DDN A³I® SOLUTIONS WITH ATOS BULLSEQUANA XH3000 SUPERCOMPUTER

Fully-integrated and optimized infrastructure solutions for accelerated at-scale AI, Analytics and HPC

Executive Summary

DDN A^{3}I Solutions are proven at-scale to deliver optimal data performance for Artificial Intelligence (AI), Data Analytics and High-Performance Computing (HPC) applications running on GPUs in an XH3000 system. This document describes fully validated reference architectures and scalable configurations. The solutions integrate DDN AI400X2 appliances with Atos BullSequana XH3000 systems and recommended Atos DLC network switches.

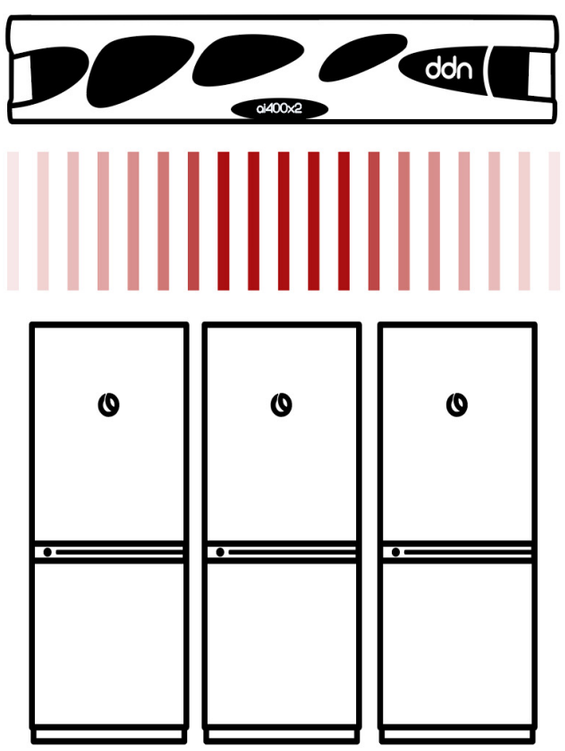
1. DDN A³I END-TO-END ENABLEMENT FOR ATOS SYSTEMS

DDN A³I solutions (Accelerated, Any-Scale AI) are architected to achieve the most from at-scale AI, Data Analytics and HPC applications running on BullSequana XH3000 Supercomputer. They provide predictable performance, capacity, and capability through a tight integration between DDN and Atos systems. Every layer of hardware and software engaged in delivering and storing data is optimized for fast, responsive, and reliable access.

DDN A³I solutions are designed, developed, and optimized in close collaboration with Atos. The deep integration of DDN AI appliances with BullSequana XH3000 ensures a reliable experience. DDN A³I solutions are highly configuration for flexible deployment in a wide range of environments and scale seamlessly in capacity and capability to match evolving workload needs. DDN A³I solutions are deployed globally and at all scale, from a single processor system all the way to the largest AI infrastructures in operation today.

DDN brings the same advanced technologies used to power the world's largest supercomputers in a fully-integrated package for BullSequana XH3000 that's easy to deploy and manage. DDN A³I solutions are proven to maximum benefits for at-scale AI, Analytics and HPC workloads on BullSequana XH3000 supercomputer.

This section describes the advanced features of DDN A³I Solutions for BullSequana XH3000.



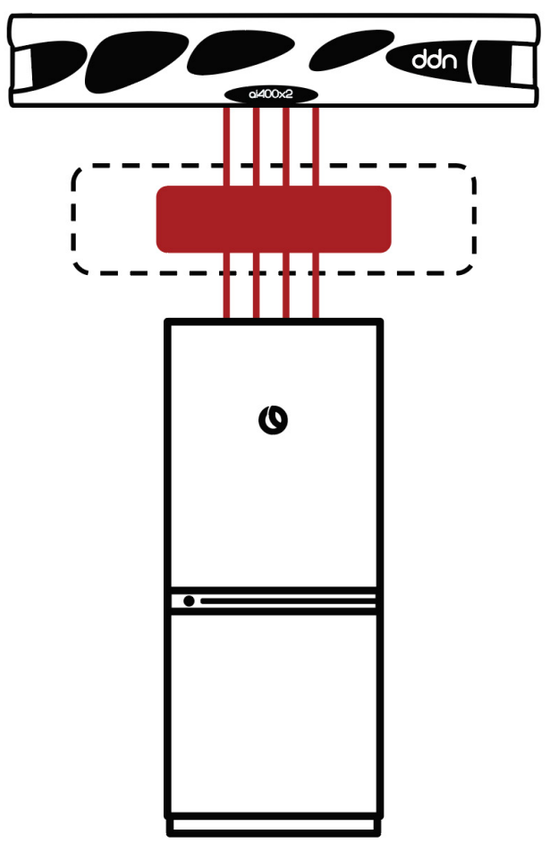
DDN A³I SHARED PARALLEL ARCHITECTURE

The DDN A³I shared parallel architecture and client protocol ensures high levels of performance, scalability, security, and reliability for DGX systems. Multiple parallel data paths extend from the drives all the way to containerized applications running on the GPUs in the DGX system. With DDN's true end-to-end parallelism, data is delivered with high-throughput, low-latency, and massive concurrency in transactions. This ensures applications achieve the most from DGX systems with all GPU cycles put to productive use. Optimized parallel data-delivery directly translates to increased application performance and faster completion times. The DDN A³I shared parallel architecture also contains redundancy and automatic failover capability to ensure high reliability, resiliency, and data availability in case a network connection or server becomes unavailable.

DDN A³I STREAMLINED DEEP LEARNING

DDN A³I solutions enable and accelerate end-to-end data pipelines for deep learning (DL) workflows of all scale running on Bullsequana XH3000. The DDN shared parallel architecture enables concurrent and continuous execution of all phases of DL workflows across multiple Bullsequana XH3000. This eliminates the management overhead and risks of moving data between storage locations.
At the application level, data is accessed through a standard highly interoperable file interface, for a familiar and intuitive user experience.

