



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

This report, licensed to Vertiv, developed and as provided by S&P Global Market Intelligence (S&P), was published as part of S&P's syndicated market insight subscription service. It shall be owned in its entirety by S&P. This report is solely intended for use by the recipient and may not be reproduced or re-posted, in whole or in part, by the recipient without express permission from S&P.

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

Copyright © 2024 by S&P Global Market Intelligence, a division of S&P Global Inc. All rights reserved.

These materials have been prepared solely for information purposes based upon information generally available to the public and from sources believed to be reliable. No content (including index data, ratings, credit-related analyses and data, research, model, software or other application or output therefrom) or any part thereof (Content) may be modified, reverse engineered, reproduced or distributed in any form by any means, or stored in a database or retrieval system, without the prior written permission of S&P Global Market Intelligence or its affiliates (collectively S&P Global). The Content shall not be used for any unlawful or unauthorized purposes. S&P Global and any third-party providers (collectively S&P Global Parties) do not guarantee the accuracy, completeness, timeliness or availability of the Content. S&P Global Parties are not responsible for any errors or omissions, regardless of the cause, for the results obtained from the use of the Content. THE CONTENT IS PROVIDED ON "AS IS" BASIS. S&P GLOBAL PARTIES DISCLAIM ANY AND ALL EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OR USE, FREEDOM FROM BUGS, SOFTWARE ERRORS OR DEFECTS, THAT THE CONTENT'S FUNCTIONING WILL BE UNINTERRUPTED OR THAT THE CONTENT WILL OPERATE WITH ANY SOFTWARE OR HARDWARE CONFIGURATION. In no event shall S&P Global Parties be liable to any party for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including, without limitation, lost income or lost profits and opportunity costs or losses caused by negligence) in connection with any use of the Content even if advised of the possibility of such damages.

S&P Global Market Intelligence's opinions, quotes and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global Market Intelligence may provide index data. Direct investment in an index is not possible. Exposure to an asset class represented by an index is available through investable instruments based on that index. S&P Global Market Intelligence assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global keeps certain activities of its divisions separate from each other to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain nonpublic information received in connection with each analytical process.

S&P Global may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P Global reserves the right to disseminate its opinions and analyses. S&P Global's public ratings and analyses are made available on its websites, www.standardandpoors.com (free of charge) and www.ratingsdirect.com (subscription), and may be distributed through other means, including via S&P Global publications and third-party redistributors. Additional information about our ratings fees is available at www.standardandpoors.com/usratingsfees.