NVIDIA Collaborates With SoftBank Corp. to Power SoftBank's Next-Gen Data Centers Using Grace Hopper Superchip for Generative AI and 5G/6G

Arm-Based Superchip and BlueField-3 DPU Power Revolutionary Architecture to Enable Generative AI-Driven Wireless Communications

COMPUTEX—NVIDIA and SoftBank Corp. today announced they are collaborating on a pioneering platform for generative AI and 5G/6G applications that is based on the NVIDIA GH200 Grace Hopper™ Superchip and which SoftBank plans to roll out at new, distributed AI data centers across Japan.

Paving the way for the rapid, worldwide deployment of generative AI applications and services, SoftBank will build data centers that can, in collaboration with NVIDIA, host generative AI and wireless applications on a multi-tenant common server platform, which reduces costs and is more energy efficient.

The platform will use the new NVIDIA MGX™ reference architecture with Arm Neoverse-based GH200 Superchips and is expected to improve performance, scalability and resource utilization of application workloads.

"As we enter an era where society coexists with AI, the demand for data processing and electricity requirements will rapidly increase. SoftBank will provide next-generation social infrastructure to support the super-digitalized society in Japan," said Junichi Miyakawa, president and CEO of SoftBank Corp. "Our collaboration with NVIDIA will help our infrastructure achieve a significantly higher performance with the utilization of AI, including optimization of the RAN. We expect it can also help us reduce energy consumption and create a network of interconnected data centers that can be used to share resources and host a range of generative AI applications."

"Demand for accelerated computing and generative AI is driving a fundamental change in the architecture of data centers," said Jensen Huang, founder and CEO of NVIDIA. "NVIDIA Grace Hopper is a revolutionary computing platform designed to process and scale-out generative AI services. Like with other visionary initiatives in their past, SoftBank is leading the world to create a telecom network built to host generative AI services."

The new data centers will be more evenly distributed across its footprint than those used in the past, and handle both AI and 5G workloads. This will allow them to better operate at peak capacity with low latency and at substantially lower overall energy costs.

SoftBank is exploring creating 5G applications for autonomous driving, AI factories, augmented and virtual reality, computer vision and digital twins.

Virtual RAN With Record-Breaking Throughput
NVIDIA Grace Hopper and NVIDIA BlueField®-3 data processing units will accelerate the software-defined 5G vRAN, as well as generative AI applications, without bespoke hardware accelerators or specialized 5G CPUs. Additionally, the NVIDIA Spectrum Ethernet switch with BlueField-3 will deliver a highly precise timing protocol for 5G.

The solution achieves breakthrough 5G speed on an NVIDIA-accelerated 1U MGX-based server design, with industry-high throughput of 36Gbps downlink capacity, based on publicly available data on 5G accelerators. Operators have struggled to deliver such high downlink capacity using industry-standard servers.

New Reference Architecture
NVIDIA MGX is a modular reference architecture that enables system manufacturers and hyperscale customers to quickly and cost-effectively build over a hundred different server variations to suit a wide range of AI, HPC and NVIDIA Omniverse™ applications.

By incorporating NVIDIA Aerial™ software for high-performance, software-defined, cloud-native 5G networks, these 5G base stations will allow operators to dynamically allocate compute resources, and achieve 2.5x power efficiency over competing products.

"The future of generative AI requires high-performance, energy-efficient compute like that of the Arm Neoverse-based Grace Hopper Superchip from NVIDIA," said Rene Haas, CEO of Arm. "Combined with NVIDIA BlueField DPUs, Grace Hopper enables the new SoftBank 5G data centers to run the most demanding compute- and memory-intensive applications and bring exponential efficiency gains to software-defined 5G and AI on Arm."

About SoftBank Corp.

Guided by the SoftBank Group's corporate philosophy, "Information Revolution – Happiness for everyone," SoftBank Corp. (TOKYO: 9434) provides telecommunications services and combines them with advanced technologies to develop and operate new businesses in Japan and globally. By fully harnessing the power of 5G, AI, IoT, Digital Twin and other key technologies, SoftBank Corp. aims to realize the "Implementation of Digitalization into Society." To learn more, please visit https://www.softbank.jp/en/.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling the creation of the industrial metaverse. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

Certain statements in this press release including, but not limited to, statements as to: the benefits, performance, impact and availability of our products and technologies, including NVIDIA Grace Hopper, BlueField-3, NVIDIA MGX, NVIDIA Spectrum and NVIDIA Aerial software; NVIDIA's collaboration with SoftBank, including the benefits and impact thereof; the demand for data processing and electricity requirements rapidly increasing; demand for accelerated computing and generative AI driving a fundamental change in the architecture of data centers; and the future of generative AI requiring high-performance, energy-efficient compute 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.

Cliff Edwards
NVIDIA Corporation
+1-415-699-2755
cliffe@nvidia.com