

## Addressing Social Challenges with Industry AI Cloud Powered by IOWN

**Akira Shimada**  
*President and Chief Executive Officer,  
NTT Corporation*

### Abstract

This article presents NTT's research and development efforts to address social challenges through the Industry AI (artificial intelligence) Cloud powered by the Innovative Optical and Wireless Network (IOWN). It is based on the keynote speech given by Akira Shimada, NTT president and chief executive officer, at the "NTT R&D FORUM 2024 - IOWN INTEGRAL" held from November 25th to 29th, 2024.

*Keywords: IOWN, artificial intelligence, AI cloud*



### 1. The rapid rise in AI utilization

The emergence of generative artificial intelligence (AI) has accelerated the adoption of AI at an unprecedented rate. For instance, ChatGPT amassed 100 million users within 2 months—a stark contrast to Facebook, which took 54 months to achieve the same milestone. This highlights the extraordinary pace at which AI is advancing.

The AI market is projected to grow 20-fold from its 2021 levels, reaching 1.8 trillion dollars (approximately 280 trillion yen) by 2030 [1]. Within businesses, AI adoption has progressed significantly, particularly for improving productivity in general-purpose tasks.

According to a 2024 survey by Japan's Ministry of Internal Affairs and Communications, more than 90% of companies in the United States have already introduced AI, and about 60% of companies in Japan have done so. While AI is an essential tool for driving digital transformation (DX), its impact on creating new services and fundamentally transforming business models remains limited. This indicates that the full potential of AI-enabled DX has yet to be achieved.

### 2. The need for AI in specialized tasks

To enhance the value of existing services and create new ones, it is vital to leverage AI in more specialized business operations. However, these specialized tasks often involve diverse factors such as data types, formats, and legal requirements, which vary by industry.

The manufacturing sector, for example, handles computer-aided design data, while the healthcare industry relies on electronic medical records and genetic information. These types of data are often highly sensitive, and companies are cautious about their potential exposure. Strict guidelines, therefore, govern their management, particularly regarding distribution and protection.

In healthcare, sensitive data owned by individuals, medical institutions, and corporations must be securely collected and analyzed in a protected environment. Within the NTT Group, NTT Precision Medicine provides genetic testing services that collect genetic information to analyze disease risks, contributing to advanced healthcare solutions (Fig. 1).

At Prime Research Institute for Medical RWD, Inc. (PRiME-R), real-world data are collected and stored,

- Data owned by individuals, medical institutions, and companies are being collected separately, and secure data-analysis environments are being developed.
- NTT Group has also been advancing specific initiatives to address these needs.

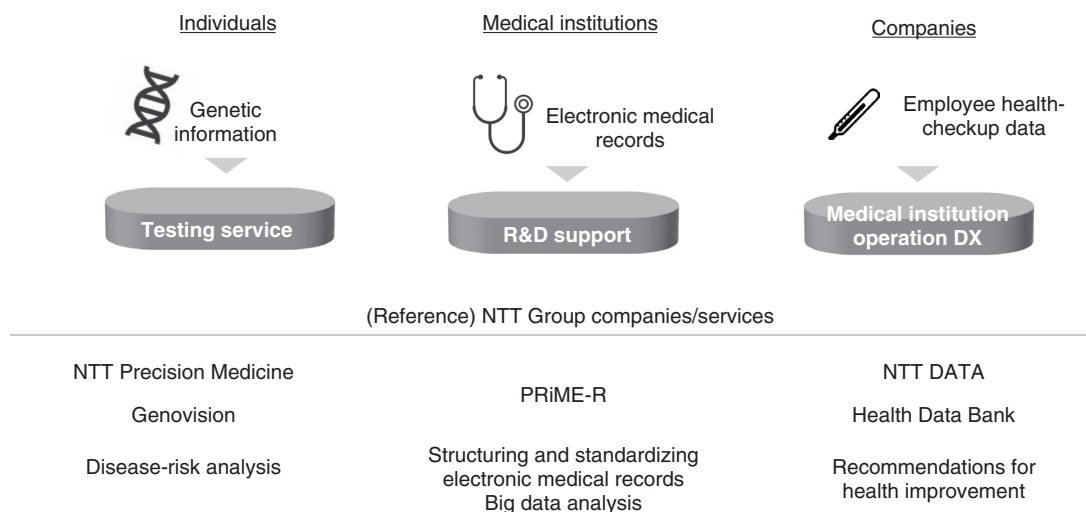


Fig. 1. AI application to specialized tasks: Case studies in the healthcare and medical fields.

while AI is used to structure electronic medical record data. These structured datasets are then analyzed to support research and development in fields such as pharmaceuticals. At NTT DATA, health-checkup information from employees is stored in the Health Data Bank, which also incorporates personal health records, such as step counts and sleep duration, to provide advice for promoting employee health.

