SOLUTION BRIEF

EXTREME FILE SYSTEM PERFORMANCE ON SUPERMICRO SERVERS

Award Winning Twin Servers with 3rd Gen Intel® Xeon® Scalable Processors

TABLE OF CONTENTS

SUPERMICRO

Supermicro (Nasdaq: SMCI), the leading innovator in high-performance, high-efficiency server and storage technology is a premier provider of advanced server Building Block Solutions® for Enterprise Data Center, Cloud Computing, Artificial Intelligence, and Edge Computing Systems worldwide. Supermicro is committed to protecting the environment through its "We Keep IT Green®" initiative and provides customers with the most energy-efficient, environmentally-friendly solutions available on the market.

Executive Summary

Unstructured data has become the backbone of IT infrastructure for consumer apps, business intelligence, financial services, media & entertainment, logistics & transportation, municipal services, education, scientific research, healthcare, government operations, and national security. At the application level, end-users may not realize how much unstructured data impacts their everyday lives and how it influences how they operate in the virtual world. Clients are relentlessly hungry for the most up-to-date and precious information, yet they expect lower latency and 100% uptime. This creates a tremendous amount of pressure on our IT infrastructure with growing data sets at a global scale. Furthermore, the dynamic changes of hot, warm, and cold data challenge the IT infrastructure's ability to meet service level agreements for data mining, processing & analytics amongst Telecommunications, Enterprises, and Technology companies, supporting business operations, mobile applications, IoT, and AI.

WekaIO™ (Weka) was founded on the idea that current storage solutions have forced IT organizations to choose complex solutions to address their highest storage need at the expense of other desirable capabilities. The three dominant architectures are block, file, and object, each servicing a different need: speed, shareability, and scalability in that order. In today's "data-as-a-service" market, organizations need a flexible infrastructure that addresses the many business needs within a single framework. The design philosophy behind the Weka file system – WekaFS - was to create a single storage architecture that runs on-premises or in the public cloud with the performance of all-flash arrays, the simplicity and feature set of network-attached storage (NAS), and the scalability and economics of object storage.

WEKA Solution Architecture Overview

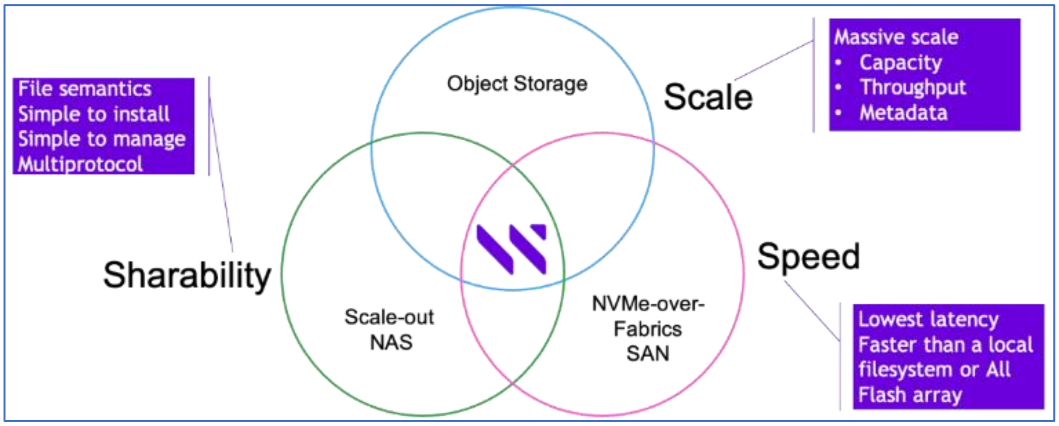


Figure 1 - Weka File System Structure

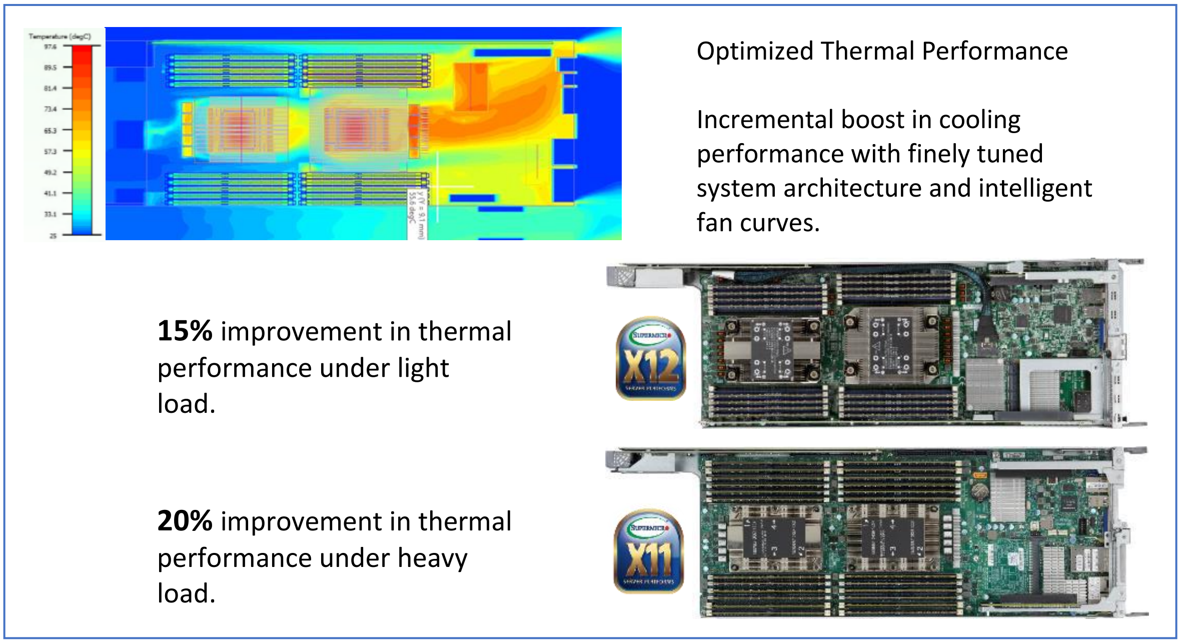
Weka's file system (WekaFS) is a fully distributed, parallel file system that was written entirely from scratch to deliver the highest performance file and object services by leveraging NVMe flash as its primary storage for persistent data across a wide range of applications. WekaFS will also, transparent to the application layer, seamlessly expand the filesystem namespace to include an extended layer built on any S3 compliant object storage system (see **Figure 3** and **Figure 4** for more details). There is no need for data migration software or complex scripts; all data resides in a single global namespace for easy access and management while maintaining the best performance. The intuitive graphical user interface allows a single administrator to quickly and easily manage hundreds of petabytes of data without any specialized storage training.

Weka's software delivers a more powerful and straightforward solution that would have traditionally required several disparate storage systems to leverage existing technologies in new ways and augment them with engineering innovations. The resulting software solution provides high performance for all workloads (big and small files, reads and writes, random, sequential, and metadata heavy). Furthermore, it is designed to run on a server infrastructure that does not rely on specialized hardware assist. As future hardware innovations come to market, WekaFS is well-positioned to leverage emerging technologies for the continued delivery of best cost and performance. The system can be expanded online to handle more demanding performance or store more capacity with no service interruption.

