DDN A3I® SOLUTIONS WITH CEREBRAS WAFER-SCALE CLUSTERS

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 Cerebras Wafer-Scale Engine processors in Cerebras CS-2 systems. This document describes fully validated reference architectures and scalable configurations for Cerebras Wafer-Scale Clusters. The solutions integrate DDN AI400X2 appliances.

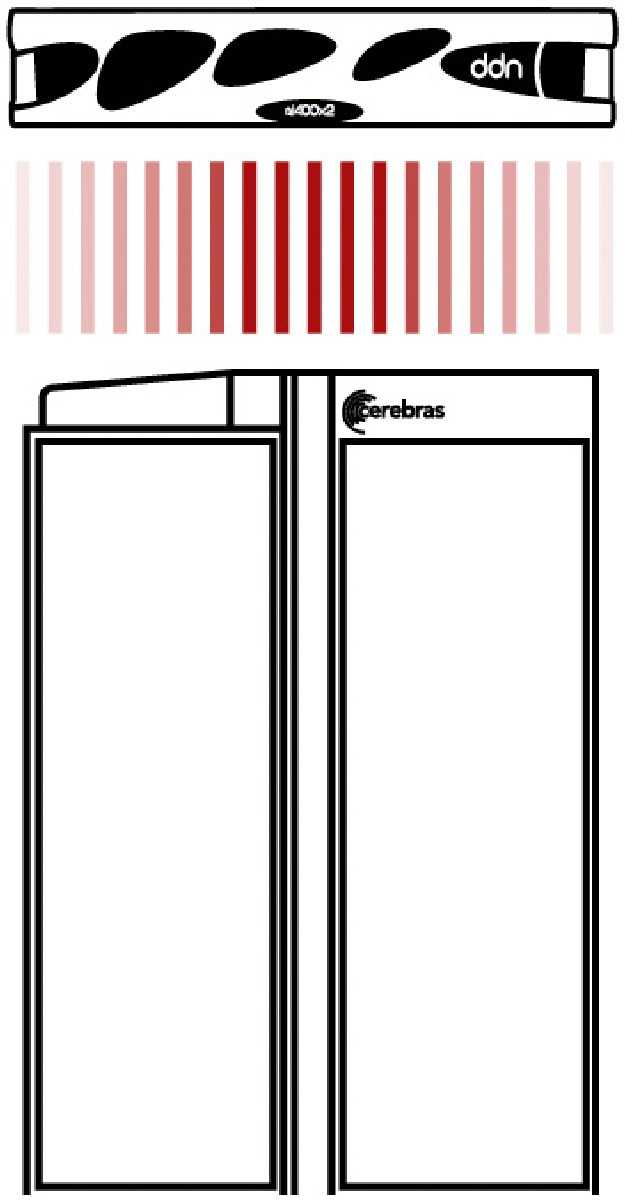
1. DDN A^{3}I END-TO-END ENABLEMENT FOR CEREBRAS SYSTEMS

DDN A^{3}I solutions (Accelerated, Any-Scale AI) are architected to achieve the most from at-scale AI, Data Analytics and HPC applications running on Cerebras CS-2 systems and Cerebras Clusters. They provide predictable performance, capacity, and capability through a tight integration between DDN and Cerebras systems. Every layer of hardware and software engaged in delivering and storing data is optimized for fast, responsive, and reliable access.

DDN A^{3}I solutions are designed, developed, and optimized in close collaboration with Cerebras. The deep integration of DDN AI appliances with CS-2 systems ensures a reliable experience. DDN A^{3}I 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 A^{3}I solutions are deployed globally and at all scale, from a single processor system all the way 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 CS-2 systems that's easy to deploy and manage. DDN A^{3}I solutions are proven to maximum benefits for at-scale AI, Analytics and HPC workloads on CS-2 systems.

This section describes the advanced features of DDN A^{3}I Solutions for Cerebras systems.