However, these initiatives remain confined to their respective domains. To create new value—such as delivering personalized medicine or accelerating drug discovery—it is essential to aggregate independently collected data and analysis results into a shared platform for broader application (Fig. 2). In developing treatments for rare conditions, for example, it is crucial to efficiently identify patients with relevant medical histories and provide this information to pharmaceutical companies. Traditionally, pharmaceutical firms or their intermediaries have had to individually verify such cases with medical institutions.

By combining and analyzing multiple datasets, it becomes possible to achieve outcomes that were previously unattainable with isolated data. This integration can lead to enhanced productivity across the industry and foster innovation by transforming traditional processes.

### 3. Co-creation through the Industry AI Cloud

General business processes have begun leveraging AI to enhance efficiency, but introducing AI into specialized operations can significantly increase their added value. By accumulating data onto a shared platform across companies and applying AI, industries can also address social challenges collectively.

This initiative, referred to as the Industry AI Cloud, involves NTT collaborating with various industry partners to foster co-creation. Below are examples of projects within this framework.

#### 3.1 Co-creating the Mobility AI Platform for a zero-traffic-accident society

On October 31, 2024, NTT announced a partnership with Toyota Motor Corporation to embark on initiatives toward achieving a society with zero traffic accidents. This effort emphasizes the importance of a triadic collaboration among humans, mobility, and infrastructure and aims to develop the Mobility AI Platform to support this vision.

When vehicles continuously collect information about humans, infrastructure, and other vehicles, blind spots can be significantly reduced. With AI learning from this data, it becomes possible to predict human and vehicle behavior with high accuracy,

- The next step involves building a dedicated AI cloud and developing a platform for cross-functional collaboration to obtain new value such as personalized medicine and faster drug development.

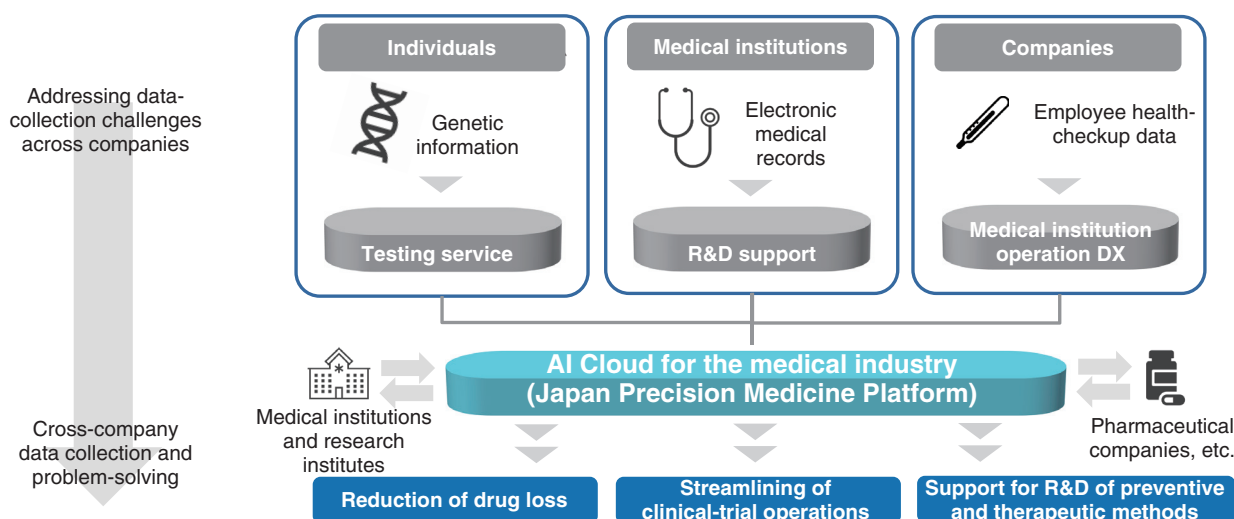


Fig. 2. Cross-industry value creation: Case study in medical industry.

enabling advanced driver assistance.

The Mobility AI Platform, for instance, could help prevent intersection collisions in urban areas, ensure smoother merging on highways, and provide autonomous driving services in rural areas to address mobility challenges. These advancements aim to enhance safety and security in diverse scenarios and expand the platform's implementation nationwide. Key components of the Mobility AI Platform are described below.

