AgilityFlexAI B200 HGX POD

Integrating NVIDIA HGX B200, DDC Hybrid-Cooled Cabinets, and IBM SS6000 Storage

Challenges

- Multi-vendor complexity
- Thermal & power limits
- Data bottlenecks

Solution

- Turnkey AgilityFlexAI B200 HGX POD
- Single-source accountability
- Expertly balanced architecture

Benefits

- Day-one optimal performance with no integration guesswork or tuning cycles
- Lower TCO & energy use
- Modular design lets you grow compute or storage independently, protecting investment as workloads evolve

AI Factory in a Box

Stand up a production-ready AI Factory in a day with Applied Data Systems' AgilityFlexAI B200 HGX POD, a turnkey high-performance computing platform for AI inference and HPC. It combines state-of-the-art GPU compute with advanced cooling and storage in a pre-engineered solution, delivering maximum performance and efficiency "done right the first time." Powered by the revolutionary NVIDIA HGX B200 Blackwell architecture, integrated with advanced DDC hybrid-cooled cabinets, and supported by IBM Storage Scale System 6000, this solution represents the perfect convergence of cutting-edge technology, expert engineering, and single-source accountability.

Our Key Value Propositions

- **Done Right the First Time:** Expertly architected and fully integrated systems that eliminate deployment risks.
- **Single Point of Accountability**: Single source for all components and support.
- **Best-of-Breed Manufacturers**: Partnerships with industry leaders ensure optimal performance and reliability.
- **Expertly Architected**: Over 50 years of combined HPC experience designing solutions for the world's most demanding environments.

Product Brief

Scaling AI Factories the Smart Way

More Available Power for More GPU Computing

Electricity has quickly become the most challenging limiting factor in GPU deployment. In typical enterprise data centers, cooling alone can account for as much as 40% of total energy consumption. DDC Hybrid-Cooled Cabinets reduce cooling energy by 30–40%, which translates to roughly 15% lower total power, freeing watts for compute so you can deploy more GPUs within a constrained power envelope.

Balanced Architecture for Optimal Performance

We right-size compute, storage, and networking to ensure GPUs stay fed without stalling.

- Balanced throughput: CPU, networking, and storage are architected to match HGX B200 performance.

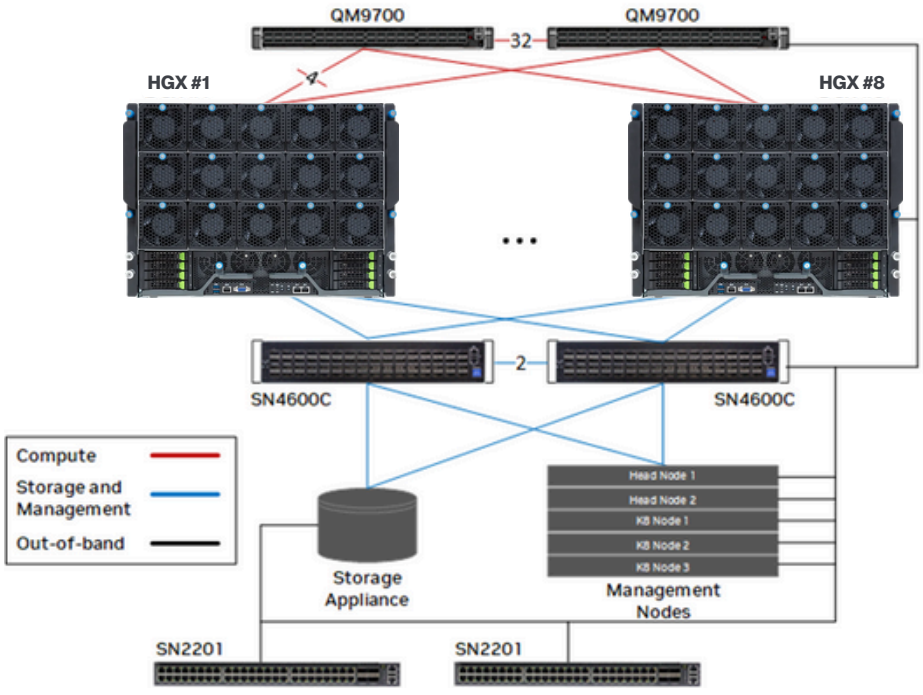
NVIDIA AI Enterprise Pre-installed, Licensed, and Validated

- Your cluster arrives fully configured with NVIDIA's enterprise AI stack.
- Includes Base Command Manager (BCM) for cluster management.
- RunAI orchestration for multi-tenant GPU scheduling, quotas, and fair-share policies for teams and projects.
- Ready-to-run templates for training, finetuning, and inference; monitoring and logging wired in.
- Comes with 3, 4, or 5-year support from NVIDIA, so customers typically see performance improvements over time as they continue to optimize the various software packages.