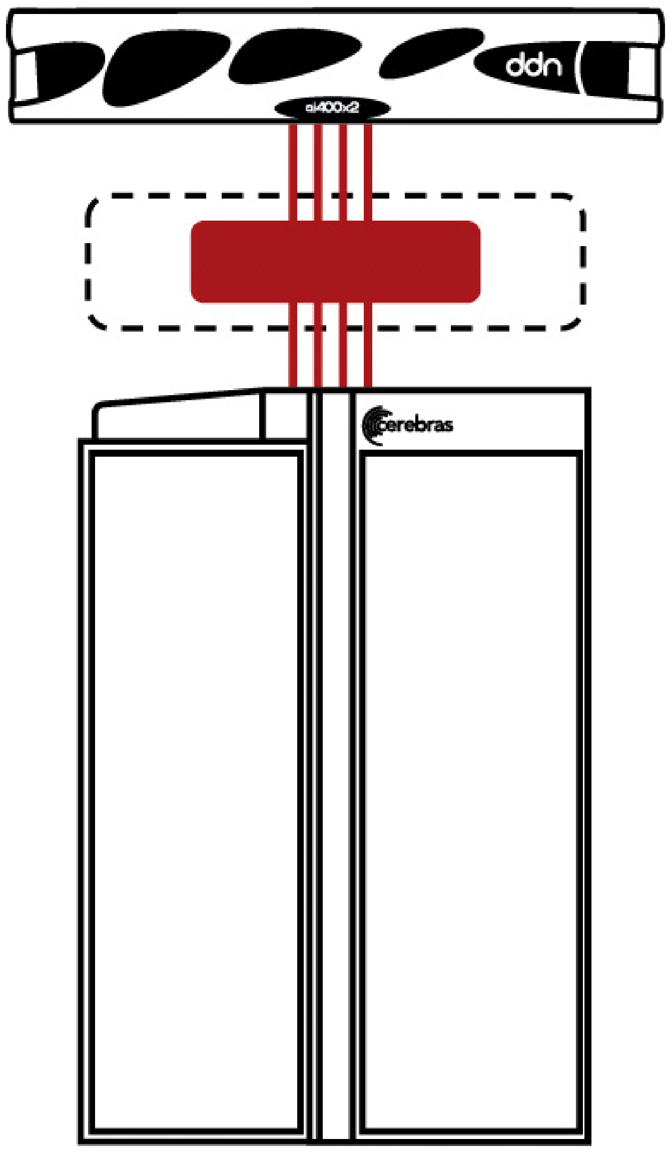


DDN A^{3}I SHARED PARALLEL ARCHITECTURE

The DDN A^{3}I shared parallel architecture and client protocol ensures high levels of performance, scalability, security, and reliability for CS-2 systems. Multiple parallel data paths extend from the drives all the way to containerized applications running on the WSEs in the CS-2 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 CS-2 systems with all WSE cycles put to productive use. Optimized parallel data-delivery directly translates to increased application performance and faster completion times. The DDN A^{3}I 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 A^{3}I STREAMLINED DEEP LEARNING

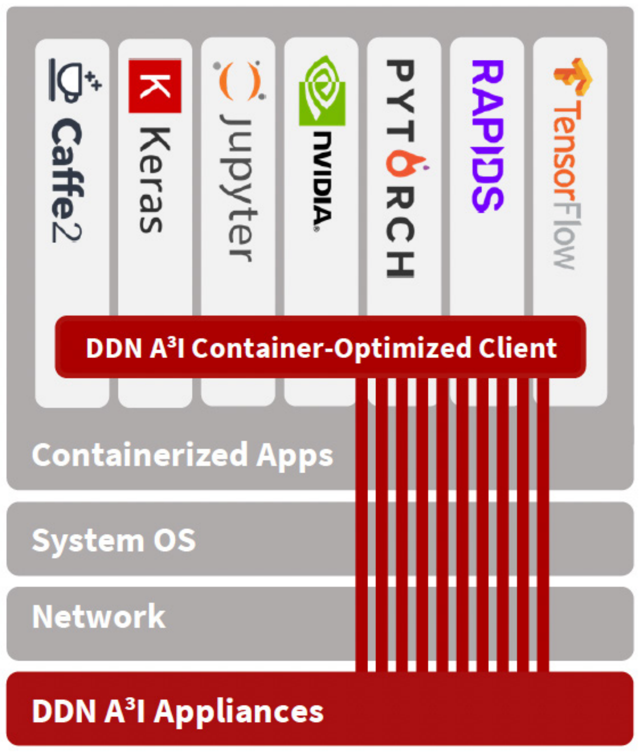
DDN A^{3}I solutions enable and accelerate end-to-end data pipelines for deep learning (DL) workflows of all scale running on CS-2 systems. The DDN shared parallel architecture enables concurrent and continuous execution of all phases of DL workflows across multiple CS-2 systems. This eliminates the 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 CS-2 systems simultaneously and engaging parallel training efforts of candidate neural networks variants. These advanced optimizations maximize the potential of DL frameworks. DDN works closely with Cerebras and its customers to develop solutions and technologies that allow widely-used DL frameworks to run reliably on CS-2 systems.



DDN A^{3}I MULTIRAIL NETWORKING

DDN A^{3}I solutions integrate a wide range of networking technologies and topologies to ensure streamlined deployment and optimal performance for AI infrastructure. Latest generation Ethernet with RoCE provides both high-bandwidth and low-latency data transfers between applications, compute servers and storage appliances. For Cerebras Wafer-Scale Clusters DDN recommends an Ethernet Network. DDN A^{3}I Multirail greatly simplifies and optimizes Cerebras Cluster networking for fast, secure, and resilient connectivity. DDN A^{3}I Multirail enables grouping of multiple network interfaces to achieve faster aggregate data transfer capabilities. The feature balances traffic dynamically across all the interfaces, andactively monitors link health for rapid failure detection and automatic recovery. DDN A^{3}I Multirail makes designing, deploying, and managing high-performance networks very simple, and is proven to deliver complete connectivity for at-scale infrastructure and deployments.