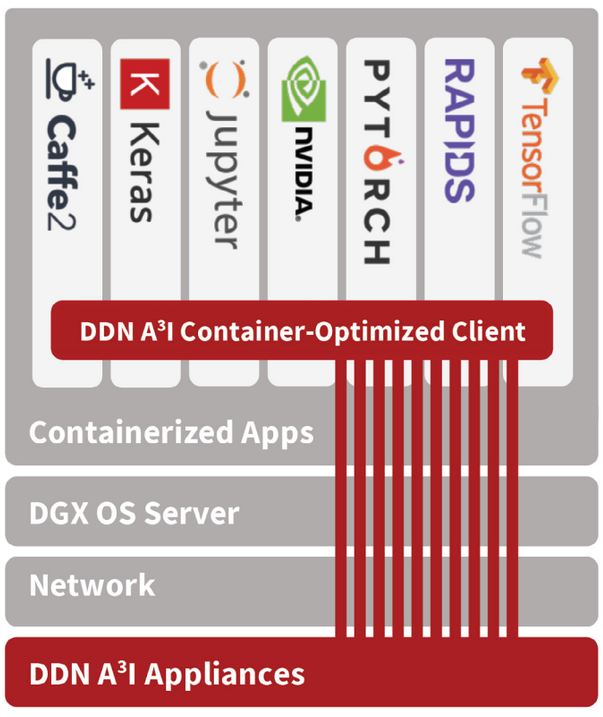
Significant acceleration can be achieved by executing an application across multiple Bullsequana XH3000 racks simultaneously and engaging parallel training efforts of candidate neural networks variants. These advanced optimizations maximize the potential of DL frameworks. DDN works closely with NVIDIA and its customers to develop solutions and technologies that allow widely-used DL frameworks to run reliably on Bullsequana XH3000.



DDN A³I MULTIRAIL NETWORKING

DDN A³I solutions integrate a wide range of networking technologies and topologies to ensure streamlined deployment and optimal performance for AI infrastructure. Latest generation InfinibBand Network and NVIDIA network switches, DDN A31 Multirail greatly simplifies and optimizes DGX system networking for fast secure, and resilient connectivity.

DDN A³I Multirail enables grouping of multiple network interfaces on a DGX system to achieve faster aggregate data transfer capabilities. The feature balances traffic dynamically across all the interfaces, and actively monitors link health for rapid failure detection and automatic recovery. DDN A³I Multirail makes designing, deploying, and managing high-performance networks very simple, and is proven to deliver complete connectivity for at-scale infrastructure for Bullsequana XH3000 supercomputer deployments.

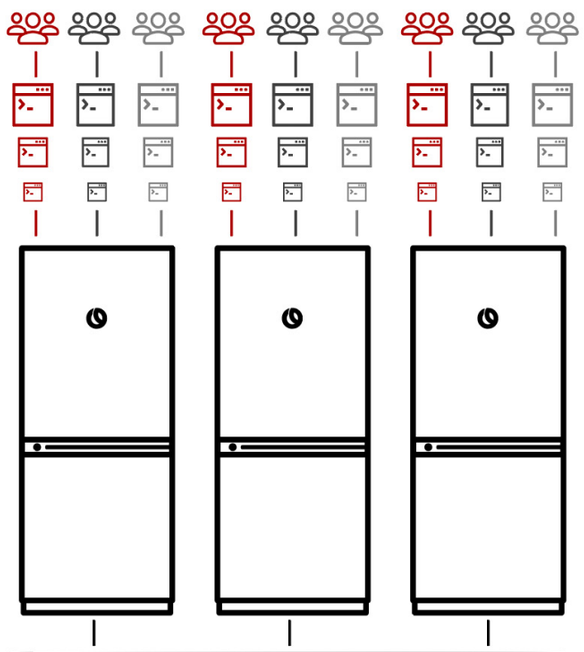


DDN A³I CONTAINER CLIENT

Containers encapsulate applications and their dependencies to provide simple, reliable, and consistent execution. DDN enables a direct high-performance connection between the application containers on BullSequana XH3000 and DDN parallel filesystem. This brings significant application performance benefits by enabling low latency, high-throughput parallel data access directly from a container. Additionally, the limitations of sharing a single host-level connection to storage between multiple containers disappear. The DDN in-container filesystem mounting capability is added at runtime through a universal wrapper that does not require any modification to the application or container.

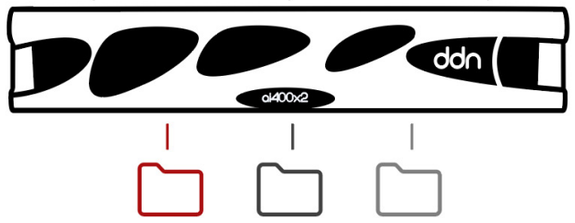
Containerized versions of popular DL frameworks specially optimized for GPUs are available available. They provide a solid foundation that enables data scientists to rapidly develop and deploy applications on BullSequana XH3000. In some cases, open-source versions of the containers are available, further enabling access and integration for developers.

The DDN A3I container client provides high-performance parallelized data access directly from containerized applications on BullSequana XH3000. This provides containerized DL frameworks with the most efficient dataset access possible, eliminating all latencies introduced by other layers of the computing stack.