Networking Built for Scale

We design GPU fabrics with high-performance Spectrum X Ethernet (100/200/400/800GbE), using proven leaf-spine/Clos topologies with RoCEv2, QoS/ECN, and congestion management. InfiniBand is available where workflows require it.

- We size and validate for high-volume North/South data ingress/egress (data lakes, checkpoints, user access) and ultra-low-latency East/West GPU-to-GPU traffic for training and inference.



AgilityFlexAI B200 HGX POD with up to eight systems with NDR400

Applied DataSystems

For more information from Applied Data Systems:
844.371.4949 | info@applieddatasystems.com | www.applieddatasystems.com

Technical Highlights

Accelerated GPU Compute (NVIDIA HGX B200)

At the heart of the solution is the NVIDIA HGX B200 platform, an 8-GPU baseboard built on NVIDIA's latest Blackwell architecture. Each HGX B200 node delivers unprecedented compute performance for AI and HPC: up to 15× higher real-time AI inference throughput than the previous-generation (HGX H100) system, while using 12X less energy. This massive leap, enabled by Blackwell GPUs' advanced Transformer Engine and 5th-Gen NVLink interconnect, means the POD can tackle multi-trillion-parameter models and complex simulations in real time.

Each 8-GPU server is equipped with dual latest-gen CPUs (Intel Xeon or AMD EPYC) and supports a 1:1 GPU-to-NIC networking ratio – for example, eight 400 Gb/s network adapters (NVIDIA BlueField-3 DPUs or ConnectX-7 InfiniBand) can be tied to the 8 GPUs.This design eliminates network bottlenecks by saturating each GPU with dedicated, high-bandwidth I/O, making it ideal for distributed deep learning or MPI-driven HPC tasks. The HGX B200 nodes can be deployed with either air cooling or liquid cooling. In AgilityFlexAI's configuration, they are integrated into a purpose-built cooling system that allows for up to 96 GPUs per rack in a dense configuration.

Thermal Efficiency with DDC Hybrid-Cooled Cabinets

To reliably operate such high-power processors at scale, AgilityFlexAI employs direct hybrid-cooled cabinet technology from DDC Solutions for thermal management. Modern AI GPUs can draw >700 W each, which quickly multiplies to hundreds of kW per rack – far beyond what traditional air cooling can handle. The DDC S-Series cabinet in this solution is a sealed 52U enclosure that combines forced-air cooling with optional liquid-to-chip cooling for extreme efficiency. It supports up to 100 kW of air cooling and liquid-to-chip-ready for any density.

In practical terms, this means the cabinet can remove heat from dozens of GPUs and high-density components without thermal throttling, even under 100% load. The hybrid-cooled design dramatically improves heat transfer efficiency (hybrid can carry heat 3,500x more efficiently than air), reducing data center HVAC strain. It also enables greater density – packing more compute per rack since cooling is no longer the limiting factor. The DDC cabinet is effectively a self-contained "mini data center," as it includes integrated coolant distribution, smart dynamic fans, DCIM monitoring, and even onboard fire suppression and security features for safe operation. No special site preparation is required to deploy it (no raised floor or external chillers are needed), making it a drop-in solution for upgrading existing facilities.

High-Performance Scalable Storage (IBM Storage Scale System 6000)

Data is the fuel for AI, and the AgilityFlexAI POD integrates a best-in-class storage infrastructure to match the compute capability. The solution features the IBM Storage Scale System 6000 (SS6000) – a high-end, all-flash parallel file system appliance engineered for AI/HPC throughput. Each SS6000 unit provides up to 1.44 PB of usable flash storage in 4U, with an astonishing 330 GB/s of sequential bandwidth and 13 million IOPS of random performance. For the AgilityFlexAI POD, storage will never be the bottleneck, even for enormous datasets or model checkpoints. The IBM system supports a broad range of data access protocols – NFS, SMB, S3 object, HDFS, and, importantly, NVIDIA GPUDirect Storage.