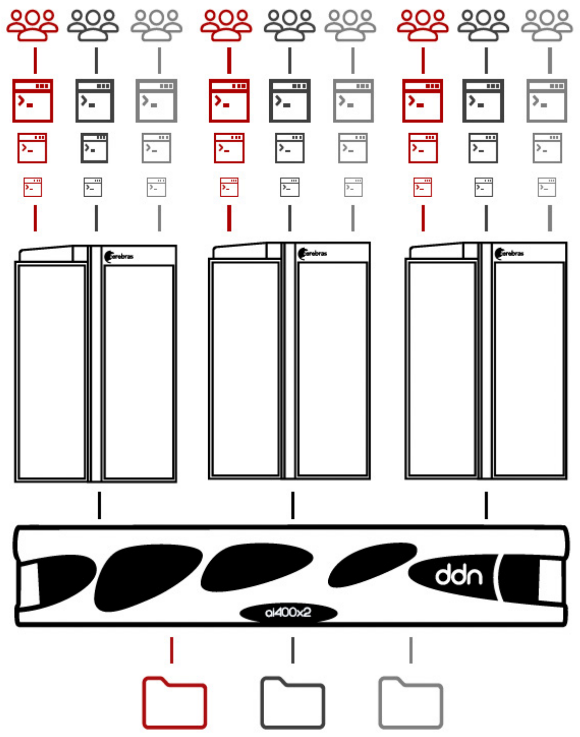
DDN A3I FULLY-PARALLEL APPLICATION AND CONTAINER CLIENT



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 the processor system and the 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.

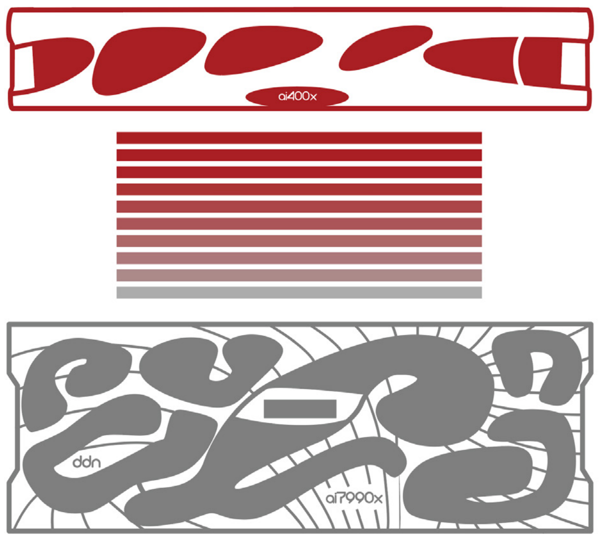
Containerized versions of popular DL frameworks specially optimized for the processor system are available. They provide a solid foundation that enables data scientists to rapidly develop and deploy applications on the processor system. In some cases, open-source versions of the containers are available, further enabling access and integration for developers. The DDN A³I container client provides high-performance parallelized data access directly from containerized applications on the processor system. 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



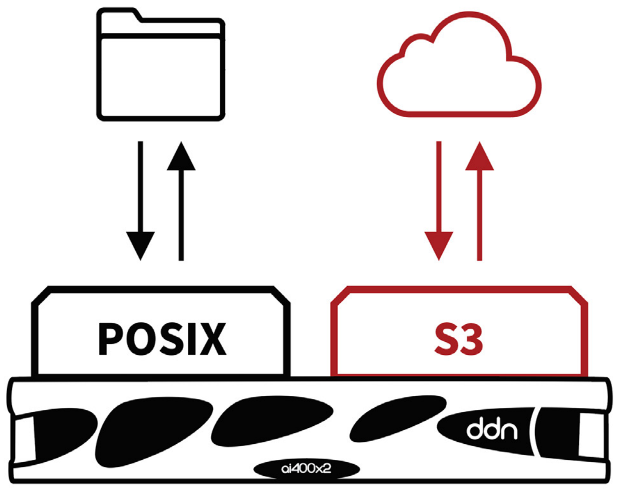
Container clients provide a simple and very solid mechanism to enforce data segregation by restricting data access within a container. DDN A³I makes it very simple to operate a secure multitenant environment at-scale through its native container client and comprehensive digital security framework. It eliminates resource silos, complex software release management, and unnecessary data movement between data storage locations. DDN A³I brings a very powerful multitenancy capability to CS-2 systems and makes it very simple for customers to deliver a secure, shared innovation space, for at-scale data-intensive applications.

