Innovating a Sustainable Future for People and Planet NTT R&D Initiatives

Akira Shimada
President and Chief Executive Officer, NTT Corporation

Abstract

This article presents the research and development (R&D) activities of NTT as it continues to innovate a sustainable future for people and the planet. It is based on the keynote speech given by Akira Shimada, president and chief executive officer of NTT Corporation, at the "NTT R&D FORUM 2023 — IOWN ACCELERATION" held from November 14th to 17th, 2023.

Keywords: IOWN, All-Photonics Network, large language model

1. Major challenges facing society

I would like to point out three of the major challenges facing society today.

The first challenge is the severe labor shortage. In addition to the decline in workforce, we in Japan are also facing the so-called "Year 2024 Problem" resulting from the enforcement of new overtime regulations, which has become a major issue in the construction and transportation industries.

The second challenge is the environmental impact of energy consumption that has become a global issue. The dramatic increase in data volume has led to a surge in electricity consumption, and energy demand, especially in urban areas, is growing. We need to harmonize the addressing of environmental and energy issues without stopping the progress of technological innovation.

The third challenge is that, with the advent of an aging society, rising healthcare costs have become a major factor contributing to the strain in Japan's fiscal situation. It is also necessary to create a well-being society that enables various people to live a healthy and fulfilling life.

We aim to address these challenges through NTT's research and development (R&D) centered on the Innovative Optical and Wireless Network (IOWN), a next-generation communication and computing infrastructure that achieves high capacity, low latency, and low power consumption, and NTT's large language model (LLM) "tsuzumi," a compact and power-saving large-scale language model with world-class language-processing capabilities.

2. IOWN

Our ultimate goal for IOWN is to increase power efficiency by 100 times and transmission capacity by 125 times and to reduce end-to-end delay by 1/200th.

As a roadmap for IOWN, we started the commercial launch of IOWN1.0 at the end of fiscal year (FY) 2022 (**Fig. 1**). I will explain some use cases of IOWN1.0 later. We plan to develop photonics-electronics convergence (PEC) devices for inter-board connection by FY2025 as IOWN2.0. Subsequently, we will develop a device for inter-chip connection as IOWN3.0 by FY2028 and aim to achieve intra-chip connectivity with PEC as IOWN4.0 by FY2032.

The important point of IOWN2.0 is to apply PEC devices to computing (**Fig. 2**). A high-capacity, low-power, and compact optical engine is key to achieving this goal. Using this optical engine and a switchboard equipped with the optical engine, the xPU (x processing unit) and memory can be connected with optics instead of electricity to achieve ultra-low power consumption IOWN computing. IOWN computing

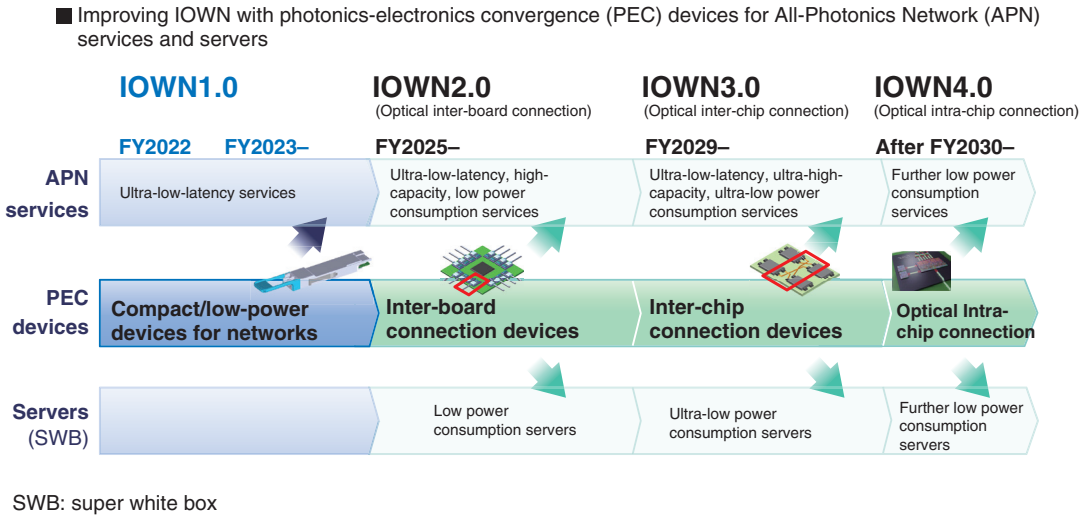

Fig. 1. Roadmap for IOWN.