- (1) AI infrastructure: Mobility AI trained with industry-specific data

By developing AI models based on collected data, this infrastructure supports the implementation of services such as autonomous driving and AI agents, contributing to the creation of new values.

- (2) Decentralized computing infrastructure: Distributed datacenters

To manage the projected 22x increase in communication volume and 150x increase in computational demands by 2030, NTT plans to deploy datacenters across Japan using its Innovative Optical and Wireless Network (IOWN) technology.

NTT currently possesses and is expanding its development infrastructure for graphics processing units (GPUs) that power large language models (LLMs) such as tsuzumi. With the growth of the Industry AI Cloud, this GPU-development infrastruc-

ture will be extended to support various industries, including mobility.

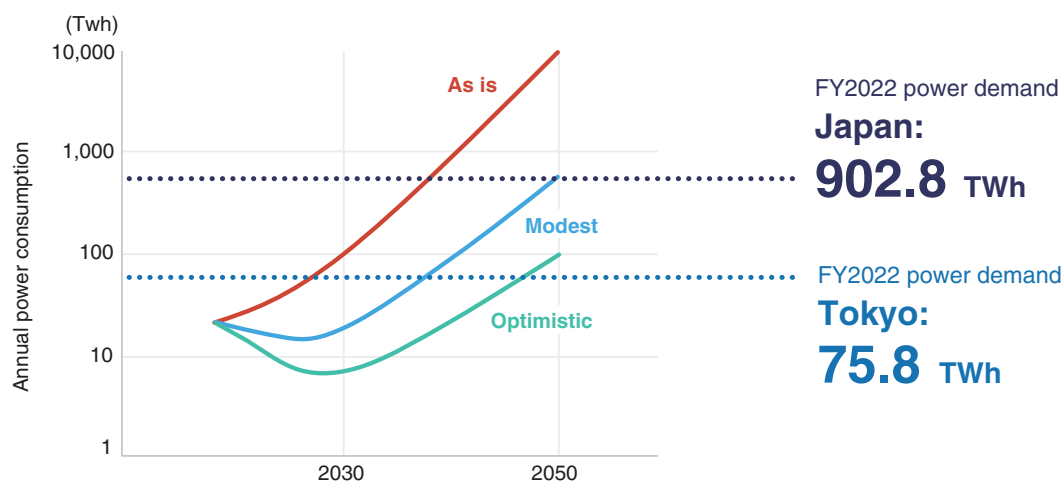
- (3) Intelligent communication infrastructure: Reliable and seamless connectivity

The final component is the intelligent communication infrastructure. By leveraging technologies such as the low-latency, high-capacity, and energy-efficient All-Photonics Network (APN) provided by NTT, AI can optimize communication, enabling seamless connectivity between humans, mobility, and infrastructure while facilitating the collection of various data. Supporting the intelligent communication infrastructure requires a variety of technologies. Cradio®, a technology that ensures consistently reliable wireless environments, is one such innovation. By predicting network quality and implementing appropriate environmental controls, it enables an optimal and natural communication environment at all times, even in dense autonomous vehicle networks.

### 3.2 Preemptive supply and demand matching through optimization of agricultural product transactions

NTT is also working on improving the supply-demand balance by optimizing agricultural transactions. Trade information in wholesale markets remains largely undigitized, with face-to-face transactions

- By 2030, the electricity demand for ICT infrastructure in Japan may exceed Tokyo's current demand, and by 2050, it may surpass Japan's total electricity demand.



Source: Created based on the "Impact of Information Society Advancement on Power Consumption Vol. 5 (February 2023)" by the Japan Science and Technology Agency (JST), National Research and Development Agency - Domestic Network Power Consumption Forecast

Fig. 3. Power consumption forecast for ICT infrastructure in Japan.

still common across Japan. This inefficiency often leads to mismatched supply and demand, increased transportation costs, reduced agricultural product quality, and food loss. Through AI-driven simulations of supply-demand and delivery planning in virtual wholesale markets, NTT aims to optimize these transactions. By aligning supply and demand in advance using collected data, inefficiencies can be mitigated. This initiative is currently undergoing pilot testing.

In August 2024, NTT launched NTT AI-CIX (AI-Cross Industry Transformation), a company dedicated to industrial efficiency and transformation through AI. This company is leading collaborations with Trial, a supermarket chain, to initially focus on shelf-space optimization and automated ordering using AI. By connecting individual AI systems, the ultimate goal is to optimize the entire supply chain of the retail and distribution industry.