Multi-tenant environments 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 A^{3}I 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 A^{3}I 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 Cerebras installations 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 CEREBRAS SYSTEMS AND CLUSTERS

The DDN A³I scalable architecture integrates Cerebras installations with DDN AI shared parallel file storage appliances and delivers fully-optimized end-to-end AI, Analytics and HPC workflow acceleration on Wafer-scale Engines. DDN A³I solutions greatly simplify the deployment of complete infrastructure while also delivering performance and efficiency for maximum processor saturation, and high levels of scalability.

This section describes the components integrated in DDN A³I Solutions for single node and scale-out Wafer-scale Clusters configurations.

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 data processing servers connected to the CS-2 systems. 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 system operators can achieve the most from at-scale 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 system 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 CEREBRAS CS-2 SYSTEM

The Cerebras CS-2 system is the industry's fastest AI accelerator. It reduces training times from months to minutes, and inference latencies from milliseconds to microseconds. And the CS-2 requires only a fraction of the space and power of graphics processing unit-based AI compute. The CS-2 features 850,000 AI-optimized compute cores, 40GB of on-chip SRAM, 20 PB/s memory bandwidth and 220Pb/s interconnect, all enabled by purpose-built packaging, cooling, and power delivery. Every design choice has been made to accelerate deep learning, reducing training times and inference latencies by orders of magnitude.

The CS-2 is easily installed into a standard datacenter infrastructure — from loading dock to users' hands in a few days rather than weeks or months that is typically required for traditional cluster provisioning.

Figure 2. Cerebras CS-2 systems installed in a datacenter.

2.3 CEREBRAS SOFTWARE PLATFORM (CSOFT)

The Cerebras software platform integrates with popular machine learning frameworks like TensorFlow and PyTorch, so researchers can use familiar tools and rapidly bring their models to the CS-2.

The Cerebras Software Platform includes an extensive library of standard deep learning primitives and a complete suite of debug and profiling tools.

The Cerebras SDK enables developers to extend the platform for their work, harnessing the power of wafer-scale computing to accelerate their development needs. With the SDK and the Cerebras Software Language (CSL), developers can target the WSE’s microarchitecture directly using a familiar C-like interface for developing software kernels.

2.4 CEREBRAS WAFER-SCALE CLUSTER

The Cerebras Cluster architecture – comprised of hardware, software and fabric technology – enables clusters of up to 192 Cerebras CS-2s to be rapidly deployed. Because these clusters run data parallel, there is no distributed computing, and they deliver near-perfect linear scaling performance, making them perfect for extreme-scale AI.

Traditionally cluster distributing work over a large cluster is punishingly difficult, even for expert teams. Trying to run a single model on thousands of devices requires the scaling of memory, compute and bandwidth. This is an interdependent distributed constraint problem.

Cerebras cluster users can launch massive training jobs from anywhere, including from a Jupyter notebook, using a Python workflow that makes interacting with teraflops of compute and millions of compute cores as simple as running an application on a laptop. CSoft, automatically, and invisibly, allocates compute resources inside a CS-2 cluster with no performance overhead and no parallel programming.

Figure 3. Cerebras Cluster installed in a datacenter.

3. DDN A3I REFERENCE ARCHITECTURES FOR CEREBRAS WAFER-SCALE CLUSTERS

DDN proposes the following reference architectures for single node and scale-out Cerebras Wafer-Scale Clusters configurations. DDN A³I solutions with Cerebras and already deployed with several joint customers worldwide.

The DDN AI400X2 appliance is a turnkey appliance for high-performance AI infrastructure deployments. DDN recommends the AI400X2 appliance as the optimal data platform for Cerebras Clusters. The AI400X2 appliances delivers optimal processor performance for every workload and data type in a dense, power efficient 2RU chassis. The AI400X2 appliance simplifies the design, deployment, and management of supplying training data to a Cerebras Cluster 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 Cerebras systems 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 an AI400X2 appliance for every CS-2 node of a Cerebras Cluster. These configurations can be adjusted and scaled easily to match specific workload requirements. For the network, Cerebras recommends 100GbE technology in a non-blocking topology for performance, with redundancy to ensure data availability.

3.1 NETWORK ARCHITECTURE

The Wafer-scale Cluster reference design includes two networks:

Storage network. Provides data transfer between the AI400X2 appliance and the data pre-processing servers connected to the CS-2 nodes. Connects eight ports from each AI400X2 appliance over 100GbE, using RoCE for optimal performance and efficiency.

Management Network. Provides management and monitoring for all Cerebras Cluster components. Connects the 1 GbE RJ45 Management port and 1 GbE RJ45 BMC port from each AI400X2 appliance controller to an Ethernet switch.