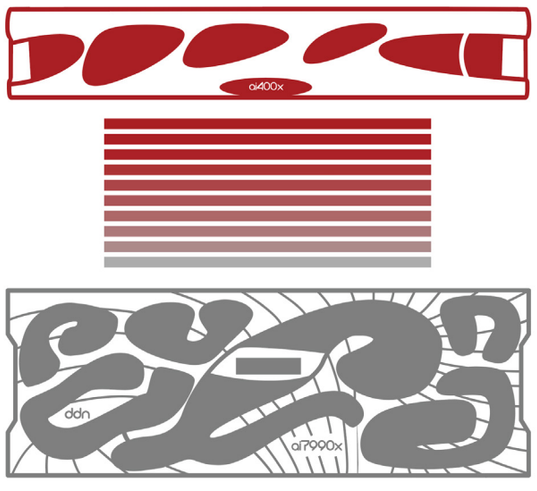


DDN A³I MULTITENANCY

Container clients provide a simple and very solid mechanism to enforce data segregation by multitenant environment at-scale through its native container client and comprehensive digital security framework. DDN A3I multitenancy makes it simple to share BullSequana XH3000 racks across a large pool of users and still maintain secure data segregation. Multi-tenancy provides quick, seamless, dynamic BullSequana XH3000 resource provisioning for users. It eliminates resource silos, complex software release management, and unnecessary data movement between data storage locations. DDN A3I brings a very powerful multitenancy capability to BullSequana XH3000 and makes it very simple for customers to deliver a secure, shared innovation space, for at-scale data-intensive applications.

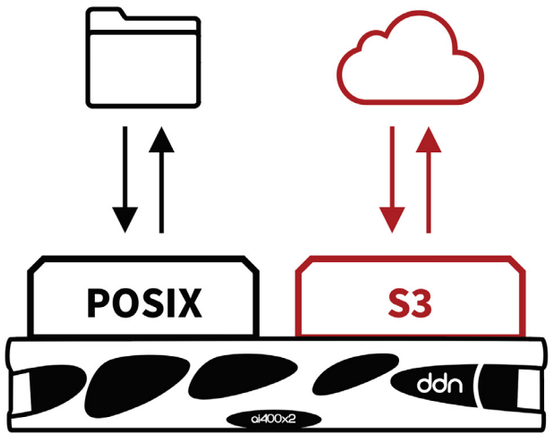


Containers bring security challenges and are vulnerable to unauthorized privilege escalation and data access. The DDN A3I digital security framework provides extensive controls, including a global root_squash to prevent unauthorized data access or modification from a malicious user, and even if a node or container are compromised.



DDN A³I HOT POOLS

Hot Pools delivers user transparent automatic migration of files between the Flash tier (Hot Pool) to HDD tier (Cool Pool). Hot Pools is designed for large scale operations, managing data movements natively and in parallel, entirely transparently to users. Based on mature and well tested file level replication technology, Hot Pools allows organizations to optimize their economics – scaling HDD capacity and/or Flash performance tiers independently as they grow.



DDN A³I S3 DATA SERVICES

DDN S3 Data Services provide hybrid file and object data access to the shared namespace. The multi-protocol access to the unified namespace provides tremendous workflow flexibility and simple end-to-end integration. Data can be captured directly to storage through the S3 interface and accessed immediately by containerized applications on BullSequana XH3000 through a file interface. The shared namespace can also be presented through an S3 interface, for easy collaboration with multisite and multicloud deployments. The DDN S3 Data Services architecture delivers robust performance, scalability, security, and reliability features.

2. DDN A³I SOLUTIONS WITH ATOS BULLSEQUANA XH3000

The DDN A³I scalable architecture integrates BullSequana XH3000 supercomputer with DDN AI shared parallel file storage appliances and delivers fully-optimized end-to-end AI, Analytics and HPC workflow acceleration on GPUs. DDN A³I solutions greatly simplify the deployment of BullSequana XH3000 systems, while also delivering performance and efficiency for maximum GPU saturation, and high levels of scalability.

This section describes the components integrated in DDN A³I Solutions for Atos Systems.

2.1 DDN AI400X2 APPLIANCE

The AI400X2 appliance is a fully integrated and optimized shared data platform with predictable capacity, capability, and performance. Every AI400X2 appliance delivers over 90 GB/s and 3M IOPS directly to GPU blades in the BullSequana XH3000 compute blades. Shared performance scales linearly as additional AI400X2 appliances are integrated to the solution. The all-NVMe configuration provides optimal performance for a wide variety of workload and data types and ensures that BullSequana XH3000 system operators can achieve the most from at-scale GPU applications, while maintaining a single, shared, centralized data platform.

The AI400X2 appliance integrates the DDN A³I shared parallel architecture and includes a wide range of capabilities described in section 1, including automated data management, digital security, and data protection, as well as extensive monitoring. The AI400X2 appliances enables BullSequana XH3000 operators to go beyond basic infrastructure and implement complete data governance pipelines at-scale.

The AI400X2 appliance integrates over IB, Ethernet and RoCE. It is available in 30, 60, 120, 250 and 500 TB all-NVMe capacity configurations. Optional hybrid configurations with integrated HDDs are also available for deployments requiring high-density deep capacity storage. Contact DDN Sales for more information.

Figure 1. DDN AI400X2 all-NVME storage appliance.

2.2 ATOS BULLSEQUANA XH3000 SUPERCOMPUTER

The greener accelerated hybrid HPC platform

By combining cutting-edge processing technologies, Atos 4th generation (DLC) Direct Liquid Cooling technology, and an architecture that is flexible, dense and secured by design, BullSequana XH3000 delivers both unprecedented performance and unrivaled efficiency.

Furthermore, with its OpenSequana program, Atos enables 3rd party technology ecosystem partners to develop a supported compute blade with their technology embedded. The result is is a future-proof platform ready for quantum accelerators.

Figure 2. Atos BullSequana XH3000 supercomputer.