#### 4. Toward a sustainable future

As AI adoption accelerates across various domains, increased power consumption will become an issue. AI computational infrastructure, built on large-scale datacenters, consumes significantly more energy than conventional servers. By 2030, it is estimated that

energy consumption by Japan's domestic datacenters alone will exceed the total energy demand of Tokyo in 2022 and exceed Japan's total energy demand by 2050 (Fig. 3).

To address this growing demand, NTT is actively pursuing research and development to create a sustainable AI-powered society. This effort includes developing tsuzumi, a lightweight, low-power AI model and implementing the IOWN 2.0 energy-efficient computing infrastructure.

In 2023, NTT launched tsuzumi, an original, lightweight LLM optimized for Japanese, and commercial deployment began in March 2024. Depending on the type of data handled in various industries, lightweight models such as tsuzumi are often better suited to specific use cases.

Since its commercialization, tsuzumi has undergone continuous advancements. To meet diverse needs, efforts are underway to enable tsuzumi to process and interpret visual data, such as photos and graphs. Beyond improving its comprehension capabilities, research is focused on enabling tsuzumi to interact with web interfaces and systems as well as executing tasks collaboratively with humans. For example, if a user requests tsuzumi to order a product from an e-commerce site, the model can navigate the website, locate the desired item, complete the purchase,

- NTT's LLM "tsuzumi" has continued to evolve since its commercialization and is now able to understand visual information. Further performance improvement is in progress.

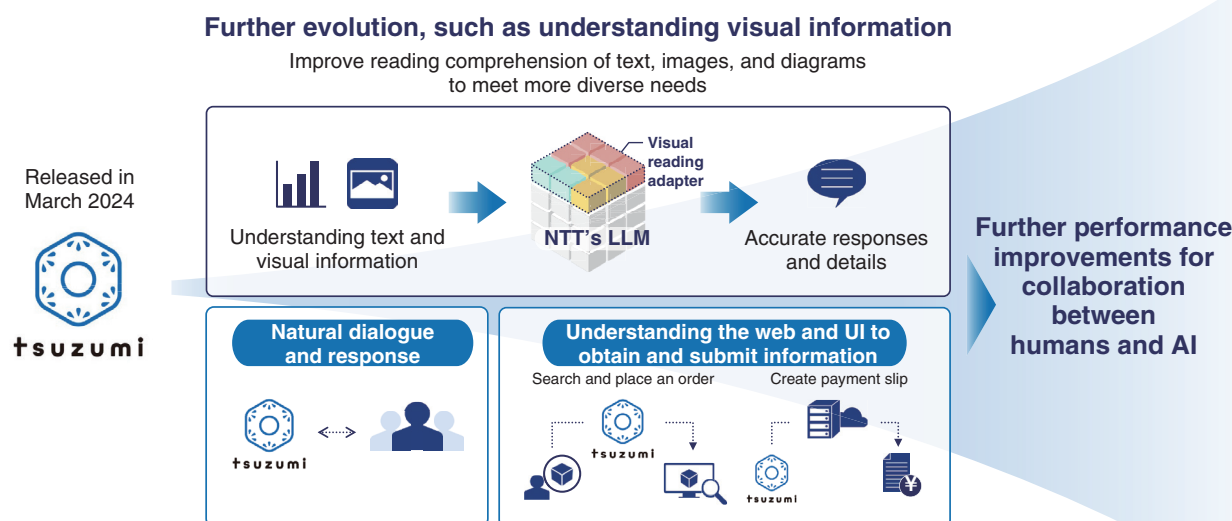


Fig. 4. Advancement of tsuzumi.

and subsequently access internal company systems to generate the necessary payment documents (Fig. 4).

The evolution of tsuzumi continues, with plans to incorporate user feedback into future updates. In a remarkable achievement, tsuzumi recently became the first Japanese LLM to be included in Microsoft's Models-as-a-Service lineup. This milestone was announced at Ignite, Microsoft's tech conference held in Chicago, and tsuzumi became available for use starting November 20, 2024.

## 5. Progress of low-power computing with IOWN

The rollout of IOWN 2.0 is slated for 2025, marking a significant step toward low-power computing. This initiative integrates photonics-electronics convergence (PEC) devices into computing infrastructure to reduce the energy consumption of servers and APN-transmission equipment.

The computing framework for the IOWN era is termed data-centric infrastructure (DCI), which reduces power consumption through the following two key approaches.