High-Speed Network Fabric

The AgilityFlexAI POD's design features a high-bandwidth, low-latency cluster interconnect, enabling all GPUs and storage to work in concert efficiently. It supports NVIDIA Quantum-2 InfiniBand (NDR: 400 Gb/s) and NVIDIA Spectrum-X Ethernet (up to 400 GbE) switching technologies, enabling the connection of GPU servers to storage and to each other. In practice, each 8-GPU node can leverage an aggregate 3.2 Tb/s of network bandwidth (8 × 400 Gb connections) to the fabric. This ensures near-linear scaling when clustering multiple GPU nodes – large AI models can be distributed across servers with minimal communication overhead, and HPC workloads using MPI benefit from sub-microsecond latencies.

Applied Data Systems Differentiators

AgilityFlexAI B200 HGX POD provides organizations a comprehensive, worry-free path to deploy AI and HPC infrastructure that is both high-performance and easy to manage. Below is a summary of key benefits and competitive differentiators:

- **Turnkey Deployment with Fast Time-to-Production:** The solution arrives as an integrated and fully configured system – rack, stacked, and thoroughly tested. This turnkey approach enables you to set up an AI/HPC cluster in a fraction of the time it would take to design and build one in-house. All components are compatible and optimized out of the box, allowing your team to focus on workloads rather than debugging infrastructure.

- **Single-Vendor Support & Accountability:** Unlike piecemeal approaches, AgilityFlexAI provides you with a single vendor to hold accountable for the entire solution's success. This "one throat to choke" model ensures streamlined troubleshooting and support. There's no confusion over whether an issue is with the servers, storage, or network vendor – Applied Data Systems supports everything as a unified platform, simplifying maintenance and SLA management.

- **Ultimate Performance with No Bottlenecks:** Every layer of this design is best-of-breed and tuned for maximum performance. NVIDIA's HGX B200 GPUs deliver cutting-edge compute acceleration (with substantial performance-per-watt gains). IBM's Storage Scale system sustains extreme I/O throughput to keep those accelerators busy, and the HDR/NDR networking ties it all together with minimal latency. The result is an infrastructure that can handle GPU-intensive inference, training, and HPC simulations at scale without one subsystem throttling the others. This balanced, high-throughput design is a key differentiator versus competing solutions that might excel in one dimension (e.g., compute) but falter in others (like I/O or cooling).

- **Flexibility and Independent Scalability:** AgilityFlexAI is built with a modular architecture, allowing independent scaling of compute and storage. Need more GPU horsepower for AI inference? Additional HGX B200 nodes can be added to the POD without requiring an overhaul of the storage. Conversely, if your data volumes grow, you can expand the IBM Storage Scale cluster independently to add capacity/throughput – all while existing applications continue running. This flexibility protects your investment and adapts to changing workload demands, in contrast to rigid "appliance" solutions that force fixed ratios of compute to storage. Furthermore, the system's cooling and power design has ample headroom to accommodate future expansion or next-gen processors, providing forward compatibility.

- **Efficiency and Cost Savings:** By leveraging advanced cooling and efficient GPUs, the AgilityFlexAI POD also yields operational cost advantages. The hybrid-cooled cabinet significantly reduces cooling power usage compared to traditional air cooling, contributing to a lower PUE (Power Usage Effectiveness) for the deployment. NVIDIA's latest GPUs not only deliver far higher performance but also do so at a fraction of the energy per compute task, resulting in lower electricity and HVAC costs for equivalent workload throughput. These efficiency gains can translate into substantial TCO reduction over the system's life. Additionally, the single-vendor integrated approach can lower administrative overhead and avoid costly downtime that might occur in a DIY multi-vendor environment, further improving the cost-benefit equation.

Applied Data Systems' AgilityFlexAI B200 HGX POD provides a robust, high-performance AI/HPC infrastructure with the convenience of a fully managed solution. It marries the world's leading hardware components (NVIDIA for compute, IBM for storage) with expert engineering and support, delivering a platform that is immediately ready to propel your AI initiatives and scientific workloads. AgilityFlexAI offers the best of both worlds – bleeding-edge capability and "done-right-the-first-time" reliability – all with one single point of accountability. By choosing this expertly designed GPU POD, organizations can accelerate innovation in AI and HPC while minimizing risk, knowing that their infrastructure is optimized, scalable, and supported from end to end.

Applied DataSystems

For more information from Applied Data Systems:
844.371.4949 | info@applieddatasystems.com | www.applieddatasystems.com