2.3 ATOS FASTML

Atos FastML is a software suite dedicated to artificial intelligence, adapting to multiple use cases, leveraging the technologies of Machine Learning and Deep Learning to enable rapid deployment of software environments for data analysis and artificial intelligence on supercomputers (HPC). Atos FastML brings to data-scientists the following features:

- Single UI for data-scientists to manage data science experiments
- Command line interface
- Secured end-users environment with Atos software SSO using Keycloak
- Leverage AI opensource tools (JupyterHub, TensorFlow, pyTorch, Keras, etc.)
- Leverage HPC environment (Slurm, Sylabs SingularityPRO, etc.)
- Target HPC & AI hardware solutions

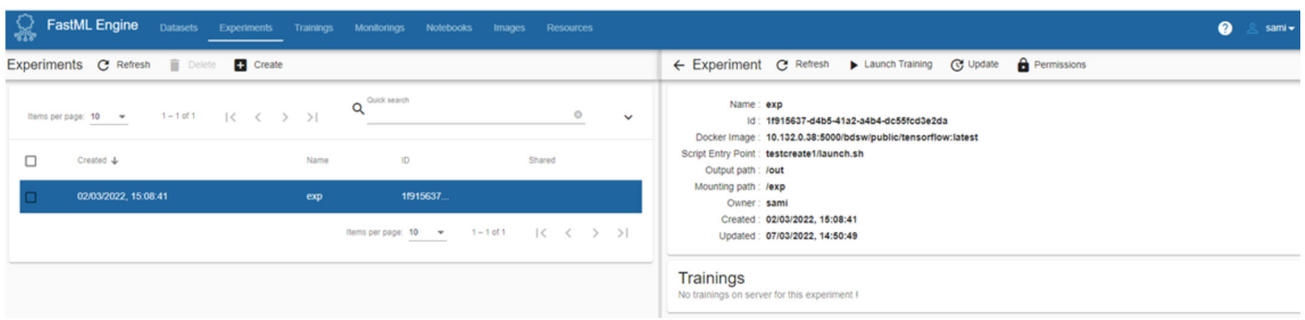


Figure 3. Atos FastML User Interface.

2.4 ATOS MANAGEMENT AND LOGIN NODES

The BullSequana SMC management servers are designed to function as a fully redundant, fault-tolerant system. It consists of two 1U servers (Bull Sequana SMC-Server), both connected to a single JBOD Chassis. A shared filesystem is setup on these disks so that the Cluster configuration is always accessible even in case of a failure of one of the two management nodes.

These servers are the main access point administering BullSequana XH3000 infrastructure. It runs Management Software and is in charge of deploying Bare metal Image on each component of the system, monitoring the infrastructure and providing alerts to the administrators of the systems.

Figure 4. Atos BullSequana SMC management server.

BullSequana X430-E7 server, the Login node, provides the entry point of all supercomputer clients, with requirements-based separation of accounts. BullSequana X430-E7 is a 2U rack-mounted 2-socket server. It is ideally suited as a service node, since its advanced connectivity features, its extended storage options and its redundancy features guarantee efficient and reliable cluster administration services

Figure 5. Atos login node - BullSequana X430-E7.

2.5 ATOS DLC NETWORK SWITCH BLADE

Atos Direct Liquid Cooling 1U NDR switch blade delivers high-speed networking to exchange data between compute blades and storage. It embeds NVIDIA Quantum-2 baseboard and delivers an unprecedented 64 ports of NDR 400Gb/s InfiniBand per port in a 1U standard chassis design. A single switch carries an aggregated bidirectional throughput of 51.2 terabits per second (Tb/s), with a landmark of more than 66.5 billion packets per second (BPPS) capacity. Supporting the latest NDR technology, NVIDIA Quantum-2 brings a high-speed, extremely low-latency and scalable solution that incorporates state-of-the-art technologies such as Remote Direct Memory Access (RDMA), adaptive routing, and NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™.Unlike any other networking solution, NVIDIA InfiniBand provides self-healing network capabilities, as well as quality of service (QoS), enhanced virtual lane (VL) mapping, and congestion control to provide the highest overall application throughput.

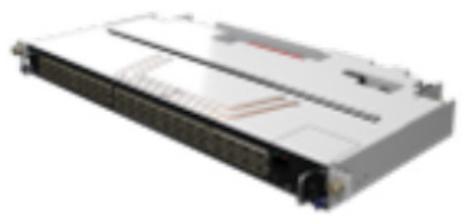


Figure 6. ATOS DLC NDR Switch blade .

3. DDN A³I REFERENCE ARCHITECTURES FOR BULLSEQUANA XH3000

DDN proposes the following reference architectures for BullSequana XH3000 system configurations. DDN A³I solutions are fully-validated with Atos and already deployed with hundreds of GPU customers worldwide.

The DDN AI400X2 appliance is a turnkey appliance for at-scale BullSequana XH3000 system deployments. DDN recommends the AI400X2 appliance as the optimal data platform for AI infrastructure solutions. The AI400X2 appliances delivers optimal GPU performance for every workload and data type in a dense, power efficient 2RU chassis. The AI400X2 appliance simplifies the design, deployment, and management of BullSequana XH3000 and provides predictable performance, capacity, and scaling. The AI400X2 appliance arrives fully configured, ready to deploy and installs rapidly. The appliance is designed for seamless integration with BullSequana XH3000 and enables customers to move rapidly from test to production. As well, DDN provides complete expert design, deployment, and support services globally. The DDN field engineering organization has already deployed hundreds of solutions for customers based on the A³I reference architectures.

As general guidance, DDN recommends two AI400X2 appliances for every BullSequana XH3000 rack . The configuration can be adjusted and scaled easily to match specific workload requirements. For the high-performance network, DDN recommends NDR and HDR200 technology in a non-blocking network topology, with redundancy to ensure data availability. DDN recommends use of at least two NDR400 connections per GPU blade to the high-performance network.