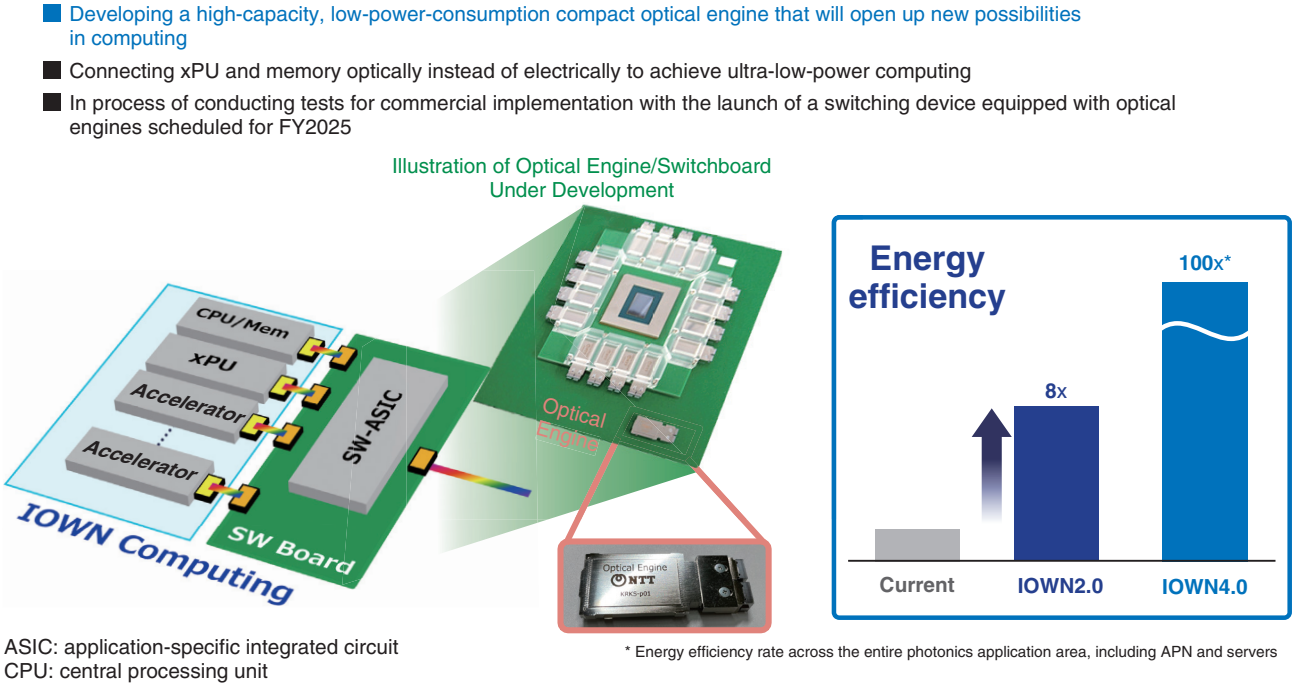

Fig. 2. IOWN2.0 – Optical-based computing.

can increase power efficiency by up to eight times compared with conventional computing.

