hyve solutions

orion

AI Rack-Level Solution Featuring the NVIDIA® HGX™ Platform

hyve solutions®

NVIDIA®

CONTENTS

EXECUTIVE SUMMARY 2
MARKET TRENDS 3
PRODUCT HIGHLIGHTS 4
INSIDE THE HYVE NVIDIA HGX PLATFORM 5
HYVE ADVANTAGES 7

EXECUTIVE SUMMARY

Hyve Orion featuring NVIDIA HGX Platform offers a cutting-edge solution for hyperscale datacenters looking to deploy and scale AI-centric infrastructure. Built on NVIDIA's HGX reference architecture, Hyve Orion provides a rack-level solution optimized for current Hopper™ and future NVIDIA Blackwell architectures. Available in two foundational 10U SKUs — a JBOG (Just a Bunch of GPUs) and a complete, monolithic server — the solution delivers exceptional flexibility and modularity to meet evolving AI computational demands.

As one of NVIDIA's Elite design partners, Hyve leverages its extensive background in hyperscaler infrastructure to enhance the HGX design. The platform's 6U graphics module ensures compatibility with air-cooled NVIDIA H100 and H200, and B100 and B200 GPUs, as well as an optional liquid-cooled configuration, offering a smooth transition path for datacenters if and when their needs expand. This forward-looking approach, combined with Hyve's comprehensive Unified Global Services (HUGS) program, provides hyperscalers with a robust, future-proof solution that minimizes deployment complexities and optimizes total cost of ownership.

Hyve Orion architecture stands out through its emphasis on validation at both rack and infrastructure levels. This "rack-ready" approach ensures that what's tested is exactly what's deployed, capturing real-world variables often missed in single-box testing. For hyperscale operators facing mounting AI workloads and power densities, Hyve's platform offers performance, scalability, and support services, positioning datacenters to meet both current and future AI infrastructure challenges with confidence.

MARKET TRENDS

Artificial intelligence (AI) and machine learning (ML) hardly need an introduction these days. From in-app assistants to multimodal industrial manufacturing systems, the broad umbrella of AI is this decade's "it" technology. The computational resources needed to develop, train, and host AI's underlying models — both proprietary and open source — continue to escalate at a breakneck pace across practically every metric.

JLL Research anticipates hyperscaler data center rack-level power density rocketing from today's 36 kW to nearly 50 kW in 2027, largely driven by AI demands. The Wall Street Journal reports that "spending in the global AI infrastructure market — including data centers, as well as networks and other hardware that supports the use of AI applications — is expected to reach $422.55 billion by 2029," reflecting a 44% compounded annual growth rate (CAGR). Gartner recently revealed that its survey of 459 technology service providers "found that 83% of respondents reported they either have already deployed or are currently piloting generative AI (GenAI) within their organizations."

Hyperscale datacenters face considerable challenges in staying abreast of these trends. New facilities must embrace next-generation power and cooling demands, while thousands of brownfield sites remain in need of more flexible, modular approaches than what was available only a few years ago. Effective, efficient rack-scale solutions for foundational AI applications will provide maximal value by taking a more modular approach to AI-centric infrastructure. A pre-validated, "rack-ready" modular approach reduces the time and complexity typical in hyperscale deployments, making future-proof AI infrastructure deployment a relative snap.

PRODUCT HIGHLIGHTS

Orion is a rack-level solution based on NVIDIA's HGX AI supercomputing reference architecture for NVIDIA Hopper and NVIDIA Blackwell - architecture. Orion is available as two core SKUs:

1. Hyve Orion JBOG : A 10U JBOG ("just a bunch of GPUs"). The unit suits organizations with legacy code tied to specific head node needs, such as a given CPU vendor or model family.
2. Hyve Orion Monolithic: A 10U monolithic server. This option pairs the JBOG (which occupies 6U within the enclosure) with a 2U server (built to NVIDIA's exact reference specifications) and 2U of power. Datacenters can deploy up to three Hyve HGX servers per rack.