Configuration

In 2018, WekaIO set a number of benchmark records using the previous generation 2U 4-Node X11 BigTwin ® [1]. Now, in 2021, Supermicro has launched next-generation X12 BigTwin hardware, featuring 3rd Generation Intel® Xeon® Scalable Processors supporting up to 40 cores, higher instructions per clock, and two 512-bit FMA units. This is a huge advantage, as AVX-512 doubles the data registers compared to the AVX2 extension for the x86 instruction set. Beyond the computing power, Supermicro X12 BigTwin can now access data faster with twice the NVMe storage & I/O performance utilizing PCI-E 4.0, increasing WekaFS performance significantly over the previous generation of NVMe drives.

- There are two WEKA reference configurations to choose from, based on Supermicro's flagship green computing platform, X12 BigTwin. Both options offer top-tier performance with 6 or 12 NVMe PCI-E 4.0 storage drives, 256GB of memory, and 32 CPU cores per host.
- Both options offer a raw capacity of up to 7.7PB/rack or 367TB/system, using 15.3TB NVMe PCI-E 4.0 storage drives
  - ○ 2U 4-Node Capacity (21 systems x 6 drives x 4 nodes x 15.3TB = 7.7PB per 42U Rack)
  - ○ 2U 2-Node Capacity (21 systems x 12 drives x 2 nodes x 15.3TB = 7.7PB per 42U Rack)
- High-performance Kioxia NVMe PCI-E 4.0 drives with measured performance of 6.9 GB/s of throughput and 1.6 MIOPs vs. 3.3 GB/s of throughput and 800 KIOPs for NVMe PCI-E 3.0 drives.
- Key Advantages:
  - ○ 2U 4-Node X12 BigTwin offers up to 20% power efficiency than four traditional 1U servers and optimized cost models for entry-level to mid-sized deployments. Improved thermal performance over X11 BigTwin, as shown in **Figure 2**.
  - ○ 2U 2-Node X12 BigTwin offers optimal performance, storage, and operational advantages for mid-sized to hyperscale deployments.



Optimized Thermal Performance

Incremental boost in cooling performance with finely tuned system architecture and intelligent fan curves.

15% improvement in thermal performance under light load.

20% improvement in thermal performance under heavy load.

Figure 2 – Optimized Thermal Performance of SYS-220BT-HNTR over SYS-2029BT-HNR

An example of a validated high-density cluster solution using SYS-220BT-HNTR quad-node servers:

Type	Description	Per System	Per Cluster
System	SYS-220BT-HNTR X12 BigTwin 2U 4-Node, 6x U.2 NVMe PCI-E 4.0	1	2
CPU	3rd Gen Intel® Xeon® Scalable Processors 4314 16C/32T 2.4G 11.2GT 135W	8	16
Memory	16GB DDR4-3200 2Rx8 ECC Registered DIMM	64	128
Boot Controller	M.2 NVMe HW RAID Controller	4	8
Boot Drive	Toshiba XG6 512GB NVMe M.2 22x80 <1DWPD	8	16
Storage Drive	Kioxia CM6 7.68TB NVMe PCI-E 4.0 2.5" U.2 SSD	24	48
NIC1 & NIC2: Data Traffic	Mellanox ConnectX-6, LP Dual-port VPI HDR 200GbE, QSFP56, PCI-E 4.0	8	16
AIOM Slot: S3 Traffic	Broadcom BCM57414, OCP 3.0 Dual-port 25GbE, SFP28, PCI-E 4.0	4	8

Table 1 – 4U Cluster Specifications with 8 Nodes

Figure 3 shows how WekaFS can seamlessly expand the filesystem namespace on S3 compliant object storage via 25Gb network interfaces on each host. This 4U cluster includes eight nodes, each with a dual-port 25Gb NIC to offer sufficient bandwidth and redundancy for S3 APIs, such as GET, PUT, COPY, DELETE, POST, etc. For the data traffic, each node has two dual-port 200G HCAs to configure optimal performance and network redundancy.

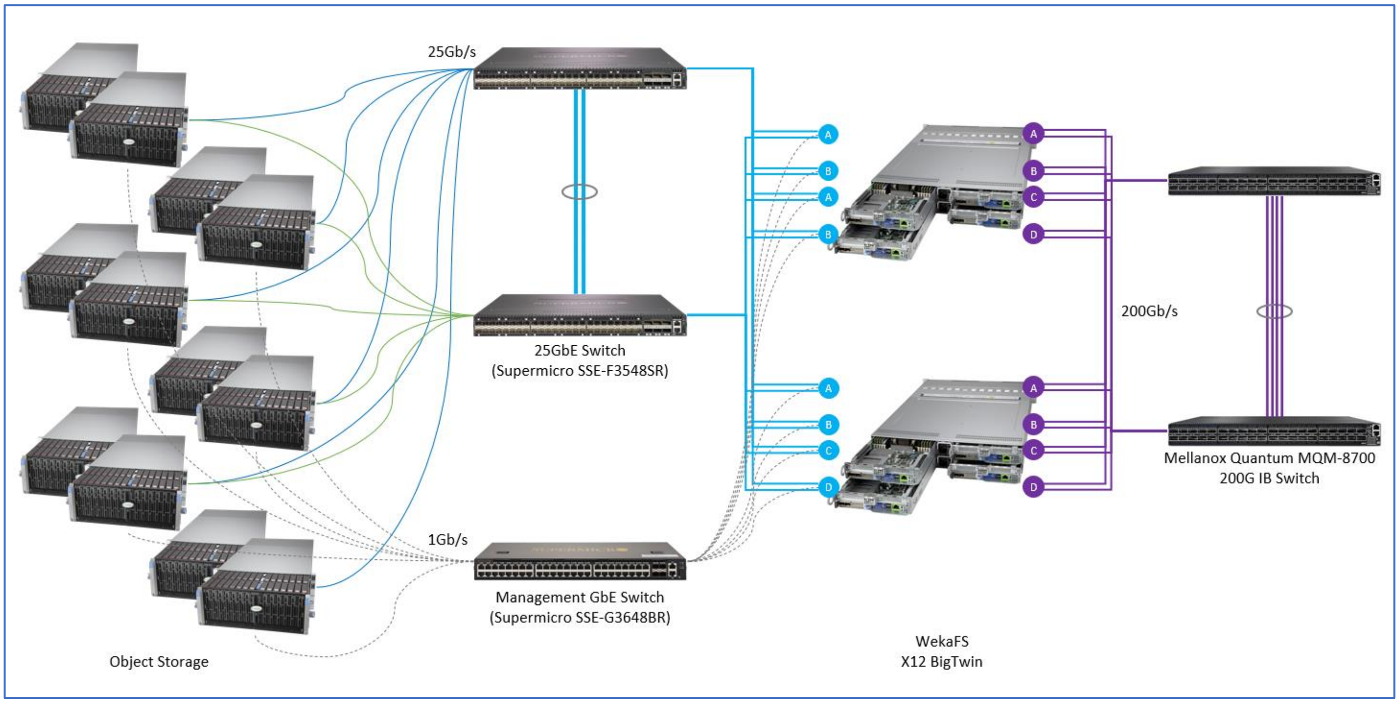


Figure 3 - Cluster Network Topology with WekaFS on SYS-220BT-HNTR

An example of a validated Twin-based cluster solution using SYS-220BT-DNTR dual-node servers:

Type	Description	Per System	Per Cluster
System	SYS-220BT-DNTR X12 BigTwin 2U 2-Node, 12x U.2 NVMe PCI-E 4.0	1	3
CPU	3rd Gen Intel® Xeon® Scalable Processors 6326 16C/32T 2.8G 11.2GT 185W	4	12
Memory	16GB DDR4-3200 2Rx8 ECC Registered DIMM	32	96
Optional Boot Controller	M.2 NVMe HW RAID Controller	2	6
Boot Drive	Toshiba XG6 512GB NVMe M.2 22x80 <1DWPD	4	12
Storage Drive	Kioxia CM6 7.68TB NVMe PCI-E 4.0 2.5" U.2 SSD	24	72
NIC1 for S3 Traffic	Broadcom BCM57414, LP Dual-port 25GbE, SFP28, PCI-E 3.0	2	6
NIC2 for Data Traffic	Mellanox ConnectX-6, LP Dual-port VPI HDR 200GbE, QSFP56, PCI-E 4.0	2	6
AIOM for Data Traffic	Mellanox ConnectX-6, OCP 3.0 Dual-port VPI HDR 200GbE, QSFP56, PCI-E 4.0	2	6

Table 2 – 6U Cluster Specifications with 6 Nodes

Figure 4 shows how WekaFS can seamlessly expand the filesystem namespace on S3 compliant object storage via a 25Gb network on each host. This 6U cluster includes six nodes, each with a dual-port 25Gb NIC to offer sufficient bandwidth and redundancy for S3 APIs, such as GET, PUT, COPY, DELETE, POST, etc. For the data traffic, each node has two dual-port 200G HCAs to configure optimal performance and network redundancy.

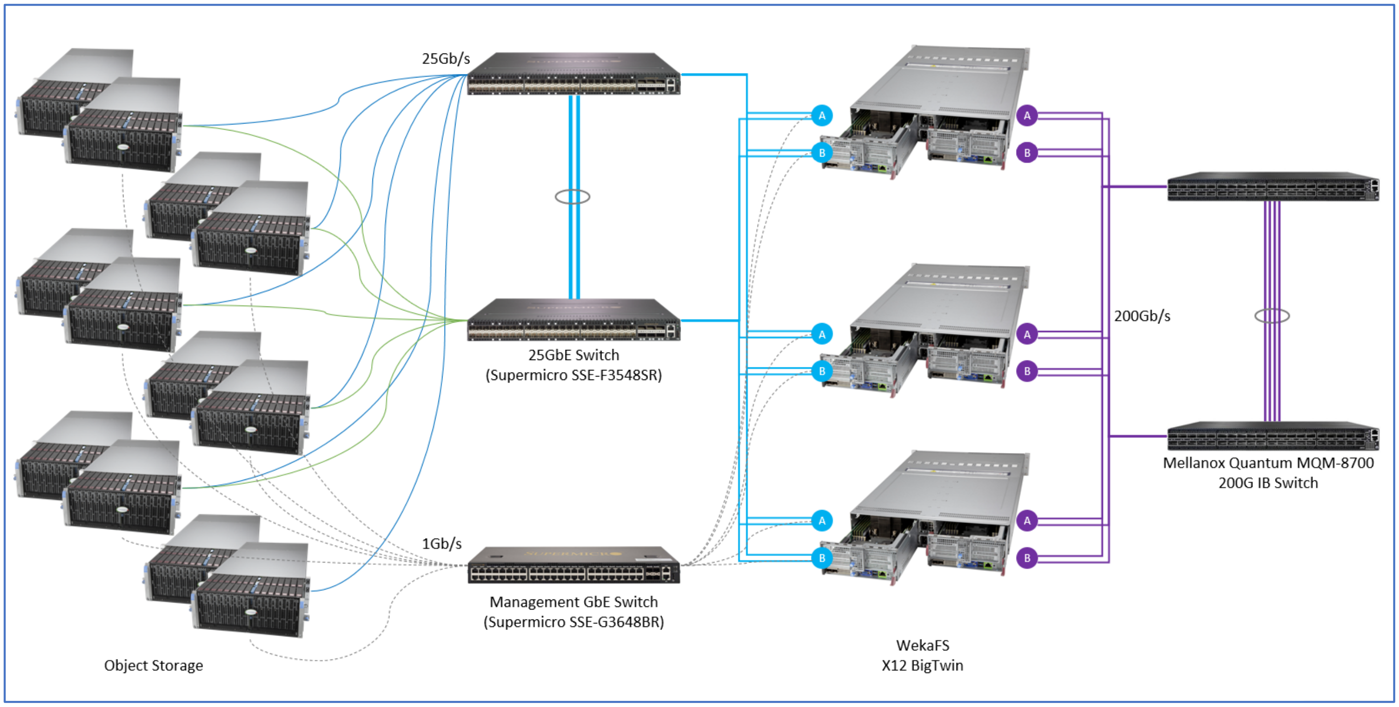


Figure 4 - Cluster Network Topology with WekaFS on SYS-220BT-DNTR

Quality, Serviceability, and Remote Management

Since Supermicro launched its Intel®-based Twin system architecture in 2007, our hardware and firmware design teams have built a variety of Enterprise features into the Twin Product Family, along with long-standing partners who continue to develop purpose-built appliances and private cloud infrastructure with X12 BigTwin. Critical use cases include hyperconverged infrastructure, scale-out object storage, scale-out block storage, and scale-out file systems. Through these significant partnerships and successes in the Enterprise server market, the design team has been able to go beyond just optimizing performance and put considerable emphasis on building quality, serviceability, and remote management.

The X12 2U 2-Node BigTwin features built-in redundancies for power, PMBus, NVMe management, and M.2 boot drives, shown in **Figure 5**. Supermicro was one of the 1st server manufacturers to support NVMe technologies and has developed advanced capabilities for power controls, re-drivers, and re-timers. On SYS-220BT-DNTR, 12 NVMe PCI-E 4.0 drives deliver balanced performance with direct connections to the dual-processors on each node, as shown in **Figure 6**.