The development of the optical engine is basically complete, and tests are being conducted for its commercial use. In FY2025, we plan to start providing switchboards equipped with optical engines. We are planning to let you experience such service using IOWN2.0 at EXPO 2025 Osaka, Kansai, Japan. The theme of NTT Pavilion is "Architecture with Emotion" (**Fig. 3**). The "living pavilion" will be represented by "cloth" covering the pavilion that moves according to the excitement of the visitors and

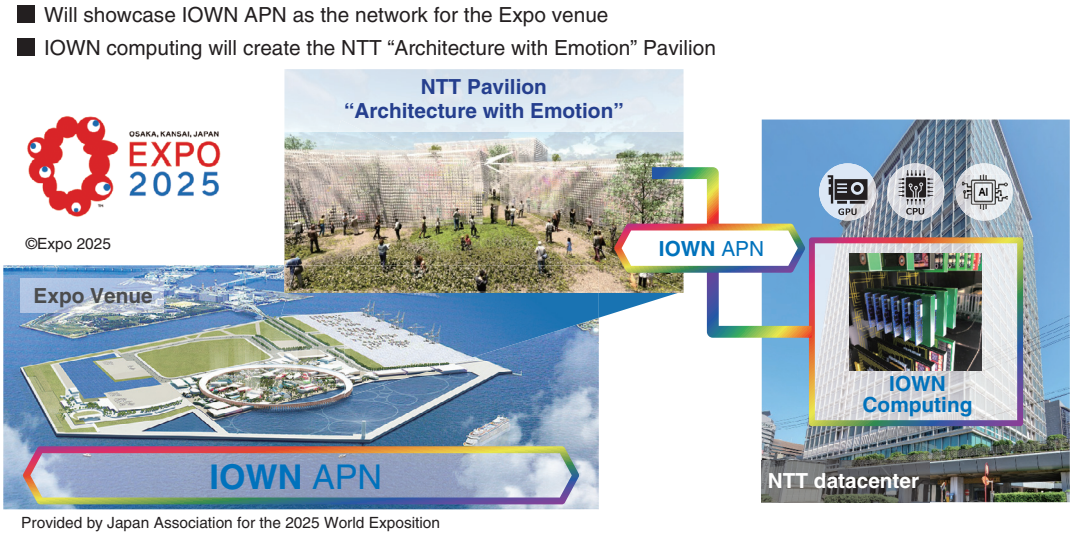

Fig. 3. IOWN at Expo 2025 Osaka, Kansai, Japan.

changes expression according to natural light and wind. These are achieved through remote artificial intelligence (AI) analysis using the All-Photonics Network (APN) and IOWN computing.

3. NTT's LLM "tsuzumi"

Next, I would like to introduce NTT's LLM, "tsuzumi," the result of over 40 years of research and expertise on natural-language-processing technology. This LLM has four key features.

The first key feature is its linguistic capabilities. It supports Japanese as well as English and a variety of other languages. We are very proud of its world-class performance in a variety of Japanese benchmarks.

The second key feature is its high level of cost performance. Demonstrating low power consumption and high graphics processing unit (GPU) performance while having the same high performance as OpenAI's Generative Pre-trained Transformer 3 (GPT-3) results in its sustainability.

The third key feature is its low cost of tuning. It is capable of frequent information updates and customization based on industry- and organization-specific data.

The fourth key feature is its functionality with various input formats such as diagrams, charts, and tables. It is the first Japanese model that can read contracts and invoices containing tables.

We compared tsuzumi's Japanese language capabilities with LLMs of other companies. It has world-class performance in Japanese, which is better than OpenAI's GPT-3.5 and significantly exceeds other domestic LLMs of the same class. It even shows English capabilities equivalent to Meta's world-class LLM and is capable of handling other languages.

A comparison of cost performance, a key feature mentioned earlier, with GPT-3-scale LLMs is shown in **Fig. 4**. Because it requires fewer GPUs, tsuzumi is able to achieve similar performance as GPT-3-scale LLMs with 1/25th the hardware costs for training. It also requires only 1/20th the cost for use. It uses less power because it requires fewer GPUs.

LLM tsuzumi will be launched in March 2024. We began expansive internal and external trials in October 2023 and are already starting to see the results. Beginning in April 2024, it will not only be able to read documents and graphics but will also be able to recognize voices and tones, such as children's voices, and will have successive releases in other languages in addition to Japanese and English.