With its extensive background in hyperscaler infrastructure design and deployment, Hyve works closely with NVIDIA to produce NVIDA HGX platform products. This partnership grants Hyve access to NVIDIA's HGX development resources and support. Hyve then produces HGX Platform hardware to NVIDIA's exacting specifications, but does so with some key enhancements for modularity, future-proofing, and total solution value via Hyve's extensive programs and services for hyperscaler datacenters.

Importantly, most available NVIDIA HGX platforms were made for NVIDIA Hopper architecture, which operated well in 4U or 5U enclosures. That condition no longer applies with air-cooled NVIDIA B100 and B200 GPUs, which require 6U to support air-cooling heatsinks. Fortunately, Hyve's 6U graphics design offers a seamless bridge from the advantages of today's NVIDIA H100 and H200 GPUs to tomorrow's NVIDIA Blackwell architecture.

This modularity and scalability, coupled with NVIDIA partner-class quality and Hyve's own roster of services, yields industry-leading flexibility and exceptionally attractive TCO for hyperscaler datacenters.

INSIDE HYVE ORION SOLUTION

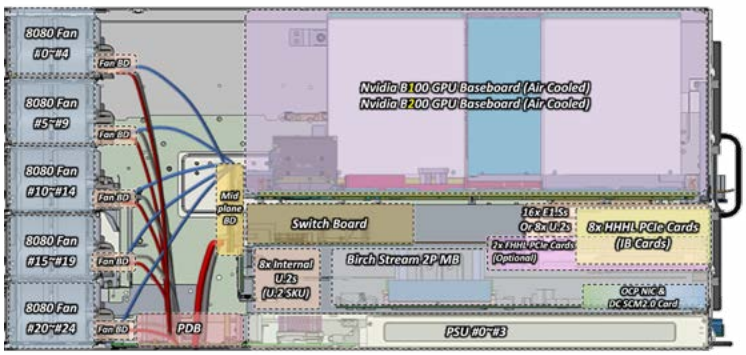
Conventional NVIDIA HGX designs typically use a 4U or 5U form factor, leaving them unable to run NVIDIA Blackwell architecture without liquid cooling, which many datacenters are not yet prepared to deploy. By building its GPU module in a 6U design, Hyve's Orion arrives ready to roll with air-cooled NVIDIA Hopper or NVIDIA Blackwell architecture, as users' needs and budgets may dictate. This forward compatibility minimizes complications associated with validating new SKUs, inventorying spares, retraining, and related issues.

Hyve Orion JBOG Featuring NVIDIA HGX Platform

Based on NVIDIA's Umbriel baseboard architecture, Hyve Orion JBOG for NVIDIA HGX platform is a rack-ready, GPU powerhouse built with the following highlights:

	Orion J200-B	Orion J100-B
GPU (max)	NVIDIA B200	NVIDIA B100
TDP (max)	1000W	700W

	Orion J200-B & Orion J100-B
Form Factor	10U, 19"
Heatsink	Air, 6U cooling solution
Socket	8S
Network Interface	OCP 3.0 NIC PCIe Gen5 x16
GPU InfiniBand	8 HHHL PCIe Gen5 x16 ConnectX-7 400 Gbps
PCIe Switches	2x PCIe Gen5 x16 switches, 1 per socket
Power Supplies	48~54Vout@110 ~ 220VAC Vin, N+N redundancy supporting
Fans	8080 x25 hot-swap



Hyve also offers a liquid-cooling based variant model.