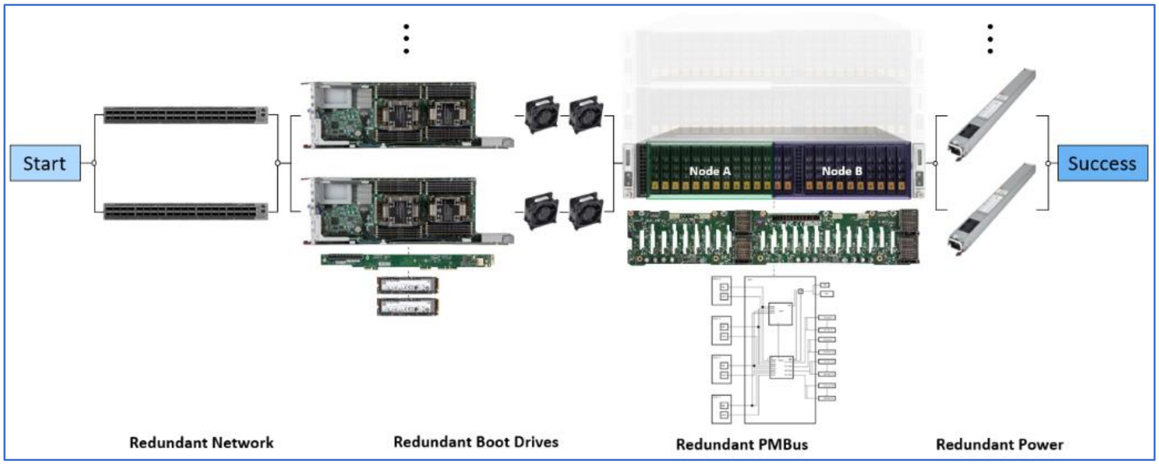


Figure 5 – 2U 2-Node System Reliability Diagram

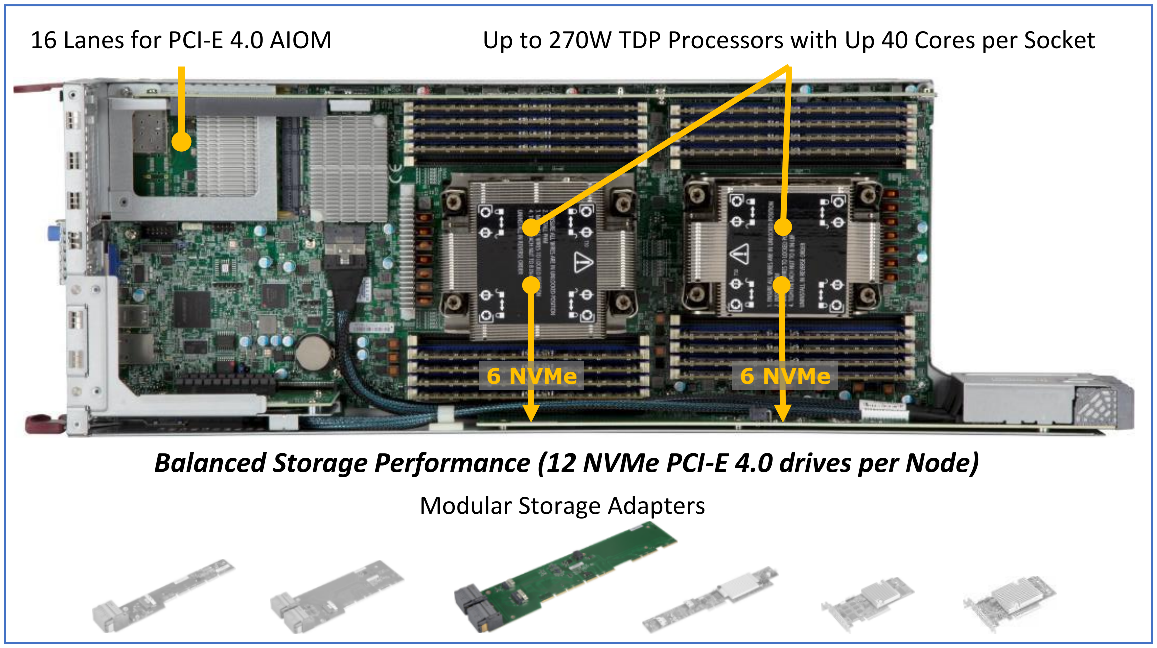
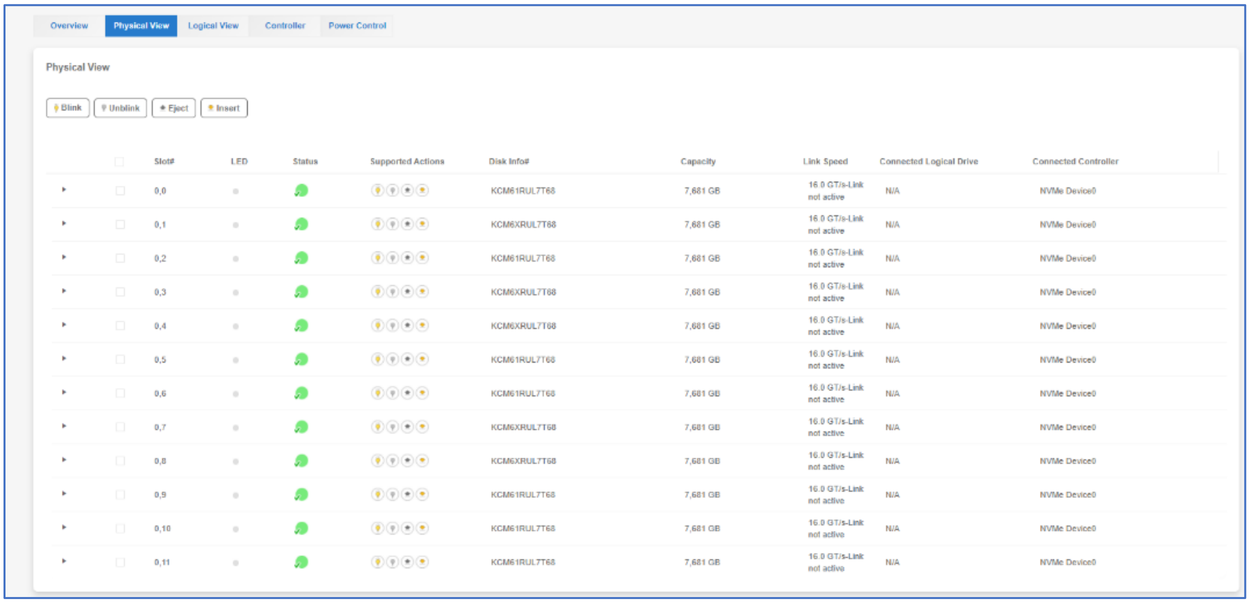

16 Lanes for PCI-E 4.0 AIOM
Up to 270W TDP Processors with Up 40 Cores per Socket
6 NVMe
6 NVMe
Balanced Storage Performance (12 NVMe PCI-E 4.0 drives per Node)
Modular Storage Adapters

Figure 6 – Balanced NVMe Performance on SYS-220BT-DNTR

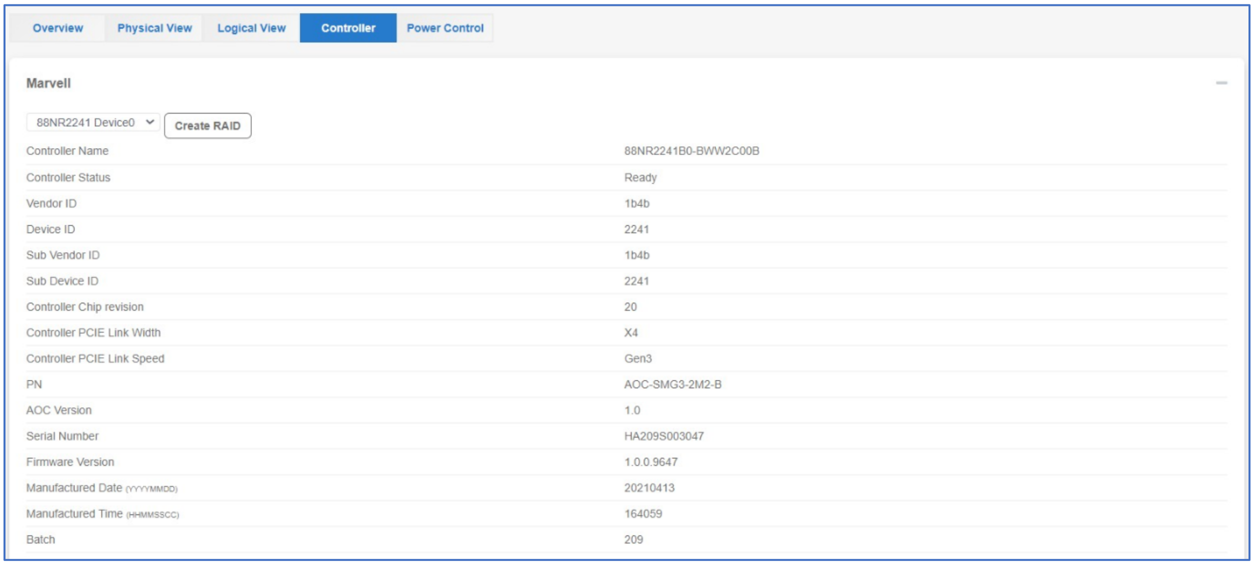
Not only do the direct NVMe connections help to deliver unrivaled storage performance, but they ensure strong signal integrity and the reliability to manage the NVMe drives through the BMC's web management interface, shown in **Figure 7**.