AI400X2 APPLIANCE NETWORK CONNECTIVITY

For Cerebras Clusters, DDN recommends ports 1 to 8 on the AI400X2 appliance be connected to the storage 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 the short ethernet cable supplied.

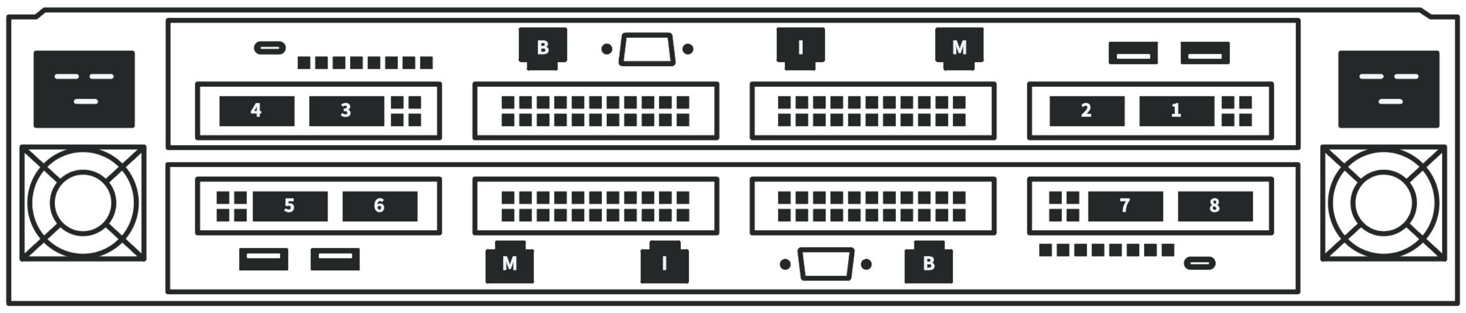


Figure 4. Recommended AI400X2 appliance network port connections.

3.2 SINGLE CS-2 SYSTEM CONFIGURATION

Figure 5 illustrates the DDN A³I architecture in a 1:1 configuration in which a single CS-2 system is connected to an AI400X2 appliance through in-rack network infrastructure. Every data pre-processing server connects to the integrated network with at least one 100 GbE link. The AI400X2 appliance connects to the integrated network via eight 100 GbE links.

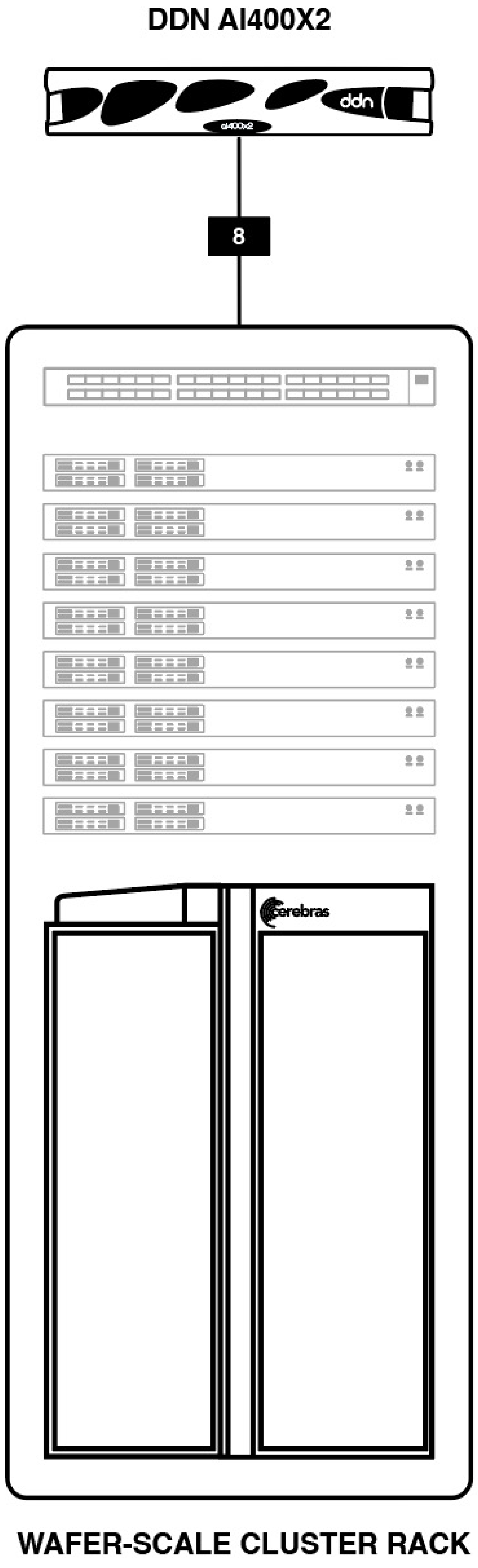


Figure 5. DDN A³I reference architecture with single CS-2 System (management network not shown).