3.1 BULLSEQUANA XH3000 SYSTEM NETWORK ARCHITECTURE

BullSequana XH3000 reference design includes two networks:

High-Performance network. Provides data transfer between the AI400X2 appliance, the GPU blades and login nodes. Connects eight ports from each AI400X2 appliance. Connects four ports from each GPU blade HCAs. Connects two ports from each login and management node HCAs.

Management Network. Provides management and monitoring for all BullSequana XH3000 rack components. Connects the 1 GbE RJ45 Management port and 1 GbE RJ45 BMC port from each GPU blade, login node, management node, and AI400X2 appliance controller to an Ethernet switch.

An overview of BullSequana XH3000 network architecture is shown in figure 7, recommended network connections for each GPU blade on figure 8, and recommended network connections for each DDN AI400X2 appliance on figure 9.

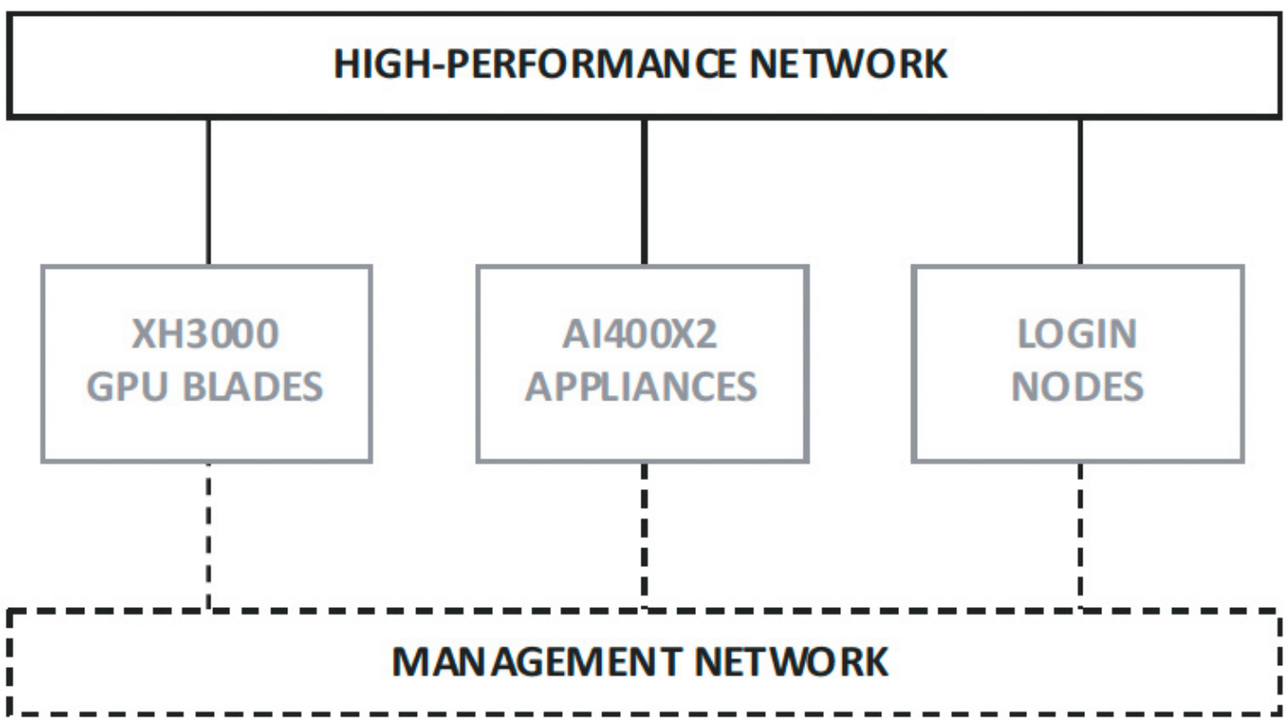


Figure 7. Overview of the XH3000 system network architecture.

BULLSEQUANA XH3000 SYSTEM GPU BLADE NETWORK CONNECTIVITY

DDN recommends ports 1 to 4 on the GPU blade be connected to the high-performance network. As well, the management (“M”) and BMC (“B”) ports should be connected to the management network.

Figure 8. Recommended GPU blade network port connections.

AI400X2 APPLIANCE NETWORK CONNECTIVITY

DDN recommends ports 1 to 8 on the AI400X2 appliance be connected to the high-performance network. As well, the management (“M”) and BMC (“B”) ports for both controllers should be connected to the management network. Note that each AI400X2 appliance requires one inter-controller network port connection (“I”) using short ethernet cable supplied.