		Slot#	LED	Status	Supported Actions	Disk Info#	Capacity	Link Speed	Connected Logical Drive	Connected Controller
▸	□	0,0	●	●		KCM61RUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,1	●	●		KCM6XRUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,2	●	●		KCM61RUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,3	●	●		KCM6XRUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,4	●	●		KCM6XRUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,5	●	●		KCM61RUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,6	●	●		KCM61RUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,7	●	●		KCM6XRUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,8	●	●		KCM6XRUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,9	●	●		KCM61RUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,10	●	●		KCM61RUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0
▸	□	0,11	●	●		KCM61RUL7T68	7,681 GB	16.0 GT/s-Link not active	N/A	NVMe Device0

Figure 7 – BMC Web UI View of Physical NVMe PCI-E 4.0 Drives

For both WEKA reference configurations, each node features a redundant boot controller for two M.2 NVMe drives, managed via a dedicated sub-panel under Storage Monitoring in **Figure 8.** The controller may be managed from the BIOS with HII (Human Interface Infrastructure) support and Redfish APIs to integrate with infrastructure orchestration tools.

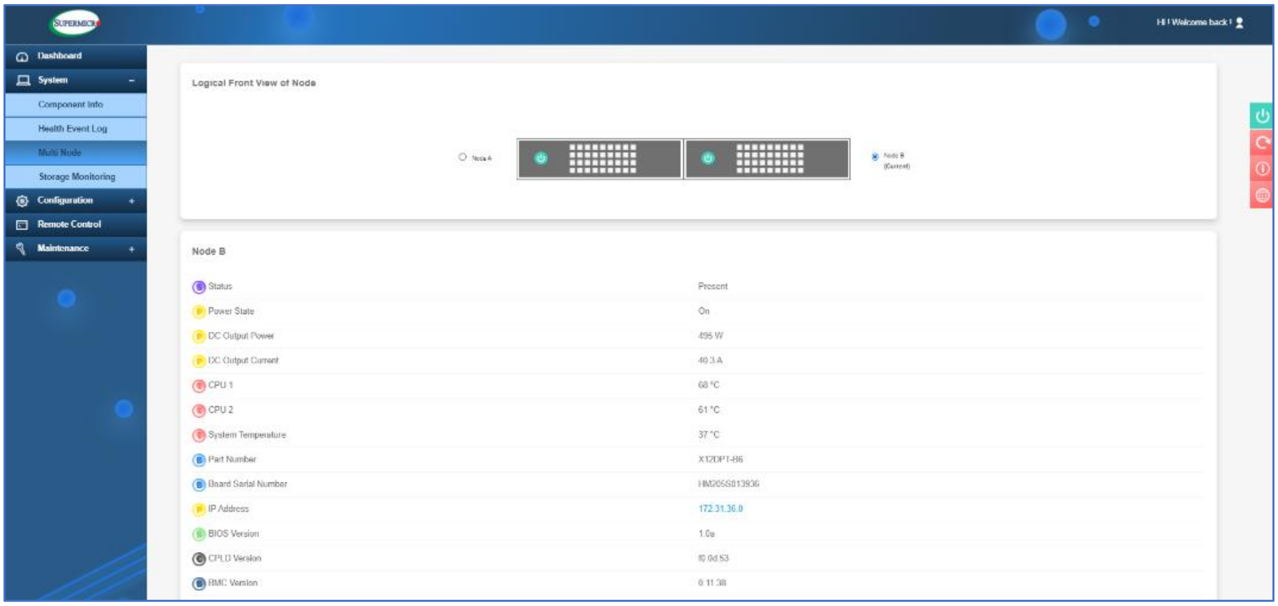


Controller Name	88NR2241B0-BWW2C00B
Controller Status	Ready
Vendor ID	1b4b
Device ID	2241
Sub Vendor ID	1b4b
Sub Device ID	2241
Controller Chip revision	20
Controller PCIE Link Width	X4
Controller PCIE Link Speed	Gen3
PN	AOC-SMG3-2M2-B
AOC Version	1.0
Serial Number	HA209S003047
Firmware Version	1.0.0.9647
Manufactured Date (YYYYMMDD)	20210413
Manufactured Time (HHMMSSCC)	164059
Batch	209

Figure 8 – BMC Web UI View of M.2 NVMe Boot Controller

With the new BMC web interface on X12 BigTwin, system administrators can quickly locate and identify each node in each 2U enclosure to streamline the deployment of WekaFS. **Figure 9** shows the Logical Front View of 2 nodes on SYS-220BT-

DNTR, which can also help system administrators communicate more effectively with field technicians managing the infrastructure.



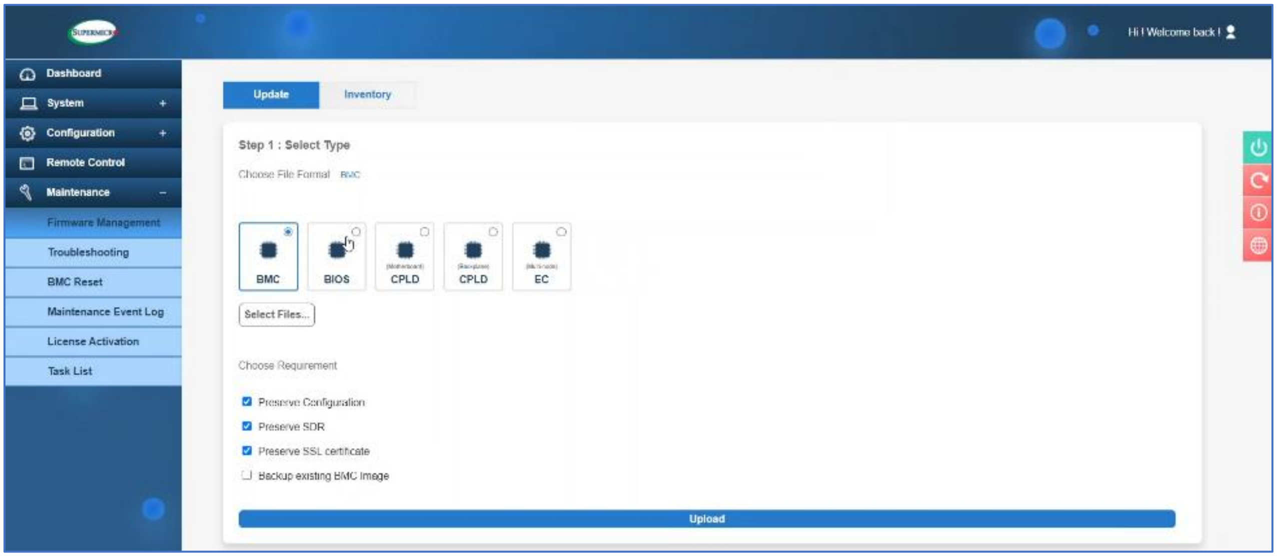
Logical Front View of Node

Node B

Status	Present
Power State	On
DC Output Power	496 W
DC Output Current	40.3 A
CPU 1	68 °C
CPU 2	61 °C
System Temperature	37 °C
Part Number	X12DPT-B6
Board Serial Number	HM205S013936
IP Address	172.31.36.0
BIOS Version	1.0a
CPLD Version	f0.0d.53
BMC Version	0.11.38

Figure 9 – BMC Web UI for Multi-Node Logical View

The backplane’s CPLD can help recover any stalled bus to ensure the backplane’s Embedded Controller (EC) can reliably manage firmware updates through the dedicated Firmware Management panel, shown in **Figure 10**.