3.3 MULTIPLE CS-2 SYSTEMS CONFIGURATION

Figure 6 illustrates a scalable DDN A³I architecture in which racks in a 1:1 configuration are connected to form a cluster. The configuration scales simply and easily simply by interconnecting network switches between racks. The DDN shared storage is presented as a single unified namespace across all appliances deployed, and data is available to all data pre-processing servers in the cluster.

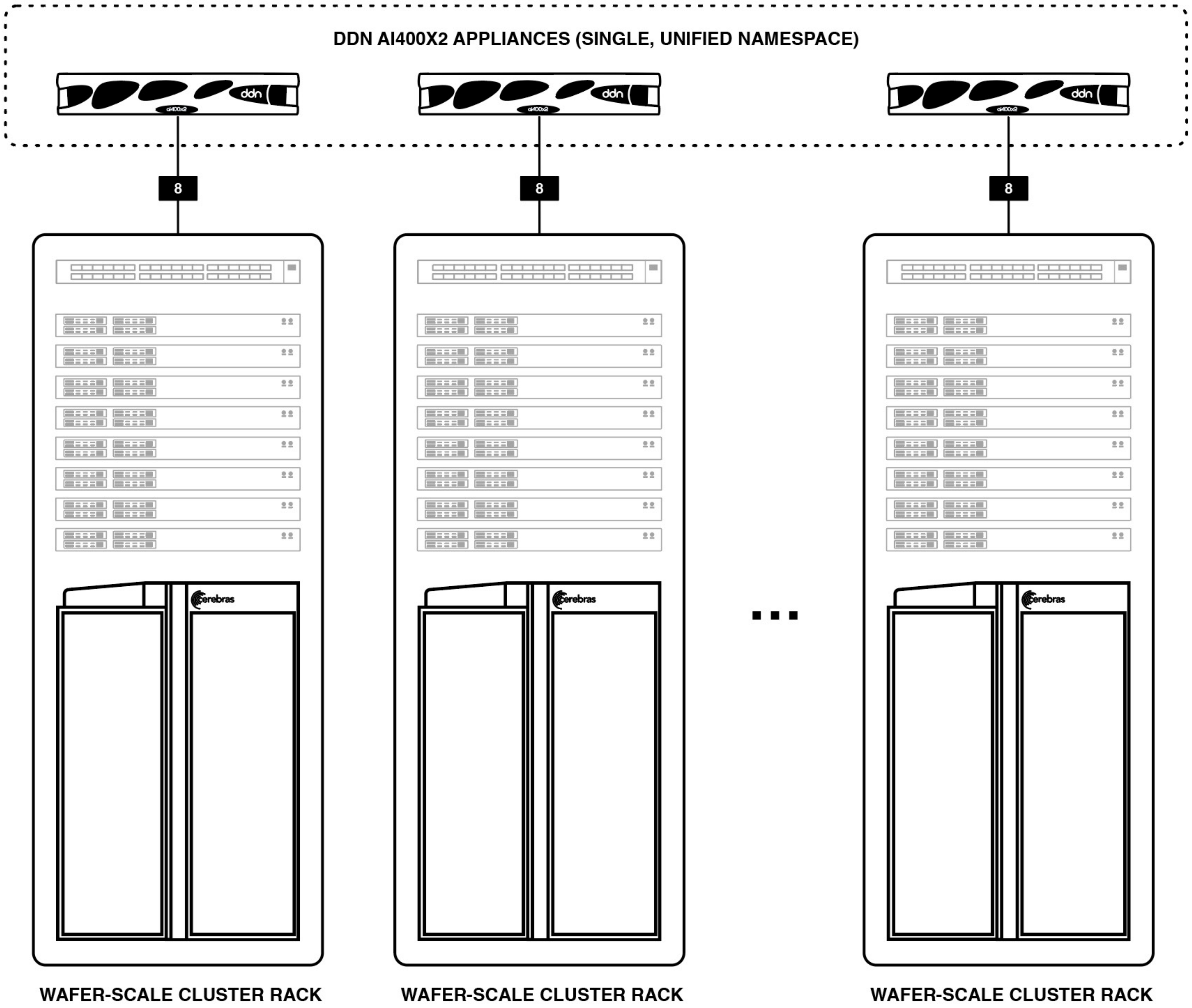


Figure 6. DDN A³I reference architecture with multiple CS-2 nodes (management network not shown).

4. DDN A3I SOLUTIONS VALIDATION

DDN conducts extensive engineering integration, optimization, and validation efforts in close collaboration with Cerebras 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 CS-2 systems.

