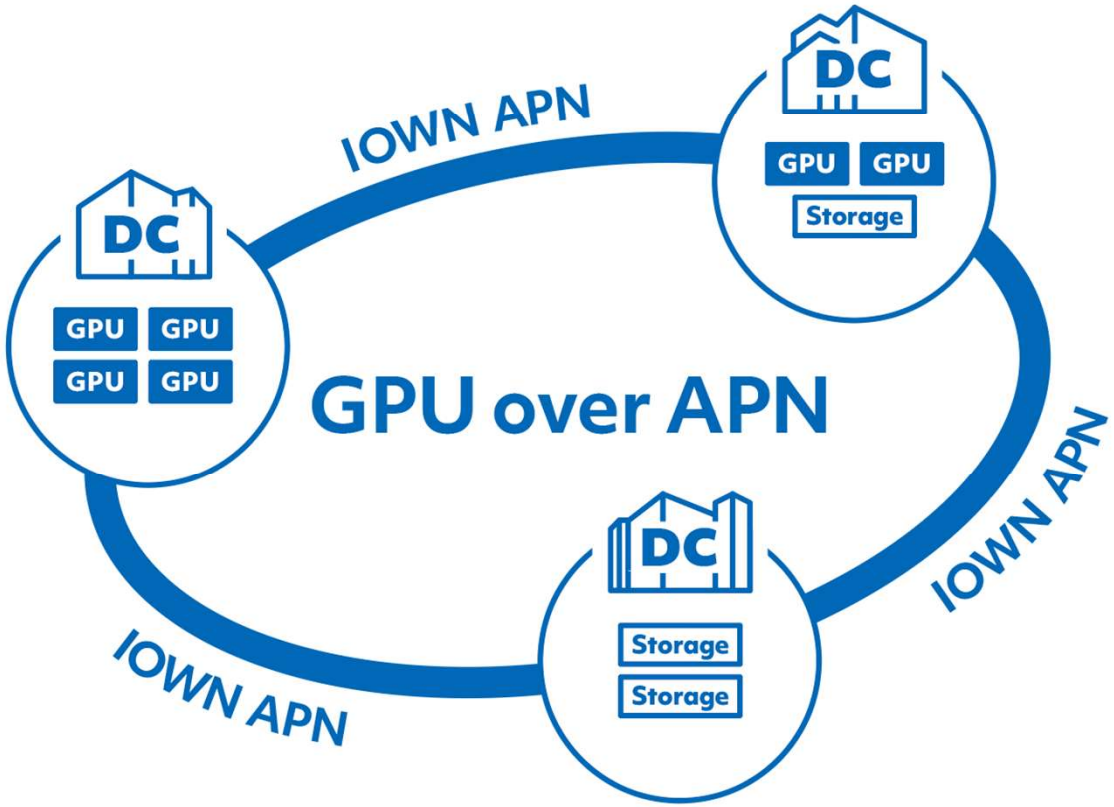
AI Training / AI Inference / Data Transfer

NTT docomo Business

The Concept of “GPU over APN”