Step 1 : Select Type

Choose File Format BMC

Choose Requirement

☑ Preserve Configuration

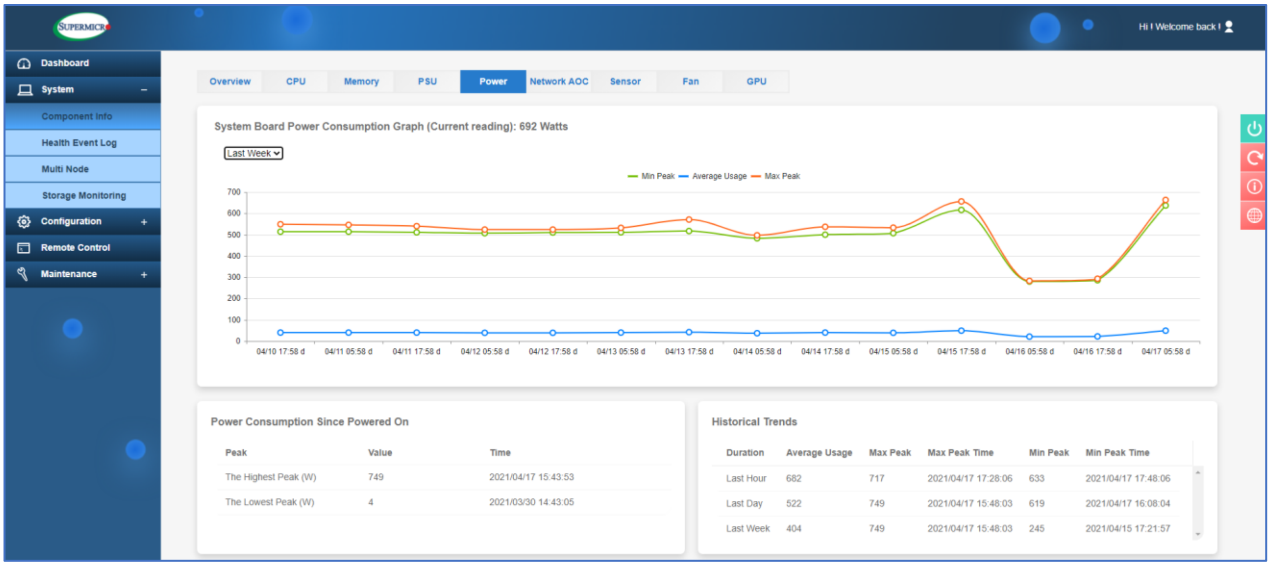
☑ Preserve SDR

☑ Preserve SSL certificate

☐ Backup existing BMC Image

Figure 10 – BMC Web UI for Firmware Management

X12 2U 2-Node BigTwin includes 2200W Redundant Power Supplies with Titanium Level 96% Power Efficiency. This is shared between the two nodes and offers about ~10% power efficiency advantage over two standard 1U servers with 12 NVMe drives. This helps to reduce e-waste by approximately 20% with its shared power, cooling system, and backplane design. The system supports Smart Ride Through (SmarT) Power to 'ride through' a momentary loss of AC power while maintaining the highest possible power supply efficiency, which can be monitored through the Power sub-panel shown in **Figure 11.** In the event of a failure, each hot-swap node is easily accessible, making service calls a breeze, e.g., swap out a memory DIMM or installing a standard low-profile card without any tools, as shown in **Figure 12.**



System Board Power Consumption Graph (Current reading): 692 Watts

Power Consumption Since Powered On

Peak	Value	Time
The Highest Peak (W)	749	2021/04/17 15:43:53
The Lowest Peak (W)	4	2021/03/30 14:43:05

Historical Trends

Duration	Average Usage	Max Peak	Max Peak Time	Min Peak	Min Peak Time
Last Hour	682	717	2021/04/17 17:28:06	633	2021/04/17 17:48:06
Last Day	522	749	2021/04/17 15:48:03	619	2021/04/17 16:08:04
Last Week	404	749	2021/04/17 15:48:03	245	2021/04/15 17:21:57

Figure 11 – BMC Web UI for Power Monitoring

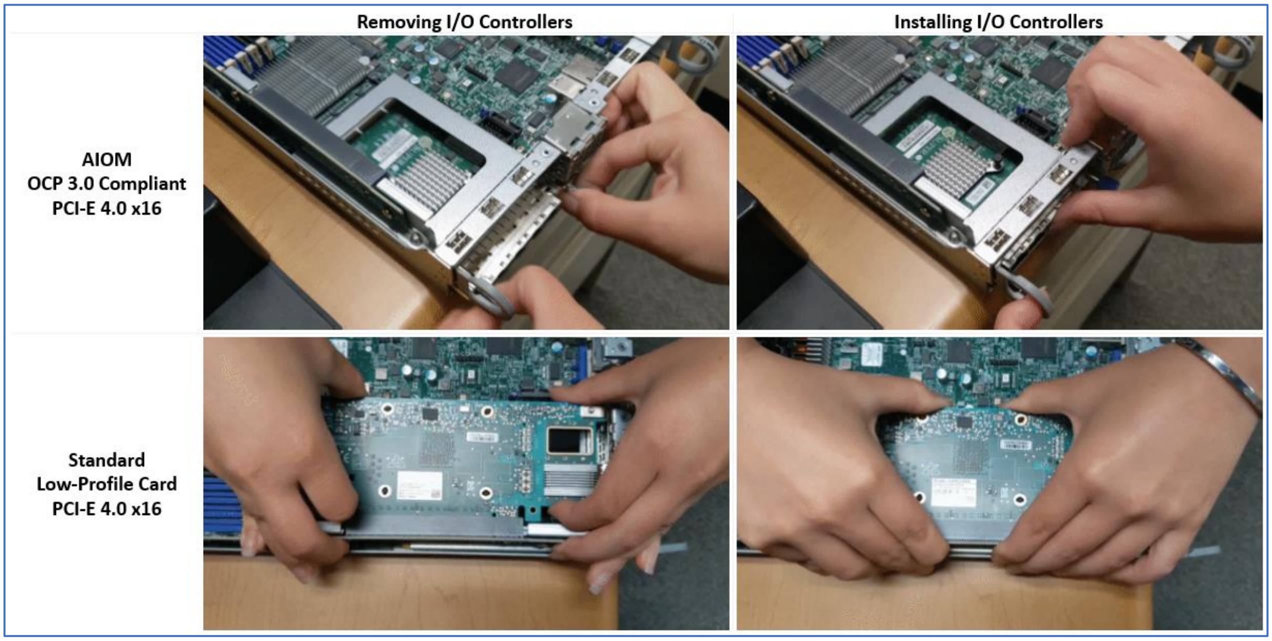
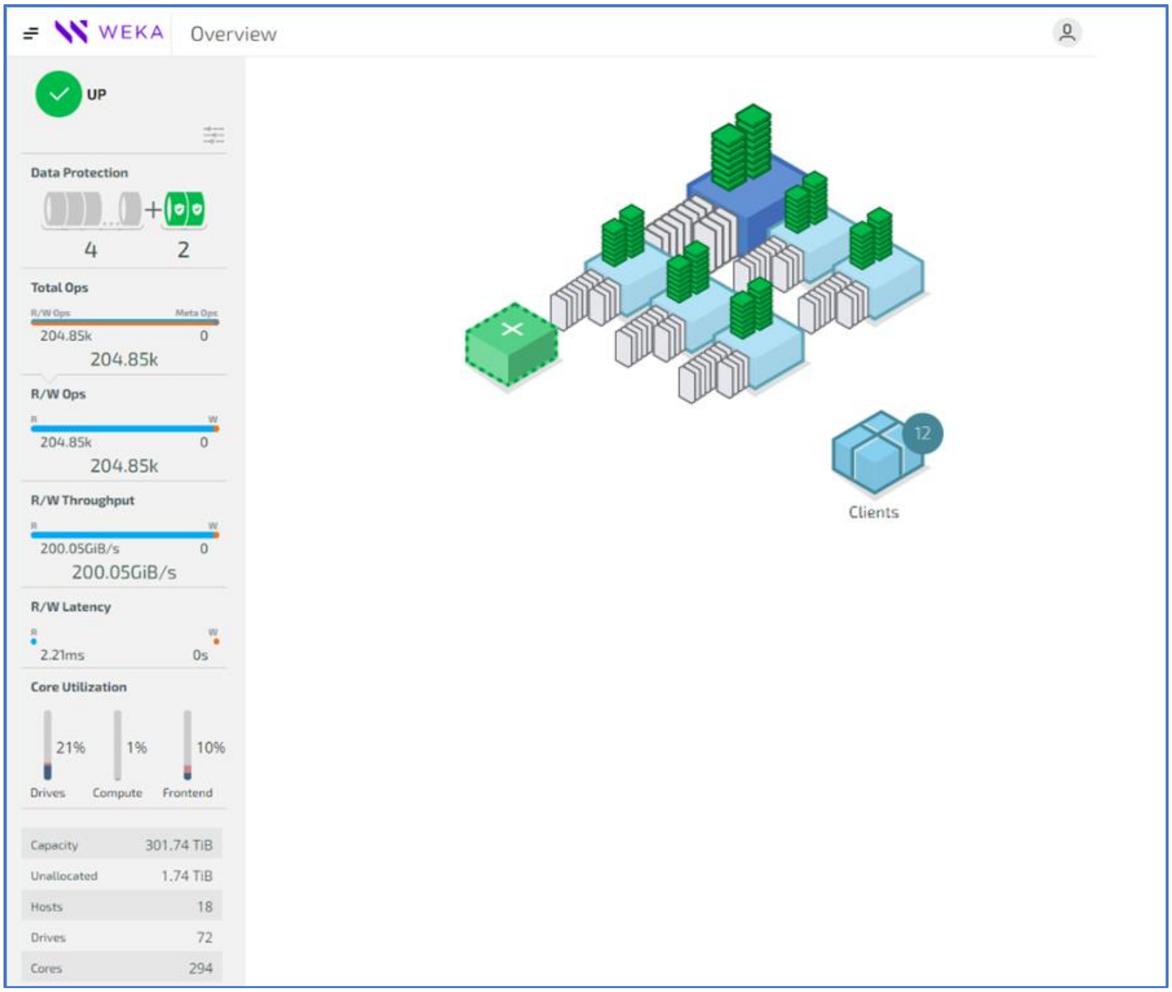