Performance testing on the DDN A^{3}I 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 A^{3}I shared parallel architecture, applications can engage the full capabilities of the data infrastructure and the CS-2 systems. Performance is distributed evenly across all the CS-2 systems in a multi-node configuration, and scales linearly as more CS-2 nodes are engaged.

This section details some of the results from recent at-scale testing integrating AI400X2 appliances with CS-2 systems.

4.1 SINGLE CS-2 SYSTEM 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 the data pre-processing servers connected to the CS-2 system are engaged.

In figure 8, test results demonstrate that DDN solution can deliver over 90 GB/s of read throughput to the data pre-processing servers, and evenly distribute the full read and write performance of the AI400X2 appliance 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 a CS-2 system.

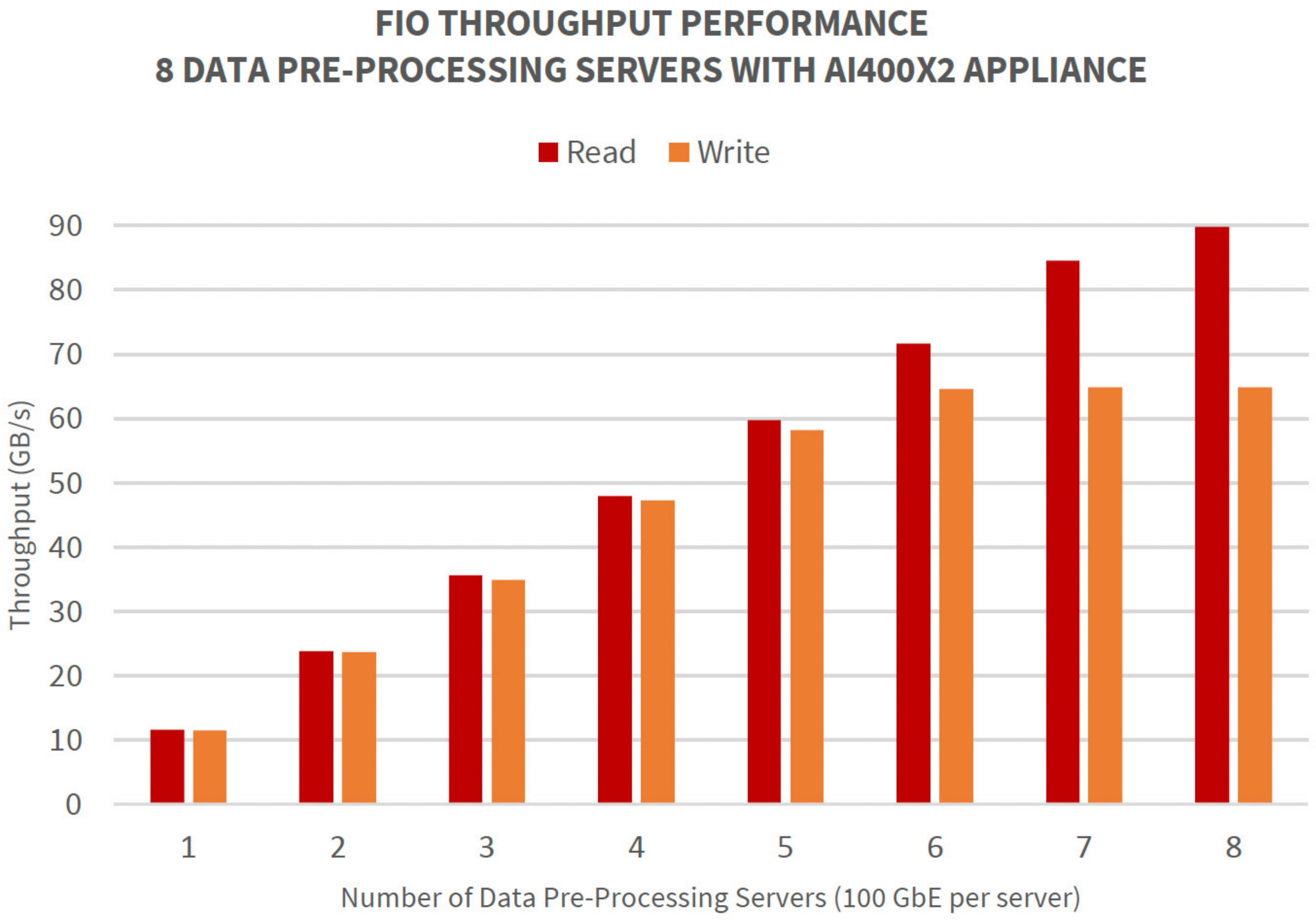


Figure 8. FIO throughput with a single CS-2 node.

4.2 SCALING PERFORMANCE WITH MULTIPLE CS-2 SYSTEMS

The DDN A^{3}I Reference Architectures for CS-2 systems 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 AI processors systems simultaneously in a production environment.