### (1) Disaggregated computing

In the conventional architecture, servers contain multiple components within a single unit. Disaggregated computing, in contrast, involves separating these components and enabling shared use. This enables optimal combinations of central processing

units (CPUs), memory, and other components based on actual needs while powering down unused components to conserve energy. The concept can be likened to breaking down and reorganizing server components for greater efficiency.

### (2) PEC

This approach replaces traditional electronic circuits that process electrical signals with optical waveguides, reducing the energy used for electrical wiring (Fig. 5).

## 6. From the IOWN concept to reality and beyond

As part of the IOWN 2.0 initiative, NTT is advancing the development of DCI-2 (DCI version 2) with the goal of commercial deployment by around 2026. DCI-2 leverages composable disaggregated infrastructure servers, which break down computing resources into board-level components. These components are interconnected using optical switches powered by PEC devices, all optimized by a DCI controller. The target is to achieve a dramatic one-eighth reduction in power consumption.

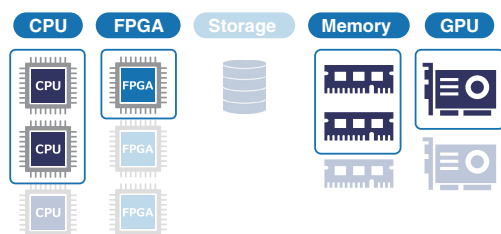
A major milestone in this effort will be showcased at the Expo 2025 Osaka, Kansai, Japan in April 2025, where NTT plans to implement these PEC-based servers in its pavilion. This marks a significant step toward actualizing the IOWN concept, offering visitors



## ■ Reducing power consumption of servers through two approaches

### Disaggregated computing

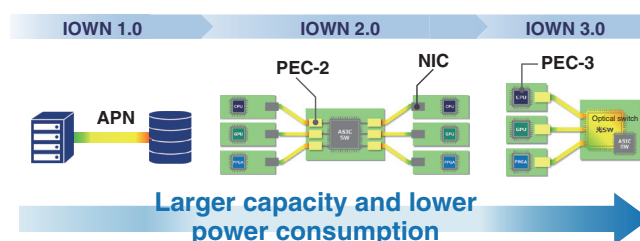
#### Optimize computing resources



- Use only the required number of components, such as CPUs and memory, by combining them as needed.
- Power off unused components to reduce power consumption.

FPGA: field programmable gate array  
NIC: network interface card

### Converting electrical wiring to optical wiring



- Connecting boards and chips through optical communication reduces power consumption previously used by electrical wiring.

Fig. 5. Low-power computing resources through DCI.

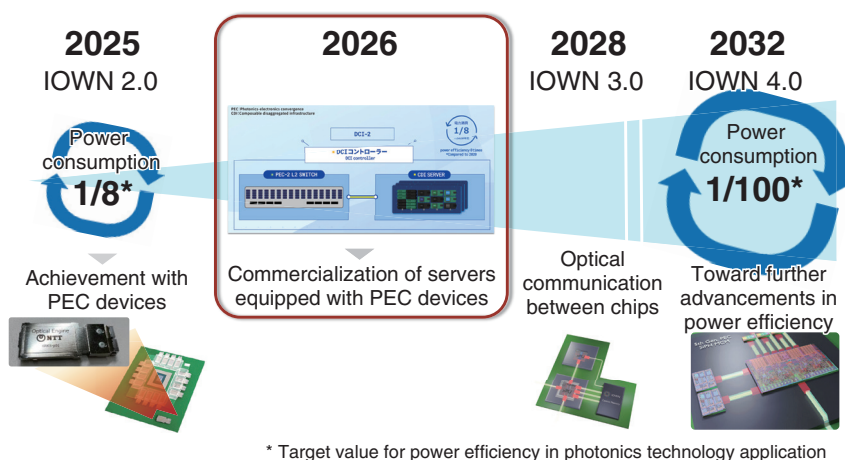


Fig. 6. IOWN from concept to reality and beyond.

the chance to witness these next-generation servers in action. Commercialization of the server is anticipated by 2026.

Looking further ahead, NTT aims to achieve optical communication between chips by 2028 and extend this innovation to intra-chip communication by 2032, ultimately targeting a 100-fold reduction in power consumption (Fig. 6).

By building on IOWN's cutting-edge technology

and harnessing the potential of the Industry AI Cloud, NTT is committed to addressing social challenges in a sustainable and transformative manner.

### Reference

- [1] Ministry of Internal Affairs and Communications, "2024 White Paper on Information and Communications in Japan," Chapter 9, 2024.