docomo Business

Scale GPUs across Multiple Sites via All-Photonics Network



Ultra-High Bandwidth

Ultra-Low Latency

Ultra-Low Jitter

GPU over APN Use cases

docomo Business

A AI Training

B AI Inference

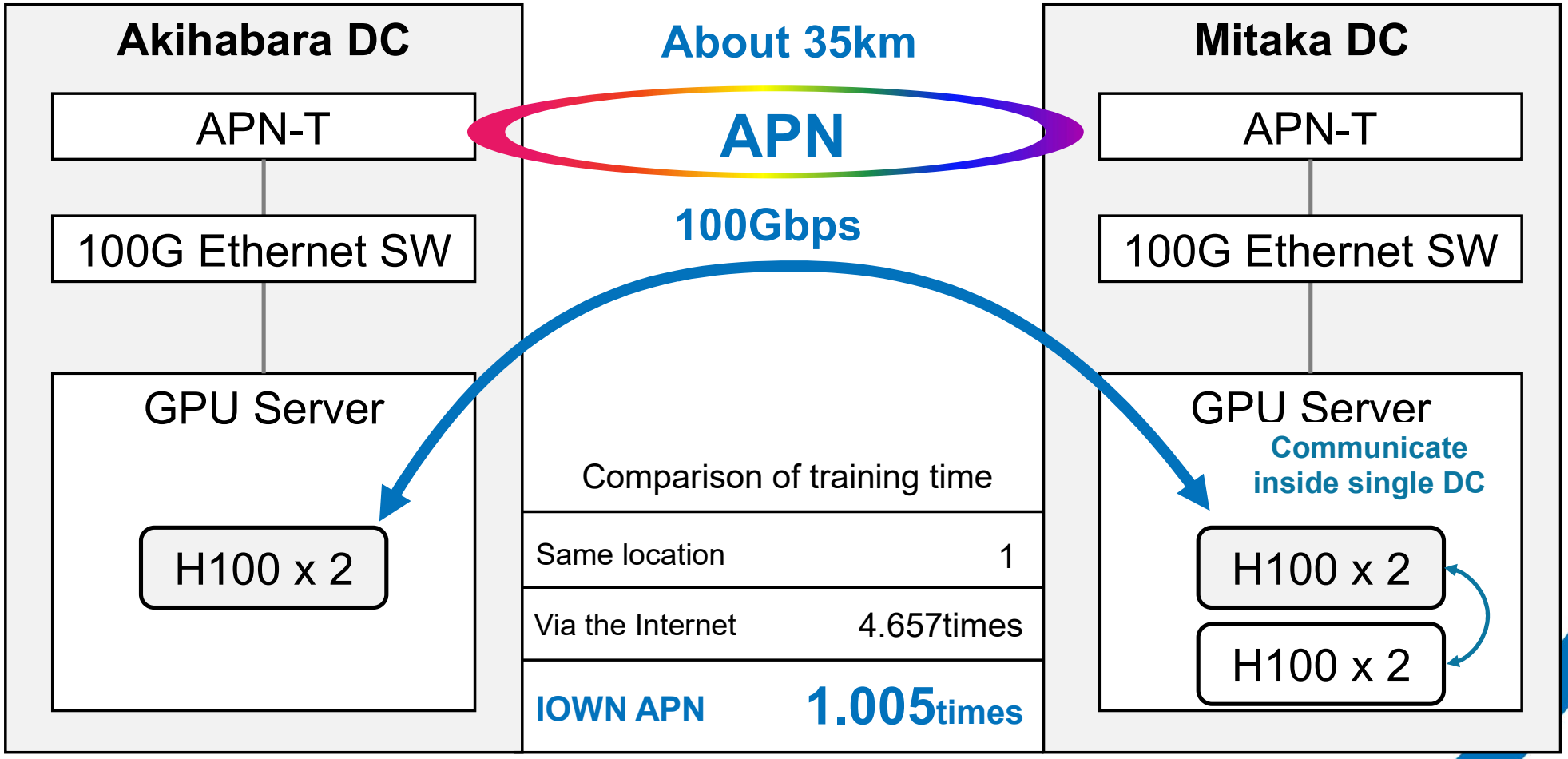
C Data Transfer

. . .

A Distributed AI Training (2 sites, Short distance)

docomo Business

Connect remote GPUs with APNs to achieve close to the same performance as having all the GPUs in the same location.