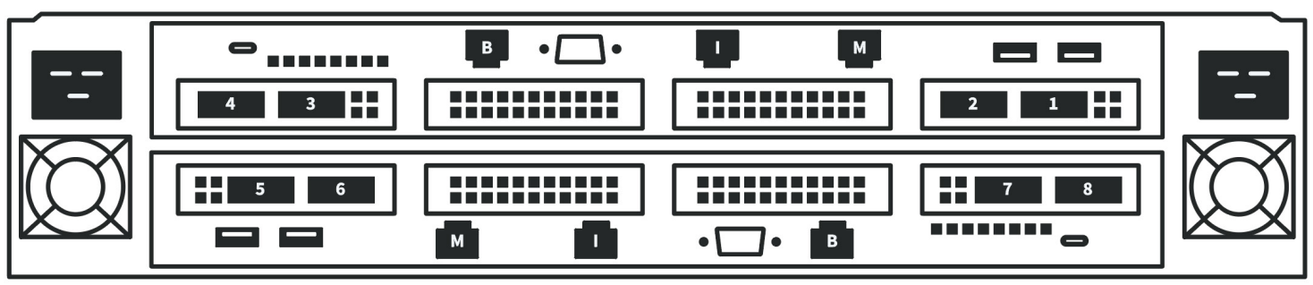
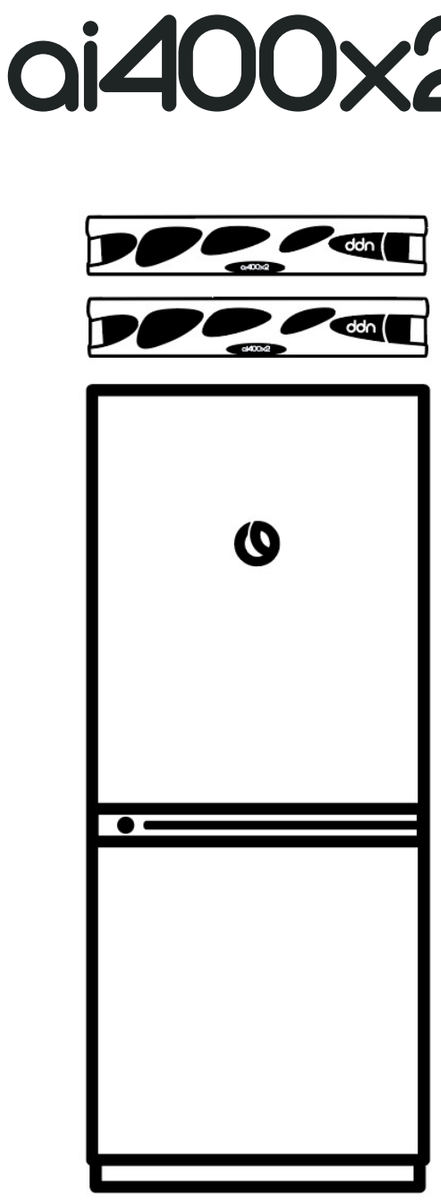


Figure 9. Recommended AI400X2 appliance network port connections



3.2 SINGLE BULLSEQUANA XH3000 RACK CONFIGURATION

Figure 10 illustrates the DDN A3I architecture in a 1:2 configuration in which a BullSequana XH3000 rack is connected with two AI400X2 appliances through high-performance network switches. Every GPU blade connects to the high-performance network switches via four NDR 400Gb/s IB links. Each AI400X2 appliance connects to the high-performance network via eight HDR 200Gb/s IB links using splitter cables to maximize switch port utilization. The management and login nodes also connect to the high-performance network.

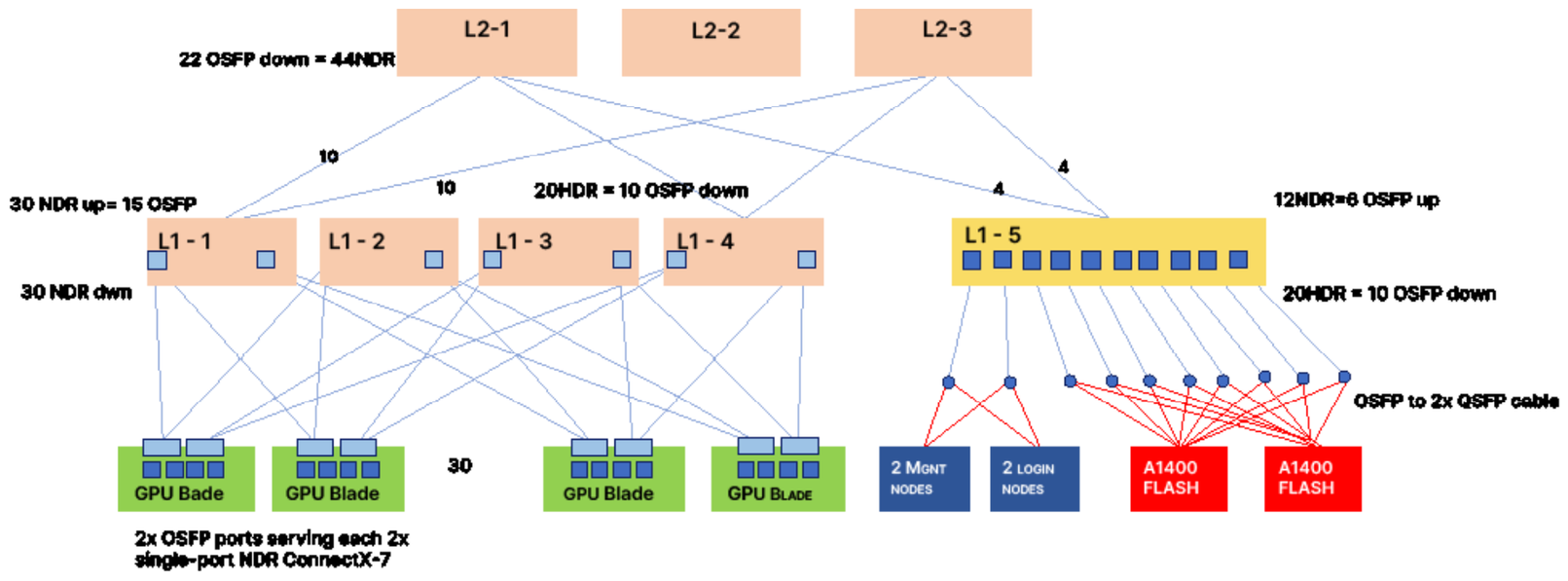


Figure 10. DDN A³I reference architecture with single XH300 systems (management network not shown)..

4. DDN A³I SOLUTIONS VALIDATION

DDN conducts extensive engineering integration, optimization, and validation efforts in close collaboration with Atos to ensure best possible end-user experience using the reference designs in this document. The joint validation confirms functional integration, and optimal performance out-of-the-box for BullSequana XH3000 supercomputers.

Performance testing on the DDN A³I architecture has been conducted with industry standard synthetic throughput and IOPS applications, as well as widely used DL frameworks and data types. The results demonstrate that with the DDN A³I shared parallel architecture, applications can engage the full capabilities of the data infrastructure and BullSequana XH3000. Performance is distributed evenly across all BullSequana XH3000 racks in a multi-node configuration, and scales linearly as more BullSequana XH3000 racks are engaged.