Figure 12 – Tool-less PCI-E Slots

WEKA Performance Overview

Using FIO IO generators on 12 clients, a massive performance of ~202GB/s throughput (33.6GB/s per host) and 8.5 million IOPS (1.4m IOPS per host) was measured on just three X12 BigTwin systems (six hosts), each with two 3rd Gen Intel® Xeon® Scalable Processors, 2x 200Gb/s ConnectX-6 PCI-E 4.0 host channel adapters, 12 Kioxia CM6 NVMe PCI-E 4.0 drives, and

WekaIO (3.11); configured with 49 Weka system cores and 3 containers per host. These results were limited by the number of clients in the test. The Weka storage cluster is capable of significantly more performance in this configuration but will require more clients to drive that performance. The eight host X12 BigTwin cluster, with two SYS-220BT-HNTR systems, is estimated to have similar performance to the six host cluster with three SYS-220BT-DNTR systems. With the same number of clients, the six host cluster will have slightly more throughput and slightly fewer IOPS at the limit when more clients are added.



Data Protection
4 + 2

Total Ops
R/W Ops 204.85k — Meta Ops 0
204.85k

R/W Ops
R 204.85k — W 0
204.85k

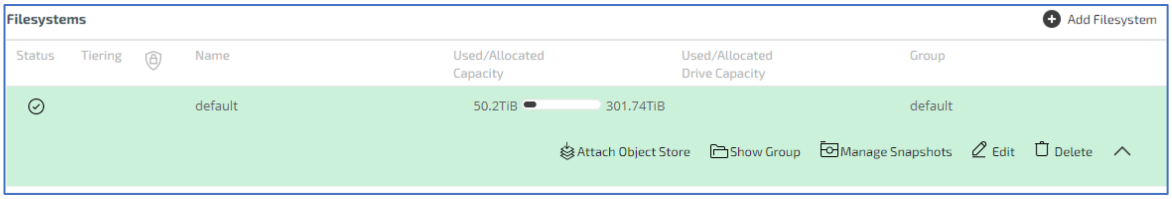
R/W Throughput
R 200.05GiB/s — W 0
200.05GiB/s

R/W Latency
R 2.21ms — W 0s

Core Utilization
Drives 21% — Compute 1% — Frontend 10%

Capacity	301.74 TiB
Unallocated	1.74 TiB
Hosts	18
Drives	72
Cores	294

Figure 13 – Weka Performance Overview



Status	Tiering		Name	Used/Allocated Capacity	Used/Allocated Drive Capacity	Group
✓			default	50.2TiB 301.74TiB		default

Figure 14 – File System with 331.76TB Usable Storage

WEKA Validated X12 Supermicro Servers

SUPERMICRO SERVER SYS-220BT-HNTR

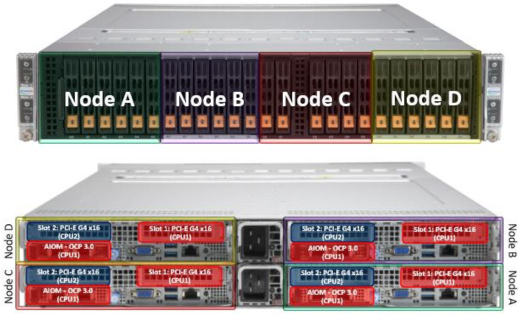
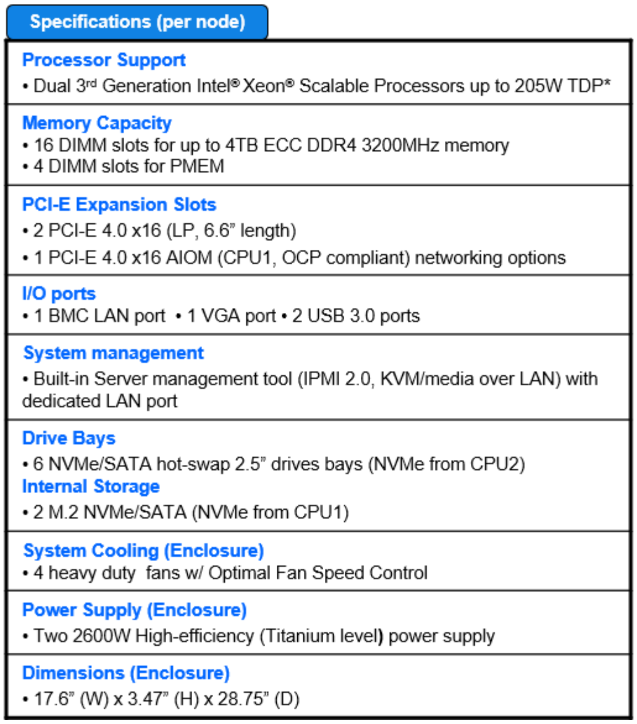