4. Addressing social challenges using NTT's R&D technologies and services

I will now introduce our efforts to address social challenges using NTT's R&D technologies and services such as IOWN and tsuzumi.

4.1 Addressing the severe labor shortage

In the construction industry, labor shortages, long working hours, and the aging of engineers are

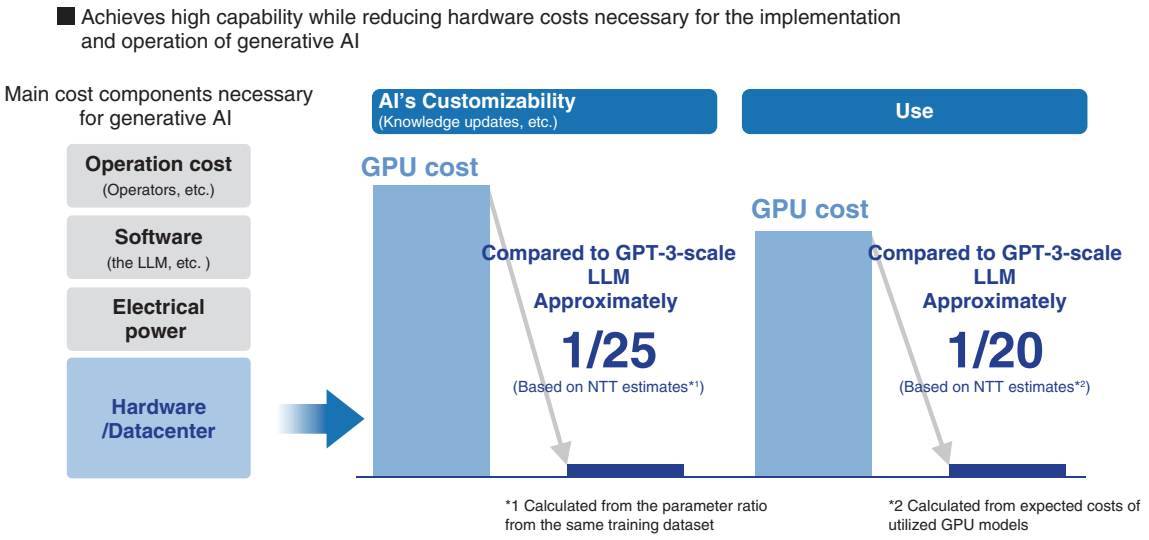

Fig. 4. High level of cost performance.

becoming more serious. Since the upper limit on overtime work will be enforced from 2024 in Japan, work-style reforms such as operational streamlining and employment diversification are required. In response, NTT, in cooperation with EARTHBRAIN, JIZAIE, and Takenaka Corporation, is promoting remote control of construction equipment for efficient and safe construction. Specifically, by using the IOWN APN—which has high-capacity, low-latency, and no lag fluctuations—in remote operations, it will become possible to operate construction equipment as if they were being operated onsite.

I will present a case study on remote content and broadcast production, which we are currently pursuing with Sony. Whenever there had been a match or event held at a stadium anywhere in the country, large-scale production spaces, personnel, equipment, broadcast vehicles, and other production items have always been needed. By using the APN to connect broadcast stations to stadiums around the country, it will become possible to achieve remote content production instead, which would reduce the amount of space, personnel, equipment, broadcast vehicles, and other production items needed at the time of the event. On November 13, we entered into an agreement with Sony to collaborate further on this.

I will now present another case study involving collaboration with Tokio Marine & Nichido Fire Insurance. We are aiming to improve the productivity of contact centers by using tsuzumi. Tokio Marine & Nichido Fire Insurance has more than 10,000 operators in the accident-response department nationwide who provide daily support for non-life insurance. Operators listen carefully to the circumstances of the accident and injuries on the phone, and after the call, they organize their responses and input necessary information into the system. This after-call work takes about 800,000 hours per year. We have already made small reductions through voice mining, etc., but by combining tsuzumi with voice mining, we can make progress in summarizing and organizing the content of the correspondence and expect to reduce the operation of after-call work by more than 50%.