This section details some of the results from recent at-scale testing integrating AI400X2 appliances with BullSequana XH3000.

4.1 SINGLE BULLSEQUANA XH3000 RACK FIO PERFORMANCE VALIDATION

This series of tests demonstrate the peak performance of the reference architecture using the fio open-source synthetic benchmark tool. The tool is set to simulate a general-purpose workload without any performance-enhancing optimizations. Separate tests were run to measure both 100% read and 100% write workload scenarios.

The AI400X2 appliance provides predictable, scalable performance. This test demonstrates the architecture's ability to deliver full throughput performance to a small number of clients and distribute the full performance of the DDN solution evenly as all GPUs in the BullSequana XH3000 compute blades are engaged.

In figure 11, test results demonstrate that DDN solution can deliver over 90 GB/s of read throughput to a single BullSequana XH3000 rack, and evenly distribute the full read and write performance of the AI400X2 appliance with up 120 GPUs engaged simultaneously. The DDN solution can fully saturate network links, ensuring optimal performance for a very wide range of data access patterns and data types for applications running on GPUs in BullSequana XH3000.

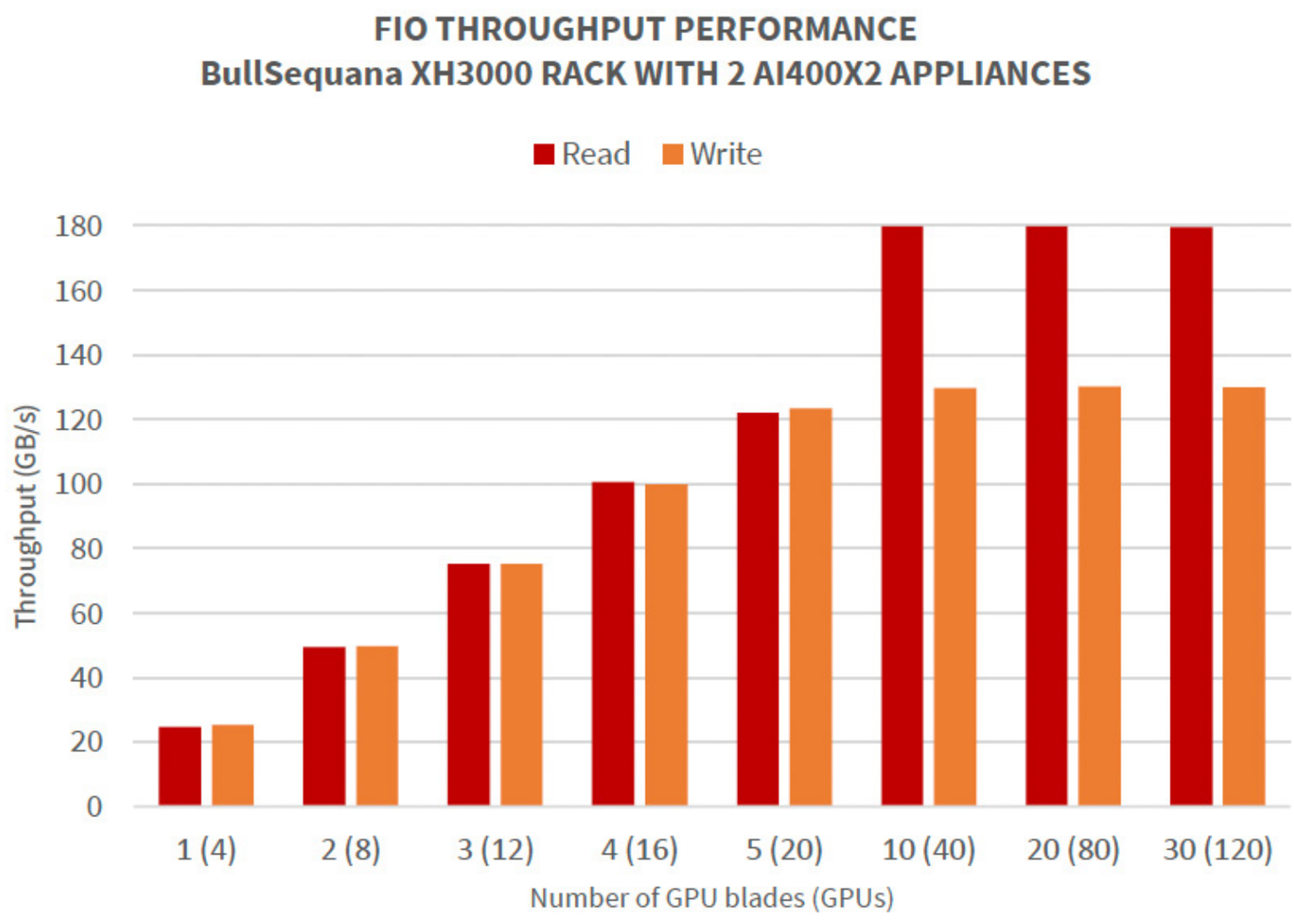


Figure 11. FIO throughput with a single XH3000 system.

4.2 SCALING PERFORMANCE WITH MULTIPLE BULLSEQUANA XH3000 RACKS

The DDN A^{3}I Reference Architectures for BullSequana XH3000 are designed to deliver an optimal balance of technical and economic benefits for a wide range of common use cases for AI, Data Analytics and HPC. Using the AI400X2 appliance as a building block, solutions can scale linearly, predictably and reliably in performance, capacity and capability. For applications with requirements beyond the base reference architecture, it's simple to scale the data platform with additional AI400X2 appliances.