Node A
Node B
Node C
Node D
Node D
Slot 2: PCI-E G4 x16 (CPU2)
Slot 1: PCI-E G4 x16 (CPU1)
AIOM - OCP 3.0 (CPU1)
Node C
Slot 2: PCI-E G4 x16 (CPU2)
Slot 1: PCI-E G4 x16 (CPU1)
AIOM - OCP 3.0 (CPU1)
Slot 2: PCI-E G4 x16 (CPU2)
Slot 1: PCI-E G4 x16 (CPU1)
AIOM - OCP 3.0 (CPU1)
Node B
Slot 2: PCI-E G4 x16 (CPU2)
Slot 1: PCI-E G4 x16 (CPU1)
AIOM - OCP 3.0 (CPU1)
Node A

Figure 15 - 2U 4-Node BigTwin Server

Recommended Scale:

Entry-level to Mid-sized Deployments,
Requiring a minimum of 4U rack space



Specifications (per node)

Processor Support
- Dual 3rd Generation Intel® Xeon® Scalable Processors up to 205W TDP*

Memory Capacity
- 16 DIMM slots for up to 4TB ECC DDR4 3200MHz memory
- 4 DIMM slots for PMEM

PCI-E Expansion Slots
- 2 PCI-E 4.0 x16 (LP, 6.6" length)
- 1 PCI-E 4.0 x16 AIOM (CPU1, OCP compliant) networking options

I/O ports
- 1 BMC LAN port • 1 VGA port • 2 USB 3.0 ports

System management
- Built-in Server management tool (IPMI 2.0, KVM/media over LAN) with dedicated LAN port

Drive Bays
- 6 NVMe/SATA hot-swap 2.5" drives bays (NVMe from CPU2)

Internal Storage
- 2 M.2 NVMe/SATA (NVMe from CPU1)

System Cooling (Enclosure)
- 4 heavy duty fans w/ Optimal Fan Speed Control

Power Supply (Enclosure)
- Two 2600W High-efficiency (Titanium level) power supply

Dimensions (Enclosure)
- 17.6" (W) x 3.47" (H) x 28.75" (D)

SUPERMICRO SERVER SYS-220BT-DNTR

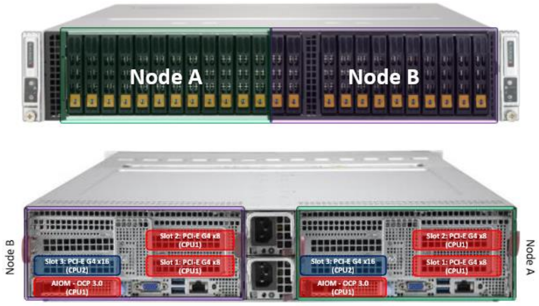
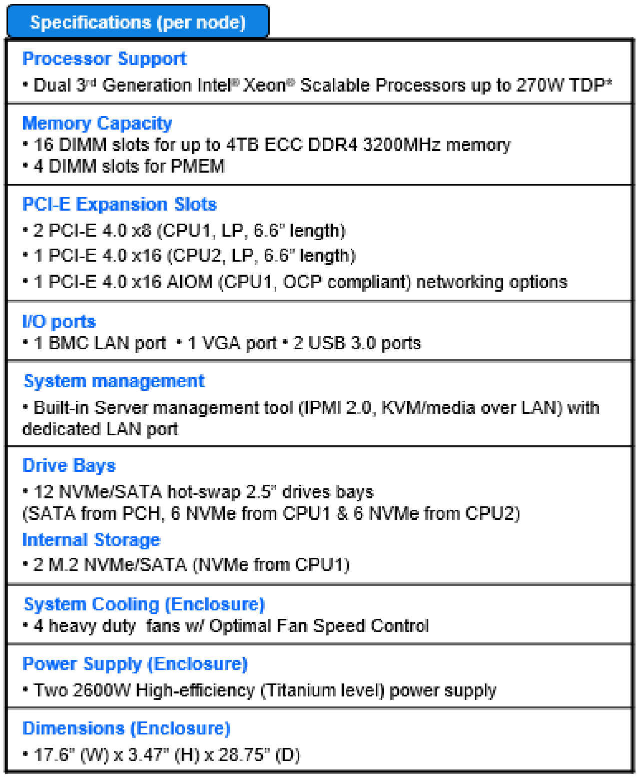

Node A
Node B
Node B
Slot 2: PCI-E G4 x8 (CPU1)
Slot 3: PCI-E G4 x16 (CPU2)
Slot 1: PCI-E G4 x8 (CPU1)
AIOM - OCP 3.0 (CPU1)
Slot 2: PCI-E G4 x8 (CPU1)
Slot 3: PCI-E G4 x16 (CPU2)
Slot 1: PCI-E G4 x8 (CPU1)
AIOM - OCP 3.0 (CPU1)
Node A

Figure 16 - 2U 2-Node BigTwin Server

Recommended Scale:

Mid-sized to Hyperscale Deployments
Requiring a minimum of 6U rack space



Specifications (per node)

Processor Support
- Dual 3rd Generation Intel® Xeon® Scalable Processors up to 270W TDP*

Memory Capacity
- 16 DIMM slots for up to 4TB ECC DDR4 3200MHz memory
- 4 DIMM slots for PMEM

PCI-E Expansion Slots
- 2 PCI-E 4.0 x8 (CPU1, LP, 6.6" length)
- 1 PCI-E 4.0 x16 (CPU2, LP, 6.6" length)
- 1 PCI-E 4.0 x16 AIOM (CPU1, OCP compliant) networking options

I/O ports
- 1 BMC LAN port • 1 VGA port • 2 USB 3.0 ports

System management
- Built-in Server management tool (IPMI 2.0, KVM/media over LAN) with dedicated LAN port

Drive Bays
- 12 NVMe/SATA hot-swap 2.5" drives bays
(SATA from PCH, 6 NVMe from CPU1 & 6 NVMe from CPU2)

Internal Storage
- 2 M.2 NVMe/SATA (NVMe from CPU1)

System Cooling (Enclosure)
- 4 heavy duty fans w/ Optimal Fan Speed Control

Power Supply (Enclosure)
- Two 2600W High-efficiency (Titanium level) power supply

Dimensions (Enclosure)
- 17.6" (W) x 3.47" (H) x 28.75" (D)

Summary

Dramatic improvements in computational power and exascale needs for storage in today's digital mediums have meant that typical file systems traditionally used to address complex workloads are often impractical or inadequate to the task. WekaIO combined with Supermicro servers provides a stunning performance, protection, and data management story for Deep Learning, High-Performance Compute, and high-throughput low-latency storage workloads. WekaIO removes your computational storage bottlenecks by leveraging the power of NVMe and task-optimized servers, along with software designed for performance, scalability, and flexibility.

The combination of Supermicro SuperSevers and WekaIO software provides customers with solutions that can leverage our building-block architecture to provide the most optimized CAPEX and OPEX. With Supermicro's professional services, our Rack Integration Team can fully rack, integrate, pre-test and tune, allowing you to be operational less than 30 minutes after receiving.

Additional Resources

For more information, please visit:
WEKAIO - https://www.weka.io/how-it-works/
Supermicro Servers - https://www.supermicro.com/en/products/bigtwin/
Contact: total_solutions@supermicro.com

Reference(s):
1 - Carol Platz, “WekaIO Delivers Record-Breaking Performance Results on SPEC 2014, Waters Communications