I will talk about autonomous driving systems. Regarding public transportation, such as local buses and taxis, the shortage of drivers in the regions has become apparent. Autonomous driving technology is expected to address various social issues. NTT has invested in May Mobility, a U.S.-based company that has strengths in autonomous driving technology, and have acquired exclusive rights to sell their autonomous driving solutions in Japan. Through cooperation with several local governments facing transportation issues, we will first provide services through community buses then expand to many other types of autonomous vehicles to address various social issues, including driver shortages.

4.2 Addressing the environmental impact of energy consumption

In a data-driven society, huge amounts of power are required to process rapidly increasing amounts of

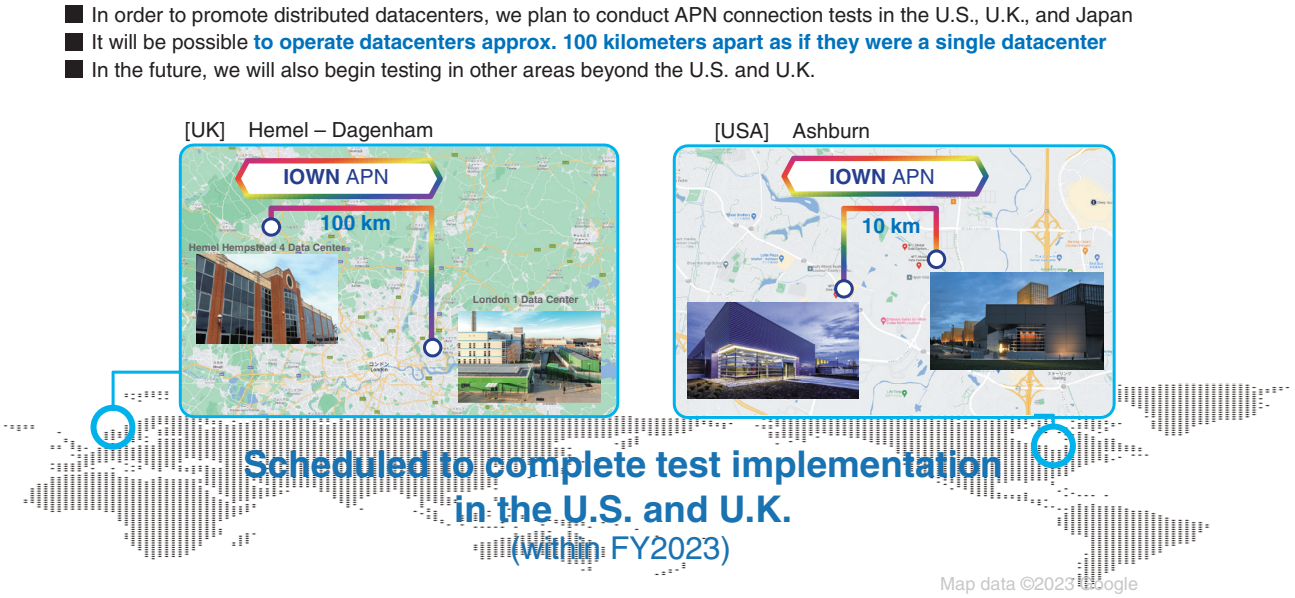

Fig. 5. Creating distributed datacenters – Implementing the APN between datacenters in overseas markets.

data. For example, the power consumption of datacenters is expected to increase by approximately 6 times in Japan and 13 times in the world from 2018 to 2030 as the volume of data handled increases. AI will continue to grow and expand, but LLMs such as ChatGPT require as much as 1300 MWh of power for one training session. This is equal or greater than the amount of electricity generated by operating one nuclear power plant for one hour. In this data-driven society, the need for datacenter computing will continue to grow and consume more power than ever before. NTT is addressing this expanding power-consumption problem with distributed datacenters using the APN. I introduce three use cases of how distributed datacenters can be used.