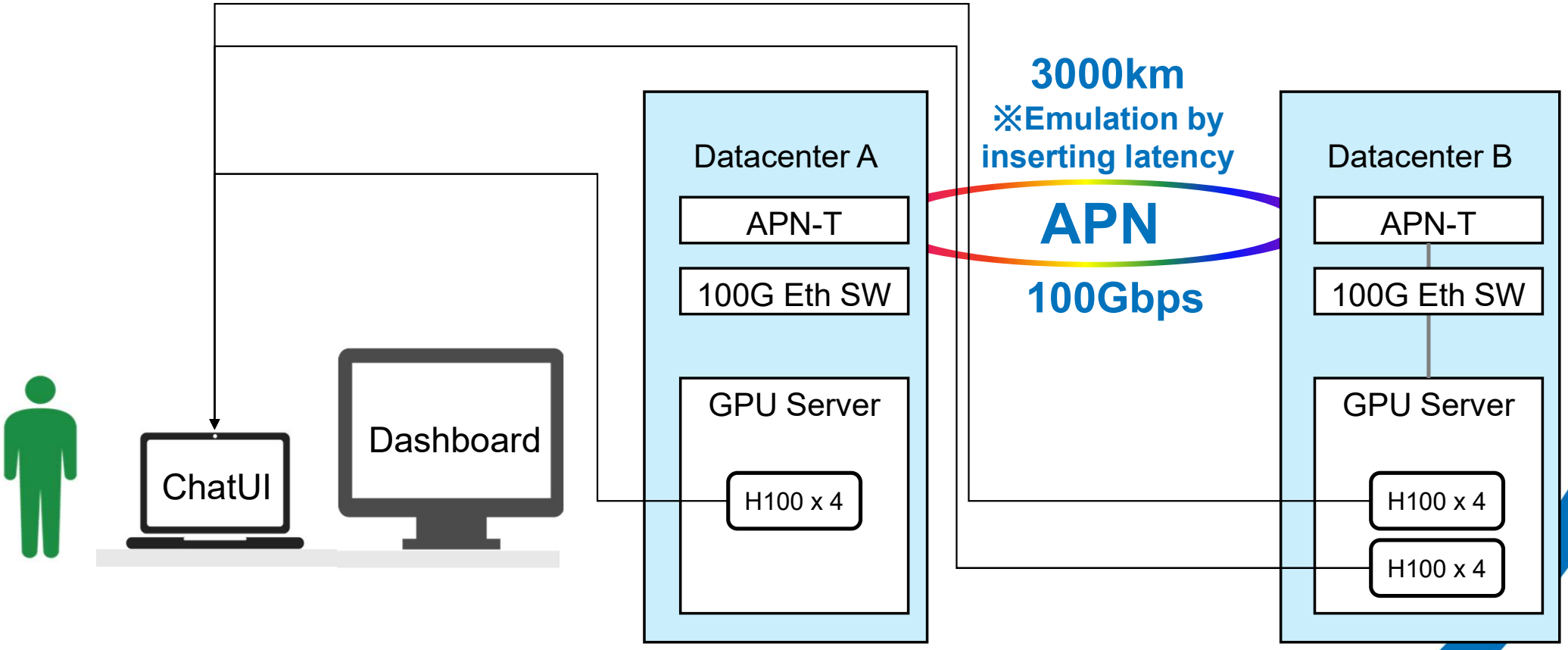


B Parallel inference over long distances

docomo Business

Assuming the inference time in a single data center is set to 1,

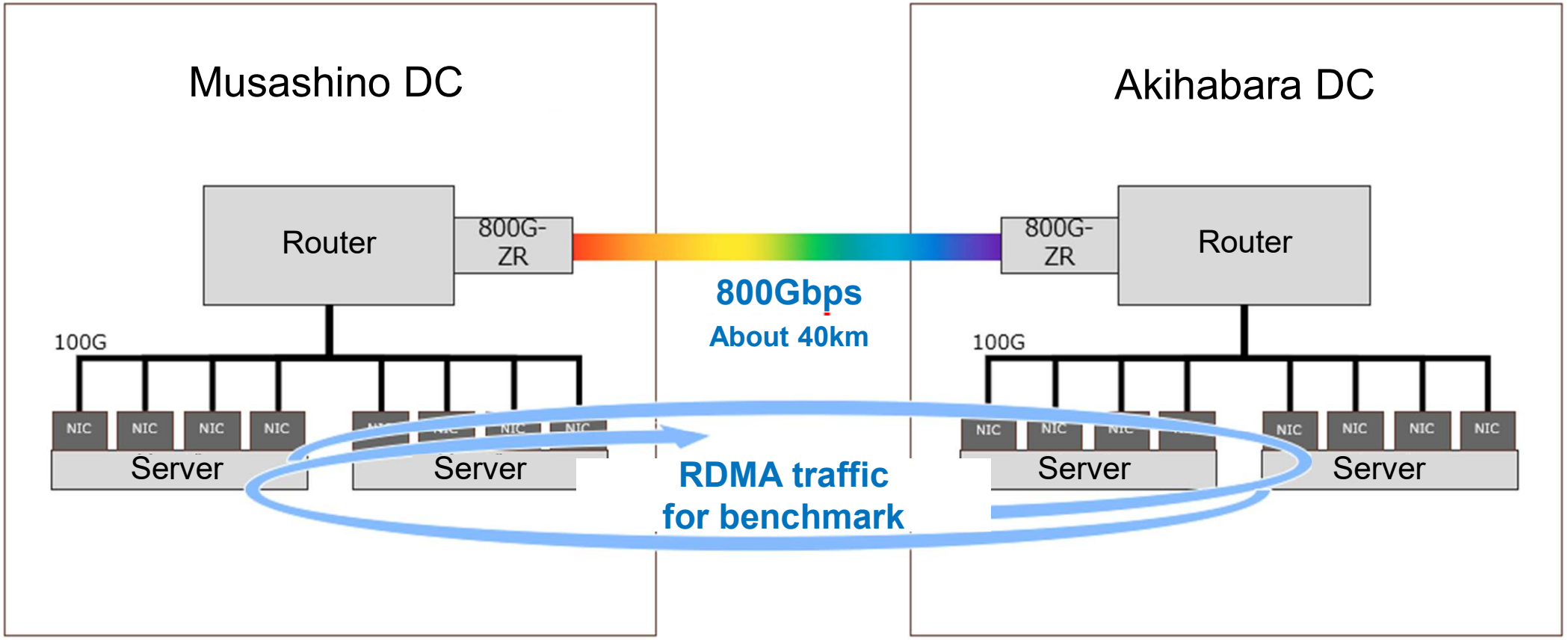
the distributed data center (via APN) takes approximately **1.009** times as long



C 800Gbps High-Bandwidth Data Transfer PoC

docomo Business

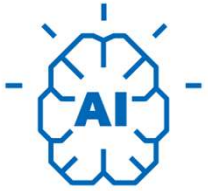
Transfer time: When using our RDMA transfer tool, the data transfer became **5.6** times faster compared to other methods.



Summary of the PoCs Results and Future Prospects

docomo Business

Distributed Training

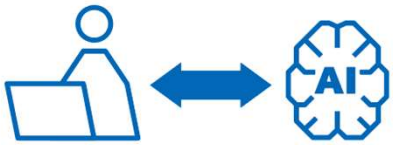


Training Time (LLM 7B)

1.005 x

vs. Single-DC

Distributed Inference

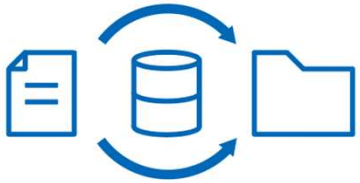


Inference Time (LLM 27B)

1.009 x

vs. Single-DC

Inter-Site Data Transfer



800G RDMA Transfer Speed

5.6 x

vs. Conventional TCP Tools

- Further expand “GPU over APN” environments and its possibilities
- Reflect the know-how gained from the experiment into our commercial services, aiming for providing customers a **one stop solution**.
- **Invite partners** who are interested and co-create PoC together.

NTT
docomo Business