In figure 9, we show a synthetic fio throughput test performed by DDN engineers similar to the one presented in section 4.1. In this example, a Cerebras Cluster with up to 16 CS-2 nodes is 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 CS-2 nodes engaged. This performance is maintained and balanced evenly with up to 16 CS-2 nodes simultaneously.

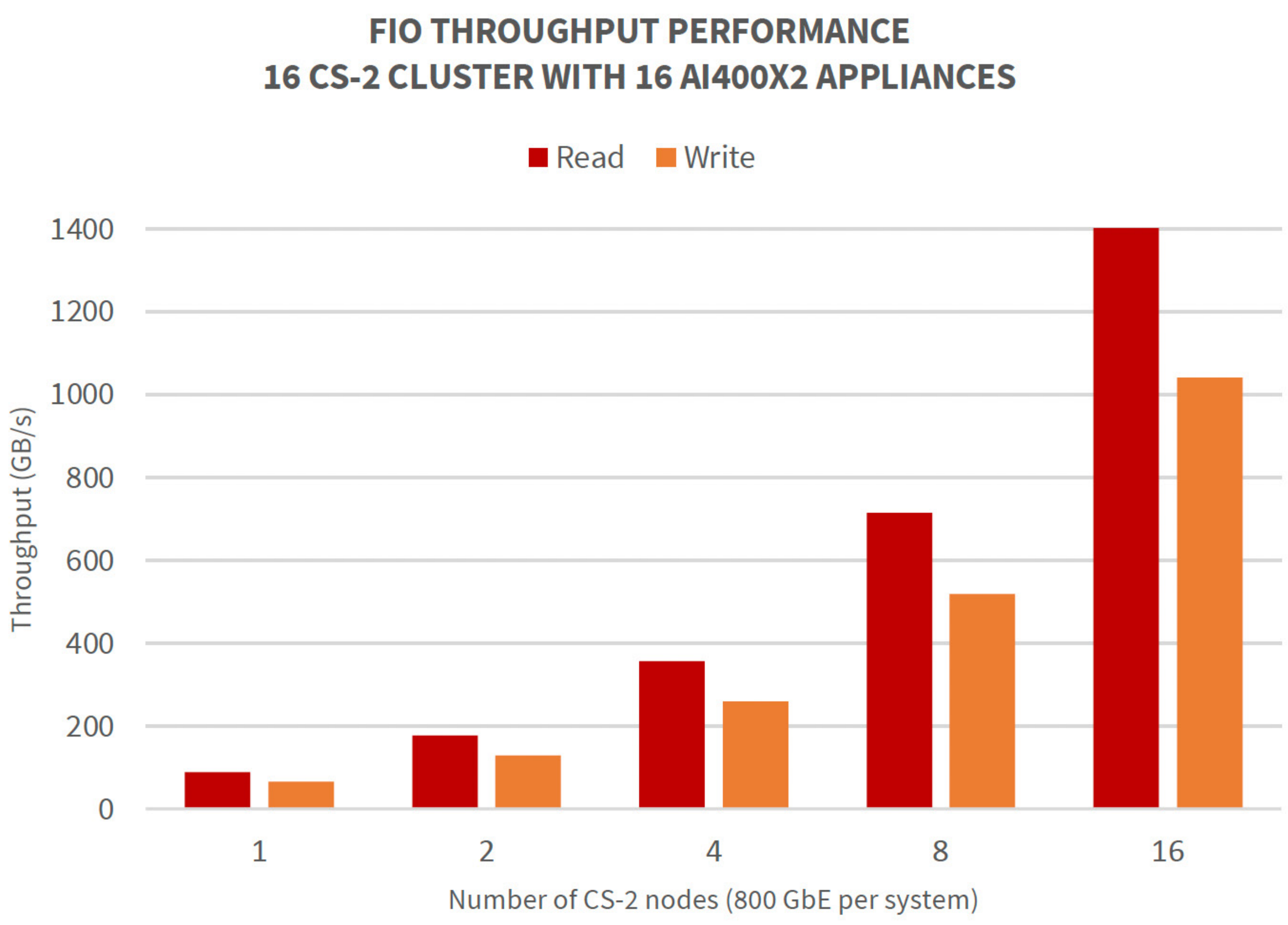


Figure 9. FIO throughput scaling with Cerebras Clusters of different sizes.

5. CONTACT DDN TO UNLEASH THE POWER OF YOUR CEREBRAS SYSTEMS

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 compute platforms to provide unique benchmarking and testing capabilities for Ai, Analytics and HPC applications.

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

About DDN

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. DDN has designed, developed, deployed, and optimized systems, software, and solutions that enable enterprises, service providers, research facilities, and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud.