

NVIDIA and SoftBank Corp. Accelerate Japan's Journey to Global AI Powerhouse

- *SoftBank Building Nation's Most Powerful AI Supercomputer With NVIDIA Blackwell for Wide Range of Sovereign AI Initiatives, Announces Plans for Grace Blackwell*
- *NVIDIA AI Aerial Enables SoftBank to Build World's First Live 5G AI-RAN, Unlocking Billions of Dollars in New Revenue Opportunities for Global Telco Industry*
- *SoftBank Uses NVIDIA AI Enterprise to Create AI Marketplace, Meeting Nation's Demand for Local, Secure AI Compute*

NVIDIA AI Summit Japan—NVIDIA today announced a series of collaborations with SoftBank Corp. designed to accelerate Japan's sovereign AI initiatives and further its global technology leadership while also unlocking billions of dollars in AI revenue opportunities for telecommunications providers worldwide.

During his keynote at NVIDIA AI Summit Japan, NVIDIA founder and CEO Jensen Huang announced that SoftBank is building Japan's most powerful AI supercomputer using the [NVIDIA Blackwell](#) platform and has plans to use the NVIDIA Grace Blackwell platform for its next supercomputer.

Additionally, NVIDIA revealed that SoftBank, using the [NVIDIA AI Aerial](#) accelerated computing platform, has successfully piloted the world's first combined AI and 5G telecom network — a breakthrough in computing that opens AI revenue streams potentially worth billions of dollars to telecom operators.

NVIDIA and SoftBank also announced that, using [NVIDIA AI Enterprise](#) software, SoftBank is aiming to create an AI marketplace that can meet the demand for local, secure AI computing. This new service, which supports AI training and edge AI inference, positions SoftBank to become the AI grid for Japan, facilitating new business opportunities for the creation, distribution and use of AI services across the country's industries, consumers and enterprises.

"Japan has a long history of pioneering technological innovations with global impact," said Huang. "With SoftBank's significant investment in NVIDIA's full-stack AI, Omniverse and 5G AI-RAN platforms, Japan is leaping into the AI industrial revolution to become a global leader, driving a new era of growth across the telecommunications, transportation, robotics and healthcare industries in ways that will greatly benefit humankind in the age of AI."

"Countries and regions worldwide are accelerating the adoption of AI for social and economic growth, and society is undergoing significant transformation," said Junichi Miyakawa, president and CEO of SoftBank. "Through our long collaboration with NVIDIA, SoftBank is leading this transformation from the forefront. With our extremely powerful AI infrastructure and our new, distributed AI-RAN solution 'AITRAS' that reinvents 5G networks for AI, we will accelerate innovation across the country and throughout the world."

SoftBank First to Receive Blackwell, Plans for Grace Blackwell

SoftBank is slated to receive the world's first [NVIDIA DGX™ B200 systems](#), which will serve as the building blocks for its new [NVIDIA DGX SuperPOD™](#) supercomputer.

SoftBank plans to use its Blackwell-powered DGX SuperPOD for its own generative AI development and AI-related business, as well as that of universities, research institutions and businesses throughout Japan.

Upon completion, SoftBank's DGX SuperPOD is expected to be Japan's most performant to date. Featuring NVIDIA AI Enterprise software and [NVIDIA Quantum-2 InfiniBand networking](#), it is also ideal for the development of large language models.

In addition to its DGX SuperPOD, SoftBank plans to build another NVIDIA-accelerated supercomputer to run extremely compute-intensive workloads. Initial plans for the supercomputer are based on an NVIDIA Grace Blackwell platform design featuring [NVIDIA GB200 NVL72](#) multi-node, liquid-cooled, rack-scale systems that combine NVIDIA Blackwell GPUs with power-efficient Arm-based NVIDIA Grace™ CPUs.

AI-RAN Reaches New Milestone

Working closely with NVIDIA, SoftBank has achieved a technology milestone — the development of a new kind of telecommunications network that can run AI and 5G workloads at the same time, known by the industry as artificial intelligence radio access network, or [AI-RAN](#).

This new breed of infrastructure has broad ecosystem support from the telecom industry, as it offers operators the ability to transform their base stations from cost centers into AI revenue-producing assets.

Through an outdoor trial conducted in the Kanagawa prefecture, SoftBank demonstrated that its NVIDIA-accelerated AI-RAN

solution has achieved carrier-grade 5G performance and was able to do so while using the network's excess capacity to run AI inference workloads concurrently.

Traditional telco networks are designed to handle peak loads and, on average, have used only one-third of that capacity. With the common computing capability provided by AI-RAN, it is expected that telcos now have the opportunity to monetize the remaining two-thirds capacity for AI inference services.

NVIDIA and SoftBank estimate that telco operators can earn roughly \$5 in AI inference revenue from every \$1 of capex it invests in new AI-RAN infrastructure.⁽¹⁾ Taking into account its opex and capex costs, SoftBank estimates it can achieve a return of up to 219% for every AI-RAN server it adds to its infrastructure.⁽²⁾

Real-World Inference Runs on AI-RAN

For the trial, SoftBank used NVIDIA AI Enterprise to build real-world AI inference applications, including autonomous vehicle remote support, robotics control and multimodal retrieval-automated generation at the edge. All inference workloads were able to run optimally on SoftBank's AI-RAN network.

SoftBank's fully software-defined 5G radio stack is optimized for NVIDIA's AI computing platform and includes L1 software enhanced by SoftBank based on [NVIDIA Aerial™ CUDA@-accelerated RAN libraries](#). SoftBank plans to incorporate [NVIDIA Aerial RAN Computer-1](#) systems, which it estimates can use 40% less power than traditional 5G network infrastructure,⁽³⁾ into its solution moving forward.

NVIDIA and SoftBank partners that contributed to the trial of SoftBank's AI-RAN solution include Fujitsu and Red Hat.

Matching Supply With Demand

Because an AI-RAN solution needs to spin compute up or down dynamically based on demand and supply without compromising carrier-grade performance in real time, SoftBank aims to build an ecosystem that connects the demand and supply of AI technology by using NVIDIA AI Enterprise serverless application programming interfaces and its in-house developed orchestrator. This enables SoftBank to dispatch external AI inferencing jobs to an AI-RAN server when computing resources are available to deliver localized, low-latency, secure inferencing services.

"Shifting from single-purpose to multi-purpose AI-RAN networks can mean 5x the revenue for every dollar of capex invested," said Ronnie Vasishtha, senior vice president of telecom at NVIDIA. "SoftBank's live field trial marks a huge step toward AI-RAN commercialization with the validation of technology feasibility, performance and economics."

"SoftBank's 'AITRAS' is the first AI-RAN solution developed through a five-year collaboration with NVIDIA. It integrates and coordinates AI and RAN workloads through the SoftBank-developed orchestrator, enhancing communication efficiency by running dense cells on a single NVIDIA-accelerated GPU server," said Ryuji Wakikawa, vice president and head of the Research Institute of Advanced Technology at SoftBank. "We are confident this AI-driven innovation, AITRAS, will pave the way for new business models in telecommunications, serving as a crucial factor in the transformation of mobile operators."

Learn more about NVIDIA solutions for [AI-RAN](#).

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the world leader in accelerated computing.

(1) Results do not guarantee actual revenue at time of implementation.

(2) Based on estimates by SoftBank.

(3) Based on estimates by NVIDIA and SoftBank. Results do not guarantee actual power reduction at time of implementation.

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, and performance of NVIDIA's products, services, and technologies, including NVIDIA Blackwell platform, NVIDIA Grace Blackwell platform, NVIDIA AI Aerial, NVIDIA AI Enterprise software, NVIDIA DGX B200 systems, NVIDIA DGX SuperPOD supercomputer, NVIDIA Quantum-2 InfiniBand networking, NVIDIA GB200 NVL72 multi-node, liquid-cooled, rack-scale systems, NVIDIA Grace CPUs, NVIDIA Aerial CUDA-accelerated RAN libraries, NVIDIA Aerial RAN Computer-1 systems, NVIDIA's full-stack AI, and NVIDIA Omniverse; our collaboration with SoftBank and the benefits and impact thereof; SoftBank using or adopting our products and technologies, the benefits and impact thereof, and the features, performance and availability of its offerings; with SoftBank's significant investment in NVIDIA's full-stack AI, Omniverse, and 5G AI-RAN platforms, Japan leaping into the AI industrial revolution to become a global leader, driving a new era of growth across the telecommunications, transportation, robotics and healthcare industries in ways that will greatly benefit humankind in the age of AI; countries and regions worldwide are accelerating the adoption of AI for social and economic growth, and society is undergoing significant transformation; through long collaboration with NVIDIA, SoftBank leading this transformation from the forefront; with its extremely powerful AI infrastructure and our new, distributed AI-RAN solution 'AITRAS' that reinvents 5G networks for AI, SoftBank accelerating innovation across the country and throughout the world; and AI-driven innovation, AITRAS, paving the way for new business models in telecommunications, serving as a crucial factor in the transformation of the mobile operators are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions;

our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

© 2024 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA, DGX, NVIDIA Aerial, NVIDIA DGX SuperPOD and NVIDIA Grace are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Kristin Bryson
Enterprise Data Center, AI/DL
+1-203-241-9190
kbryson@nvidia.com