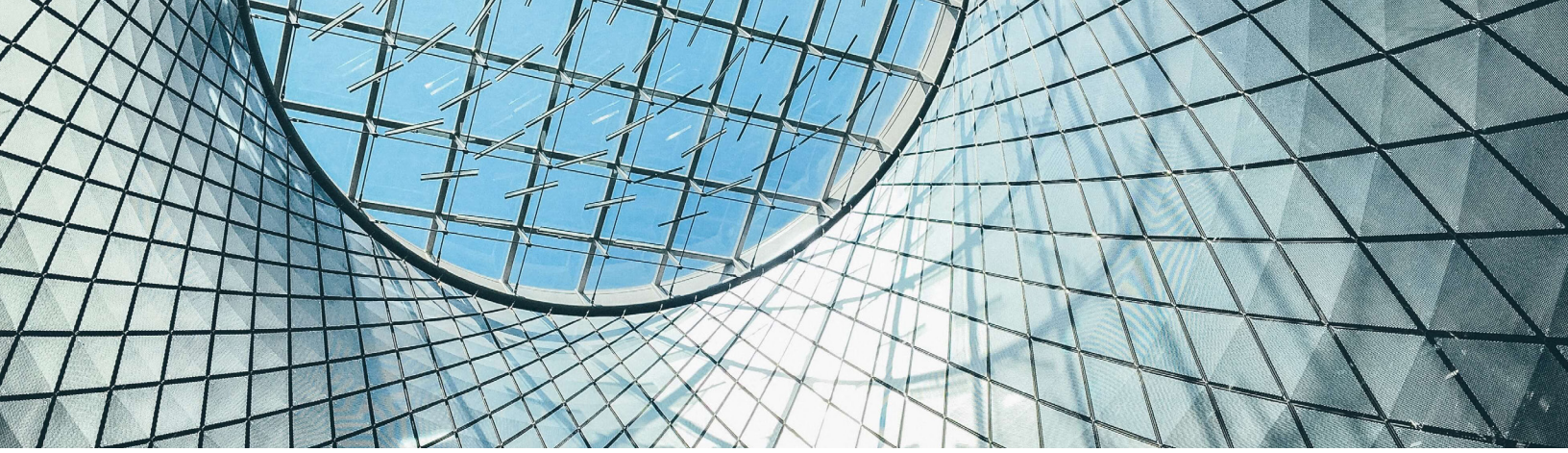

451 Research Market Insight Report Reprint

Vertiv unveils its reference designs for NVIDIA's GB200 NVL72 platform

November 4, 2024

by Perkins Liu

The company's reference design for the NVIDIA GB200 NVL72 is regarded as a major advancement in supporting AI-driven datacenters. This collaboration addresses the demand for high-density liquid-cooled solutions, optimizing deployment speed, scalability and energy efficiency, and meeting AI workloads' cooling and power demands.

S&P Global
Market Intelligence

Introduction

Vertiv, a global provider of power and cooling solutions, unveiled its reference design for the NVIDIA GB200 NVL72 platform in October. Co-developed with NVIDIA Corp., this liquid-cooled architecture is tailored for AI-driven datacenters by offering a comprehensive blueprint for critical power and cooling systems, supporting high-density workloads up to 132 kW per rack, and scalable to 7 MW. The design enables rapid deployment, energy-efficient cooling and improved power management, addressing the rising demand for AI infrastructure. This collaboration aims to transform traditional datacenters into high-efficiency AI factories, accelerating the adoption of AI technologies across industries.

THE TAKE

Vertiv's release of its reference design for the NVIDIA GB200 NVL72 is regarded as a major advance in supporting AI-driven datacenters. This collaboration addresses the demand for high-density liquid-cooled solutions, optimizing deployment speed, scalability and energy efficiency, and meeting the cooling and power demands of AI workloads such as generative AI and machine learning. It is a strategic move for Vertiv, enhancing its position as a leader in AI infrastructure while providing future-ready designs that enable adaptability, operational flexibility and sustainability across industries. For the industry, it is a sign that upstream and downstream partners in the same ecosystem need to work more closely to address the challenges that AI has brought.

Details

The rapid growth of AI deployment has increased the demand for high-power, high-density infrastructure, necessitating the use of liquid cooling. After an almost linear climb of rack density from 4 kW to 12 kW from 2010 to 2022, rack density has risen dramatically to over 100 kW thanks to the emerging AI deployment in the past two years. Datacenter infrastructure is under pressure to respond to this unprecedented increase in density. NVIDIA, a dominant player in the GPU market, faces significant challenges with its next-generation GB200 NVL72 architecture, calling for collaboration with infrastructure vendors to meet these needs.

The NVIDIA GB200 NVL72 architecture is a rack-based platform designed to support AI-driven datacenters, featuring 72 of NVIDIA's Blackwell GPUs and 36 Grace CPUs. In a 48-U high and 19-inch or 21-inch wide OCP rack, there are 18 compute trays, nine NVswitch trays and six power trays. Each tray is 1-U high. Each compute tray is considered as one compute node, which includes four GPUs and two CPUs, for power consumption of 5.4 kW. The entire rack has power consumption of 132 kW. The system integrates NVIDIA's NVLink technology, allowing for seamless connectivity across its components with bandwidth of up to 130 TBps, enabling dynamic workload management. This system is engineered to deliver impressive computational power, boasting 720 petaflops for training and 1.4 exaflops for inference tasks. To efficiently manage high-density workloads, the GB200 NVL72 employs a fully liquid-cooled design, utilizing coolant temperatures ranging from 45°C (113°F) for inlet to 65°C (149°F) for outlet.

To embrace the challenge AI brought to infrastructure, Vertiv launched its 360AI solution in April. This new suite of integrated power and cooling solutions supports the growing power and cooling demands associated with AI and high-performance computing. It features a full portfolio of power, cooling, enclosures and structures, digitized management and life-cycle services. It allows for flexibility and customization, retrofitting existing datacenters or buildouts of hyperscale green fields.

The reference design for the NVIDIA GB200 NVL72 is a key component of Vertiv's 360AI portfolio. With capabilities to support up to 132 kW per rack and scalable up to 7 MW in 1.1 MW increments, consisting of six 1.1 MW SuperPODs, this architecture addresses the growing need for efficient power and cooling solutions in environments where AI applications are rapidly evolving. It integrates with the NVIDIA Blackwell platform to streamline the deployment of AI workloads in datacenters, both new and retrofitted. It aims to minimize risks and promotes site consistency by aligning AI clusters with datacenter capacity blocks, thus reducing stranded power. The system employs a hybrid cooling approach that combines liquid and air cooling to manage high-density heat removal effectively. Additionally, it offers guidance for optional systems inspired by the Open Compute Project, such as DC power shelves.

CONTACTS

Americas: +1 800 447 2273
Japan: +81 3 6262 1887
Asia-Pacific: +60 4 291 3600
Europe, Middle East, Africa: +44 (0) 134 432 8300

www.spglobal.com/marketintelligence
www.spglobal.com/en/enterprise/about/contact-us.html