The same AI400X2 appliance and shared parallel architecture used in the DDN A^{3}I Reference Architectures are also deployed with very large AI systems. The AI400X2 appliance has been validated to operate properly with up to 5000 GPUs simultaneously in a production environment.

In figure 12, we show an fio throughput test performed by DDN engineers similar to the one presented in section 4.1. In this example, up to 8 BullSequana XH3000 racks and 960 GPUs are engaged simultaneously with 16 AI400X2 appliances. The results of the test demonstrate that the DDN shared parallel architecture scales linearly and fully achieves the capabilities of the 16 AI400X2 appliances, over 1.4 TB/s throughput for read and 1 TB/s throughput for write, with 16 BullSequana XH3000 racks engaged. This performance is maintained and balanced evenly with up to 16 BullSequana XH3000 racks simultaneously.

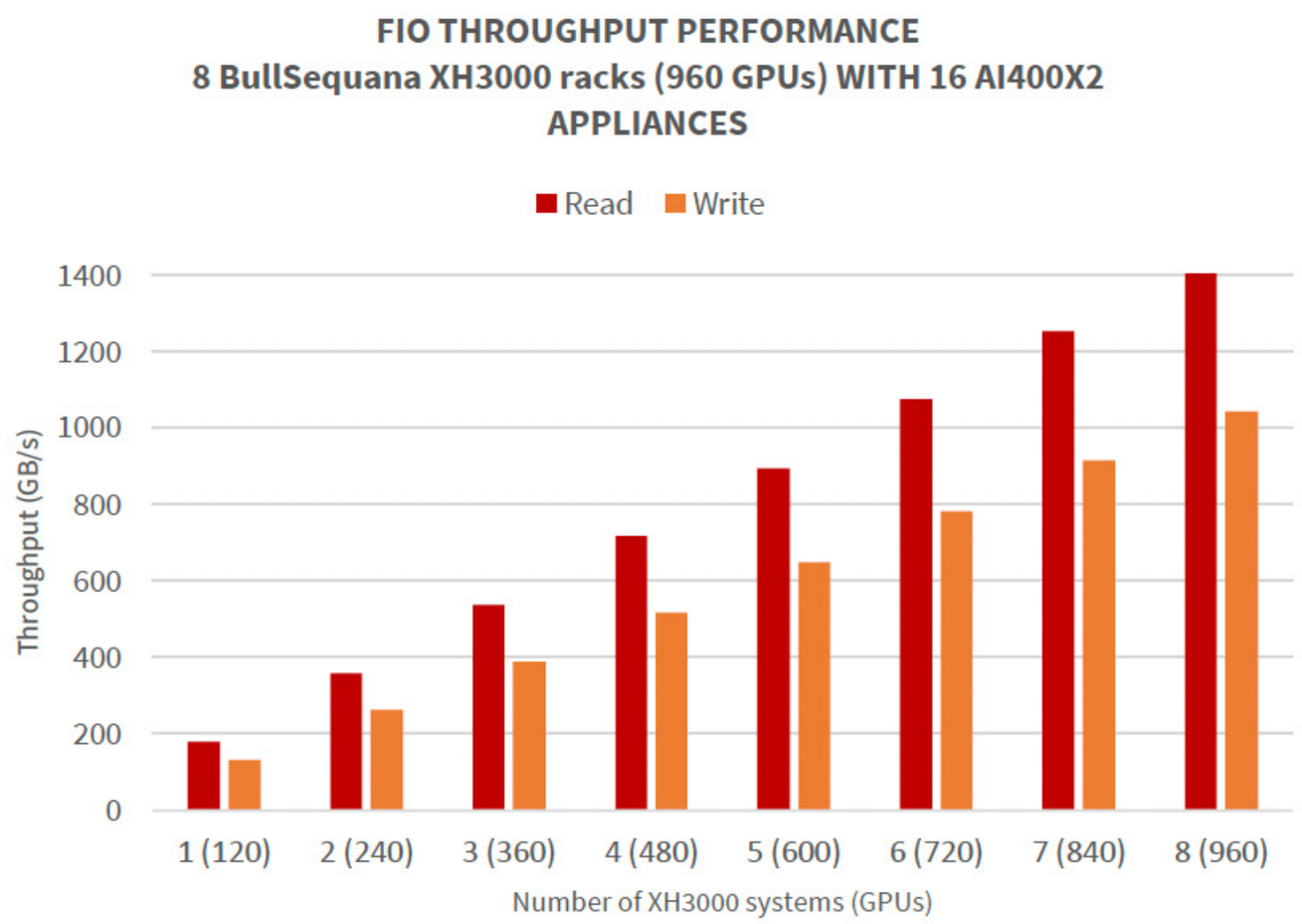


Figure 12. FIO throughput scaling with a very large number of XH3000 systems.

5. CONTACT DDN TO UNLEASH THE POWER OF YOUR BULLSEQUANA XH3000

DDN has long been a partner of choice for organizations pursuing at-scale data-driven projects. Beyond technology platforms with proven capability, DDN provides significant technical expertise through its global research and development and field technical organizations.

A worldwide team with hundreds of engineers and technical experts can be called upon to optimize every phase of a customer project: initial inception, solution architecture, systems deployment, customer support and future scaling needs.

Strong customer focus coupled with technical excellence and deep field experience ensures that DDN delivers the best possible solution to any challenge. Taking a consultative approach, DDN experts will perform an in-depth evaluation of requirements and provide application-level optimization of data workflows for a project. They will then design and propose an optimized, highly reliable and easy to use solution that best enables and accelerates the customer effort.

Drawing from the company's rich history in successfully deploying large scale projects, DDN experts will create a structured program to define and execute a testing protocol that reflects the customer environment and meet and exceed project objectives. DDN has equipped its laboratories with leading GPU compute platforms to provide unique benchmarking and testing capabilities for AI and DL applications.

Contact DDN today and engage our team of experts to unleash the power of your AI projects.