The first case study is using the APN to connect datacenters in Japan for use in training LLMs. To train tsuzumi, we used the APN to build a collaborative cloud and on-premises environment. We have a large volume of training data at our Yokosuka laboratory, but due to power issues, we found that it was difficult to install GPU equipment in that area. We therefore used the APN to connect the GPU cloud in our datacenter in Mitaka with our training data storage in Yokosuka to conduct the training. Therefore, we were able to create an environment that was completely comparable with the local environment.

The second case study is in collaboration with Oracle. We are currently conducting testing to use the APN to connect Oracle's cloud with NTT's datacenters. This will be possible to keep important data at hand while linking only the data necessary for analysis to the cloud in real time.

The third case study is an example of implementing the APN between NTT datacenters overseas. To achieve distributed datacenters, we are advancing APN-connectivity test preparations between datacenters, not only in Japan but also overseas, initially in the United Kingdom and United States. In the U.K. for example, it will be possible to operate a datacenter in London and another approximately 100 kilometers outside the city through transmission lines as if they were a single datacenter (**Fig. 5**). We plan to complete this testing during FY2023. We also plan to expand this to other regions of Asia and other areas beyond the U.S. and the U.K.

4.3 Addressing rising healthcare costs due to an aging society and the pursuit of a society with greater well-being

I would like to talk about the challenges in the medical field. While the introduction of digital patient records has been advancing in Japan, patient record-keeping methods for even the same symptoms, for example, differ among hospitals and doctors, making it extremely difficult to collect and use patient record data. As tsuzumi is lightweight, flexible, and capable of learning patient data securely, it

can interpret medical data recorded by doctors and arrange them in appropriate expressions and a uniform format to make data more suitable for analysis. At Kyoto University Hospital, tsuzumi is already being used to structure the data from digital patient records. Dr. Manabu Muto from Kyoto University said, "As digital patient record structuring and analysis advances by utilizing tsuzumi, it becomes possible to deliver effective personalized medical treatment for each individual, or what is called precision medicine." This will lead to the optimization of medical expenses across society as a whole. With structured digital patient record data, it will also be easier to analyze medical data relating to the effects and side effects of medication. We believe this will lead to the effective development of pharmaceuticals by reducing development time and costs.

Finally, I introduce our efforts to achieve a well-being society. DJ MASA, or Mr. Masatane Muto, is a former employee of the advertising firm Hakuhodo and has been active in music as a DJ. In 2014, at the age of 27, he was diagnosed with amyotrophic lateral sclerosis (ALS), an incurable disease in which the motor nerves that enable the body to move begin to break down, leading to gradual loss of movement. One can hear but is unable to respond. His first thought was, "Is this it, is my life over, why me?" Then he thought, "Even if my body is disabled, there has to be a way that I can express myself with technology." He now participates in many events as a DJ by playing music with gaze-control. NTT wanted to collaborate with him, so we asked, "What would you like to do if you could move your body?" To which he responded, "I would like to party with the audience." To make this a reality, we combined the virtual and real worlds to enable him to engage the audience using an avatar. At the Ars Electronica Festival, a world-renowned media art festival, DJ MASA communicated with the audience in English using his voice and performed live through an avatar. The audience responded enthusiastically when the avatar's hands were raised by him; thus, the performance was a great success.

NTT is researching and developing technologies that will give people with even serious disabilities, such as ALS, the ability to communicate. For people with serious disabilities, physical expressions can be difficult, but by using motor-skill-transfer technology that can respond to even small amounts of muscle movements and brainwaves, an avatar can be used to produce physical expressions. For people who have lost their ability to speak, we are working on cross-lingual speech-synthesis technology that can synthesize the voice they lost and make it possible for them to not only converse with their voice in Japanese but also in English and other languages.

Going forward, NTT will continue to take on challenges to innovate a sustainable future for people and the planet.