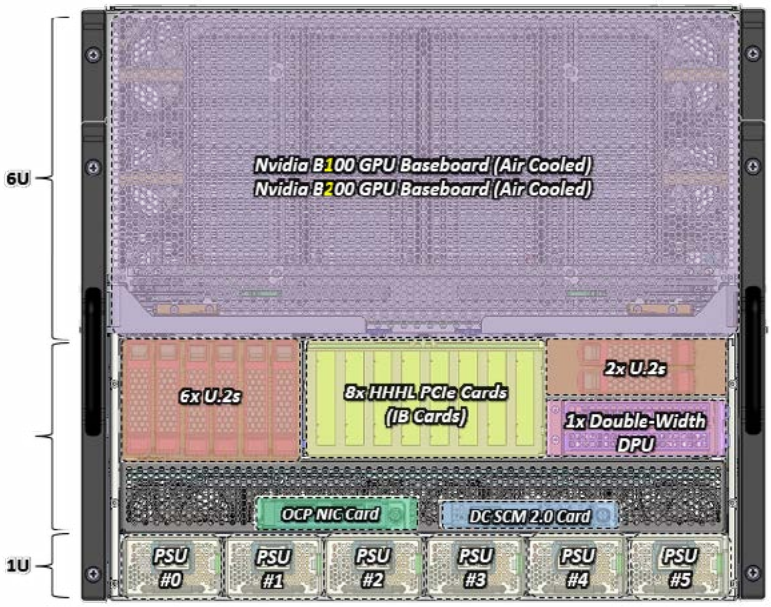
Hyve Orion JBOG is fully compatible across NVIDIA H100, H200, B100 and B200 GPUs and is agnostic with any compatible head server (e.g., Intel®, AMD, ARM) in the existing infrastructure.

Hyve Orion Monolithic Featuring NVIDIA HGX Platform

Hyve Orion monolithic NVIDIA HGX server builds on the JBOG foundation. The server stack provides a turnkey solution that quickly and easily drops into existing infrastructure, either as a 10U enclosure for existing rack resources or as a multitude of rack-mounted servers in a supercomputing cluster.

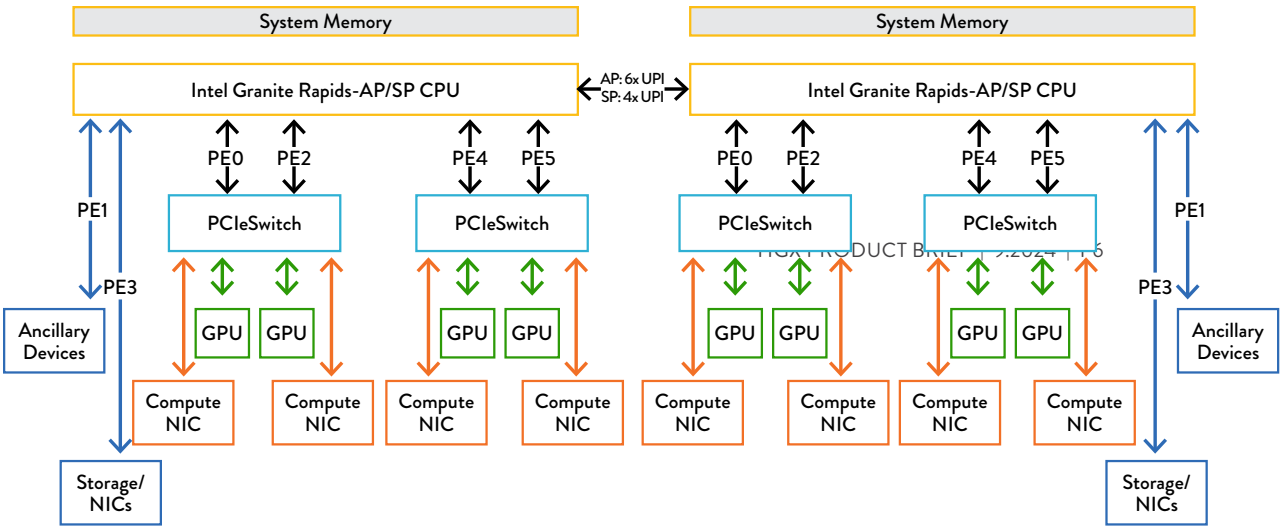
Beyond the JBOG's specifications, Hyve's Orion Monolithic server adds:

	Orion M200-H
Form Factor	10U, 19"
CPU	Intel Granite Rapids-SP
Socket	2S
TDP (max)	350W
Memory	DDR5, 8-ch, 5200/6400
Mem. Capacity	8TB
Storage	M.2 NVMe boot, PCIe Gen5 16x (E1.S and U.2 options)
NIC	OCP 3.0 NIC PCIe Gen5 x16
GPU InfiniBand	8 HHHL PCIe Gen5 x16 ConnectX-7 400 Gbps
PCIe Switches	2x PCIe Gen5 x16 switches, 1 per socket
Power Supplies	48~54Vout@110 ~ 220VAC Vin, N+N Support
Fans	8080 x25 hot-swap



With Hyve Orion solution validation services, datacenters can essentially "plug and play" these monolithic servers into their current infrastructure. In doing so, they achieve proven, stable hardware for demanding AI applications, such as LLM training and hosting.

Hyve Orion Monolithic Architecture



HYVE ADVANTAGES

We mentioned earlier that, as an Elite partner in the NVIDIA partner network, Hyve had earned access to the design collateral and resources needed to bring these visionary reference designs to hyperscale datacenters. This recognition grew from Hyve's many years of successful growth and innovation with modular hyperscaler datacenter hardware. Moreover, this growth stemmed from a long, ever-broadening stream of deep customer relationships, many of which involved significant customization and engineering.

The centerpiece of Hyve's post-sale offerings is its Hyve Unified Global Services (HUGS) program. Designed for tier-one clients and now available to all hyperscale clients globally, HUGS delivers a host of options that mesh perfectly with AI-centric operations, including:

- Supply Chain Services. From spare provisioning to fulfillment to recycling, Hyve leverages hundreds of supplier and logistics relationships to ensure that product flows remain timely and reliable.
- RMA. Hyve can adapt to your policy needs with flexible warranty periods and return location accommodations, including advanced swaps, on-site repair, and high-count failure analysis.
- On-site deployment across 150 countries worldwide, from single racks to sprawling clusters.
- Integration. Hyve got its start building rack infrastructure, and the company has only expanded on this base. HUGS integration means building, testing, and validating a full rack solution with the customer's own software configuration at Hyve, then deploying that exact solution into the customer's existing infrastructure.
- Global Logistics. Hyve possesses significant warehouse presence in North America, Asia, and Europe as well as importer of record (IoR) capabilities in over 80 countries, enabling Hyve to drop ship into every corner of the world.
- Refurbishing. Every server deserves a second chance, and HUGS refurbishing can bring failed or decommission hardware back into "like new" condition for a fraction of buying off the shelf.

For the type of datacenters deploying NVIDIA HGX platforms, Hyve's validation services should be an exceptional draw. Validation sits at the core of Hyve's "rack-ready" value proposition. This is hardware/software validation at rack and infrastructure levels, not simply a server on a test bench.

Far too many providers take the approach of only testing a single box. This is like assessing a central AC unit disconnected from its target building's ducting and power. Doing so misses the impact of ambient conditions and interactions. Hyve rack- and infrastructure-level validation captures network management, power, and a myriad of other real-world variables. And again, critically, what Hyve tests is exactly what arrives and deploys on site.

Orion begins with NVIDIA's HGX reference design. Hyve's design capabilities then creates modular and scalable reference designs targeted at datacenters needing to span from today's NVIDIA Hopper architecture into the coming NVIDIA Blackwell architecture. Hyve services complete this AI-centric stack, allowing hyperscalers to deploy with shorter time to market, redoubled quality control, best-in-class performance, and confidence in infrastructure built for next-generation opportunities.control, best-in-class performance, and confidence in infrastructure built for next-generation opportunities.

CONTACT US

855.869.6873 | sales@hyvesolutions.com | 44201 Nobel Drive, Fremont, CA 94538 | hyvesolutions.com