WHITE PAPER

HIGHLY AVAILABLE SUPERSERVER SSG-221E-DN2R24R PROTECTED BY XIRAID

Supermicro SuperServer SSG-221E-DN2R24R

TABLE OF CONTENTS

Executive Summary

Supermicro, a global leader in application-optimized total IT solutions, and Xinnor, a software development company specializing in high-performance storage, have teamed up to develop a highly reliant and fast storage solution by combining Supermicro SuperServer SSG-221E-DN2R24R with Xinnor xiRAID Classic software RAID, supporting high availability between two server nodes. This solution enables building an extremely fast cluster-in-a-box, capable of surviving multiple drive failures and a complete server node failure without data loss or downtime. The solution achieves unprecedented resiliency at the extreme performance levels of PCIe Gen5 NVMe drives.

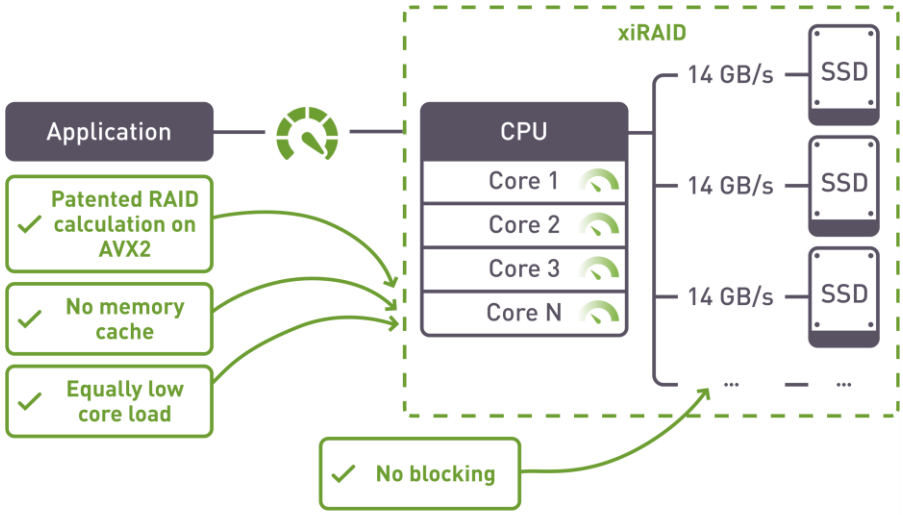
xiRAID Classic is a software RAID engine specifically designed to handle the high level of parallelism in NVMe SSDs. Its innovative data path exploits the full potential of NVMe drives while offering superior resiliency. xiRAID can be highly available when paired with the Supermicro SuperServer SSG-221E-DN2R24R.

The Supermicro SuperServer SSG-221E-DN2R24R features two independent server nodes that access 24 shared NVMe drives. In case of node failure, xiRAID can fail-over and fail-back from one server node to the other, ensuring no impact on data availability.

The solution offered by Supermicro SuperServer SSG-221E-DN2R24R and Xinnor's xiRAID Classic achieves up to 226GB/s sequential read, over 64GB/s sequential write, 12M IOPS in random read, and over 1.5 M random write with 24 Kioxia CM7-R Series drives. These drives are read-optimized, meaning write performance can be further increased when deployed by Kioxia CM7-V or other equivalent mixed-use drives. These performance metrics were achieved by implementing three groups of 8 namespaces in RAID level 6 on each server node, protecting against up to two logical drive failures in every RAID group.

About xiRAID

With their exceptional performance, NVMe PCIe Gen5 drives are ideal for mission-critical applications like databases and fast storage for Artificial Intelligence computational systems. In these applications, performance and hardware resiliency are equally crucial. At the same time, RAID remains the de-facto standard technology for drive failure protection; traditional RAID implementations were designed for slower SATA and SAS storage media. With the adoption of NVMe drives, traditional hardware RAID becomes the bottleneck, as it operates via a 16-lane PCIe bus. Since each NVMe drive uses 4 PCIe lanes, hardware RAID can only address four drives at full speed. Xinnor developed xiRAID, an innovative software RAID engine specifically designed for NVMe drives, to address this limitation.

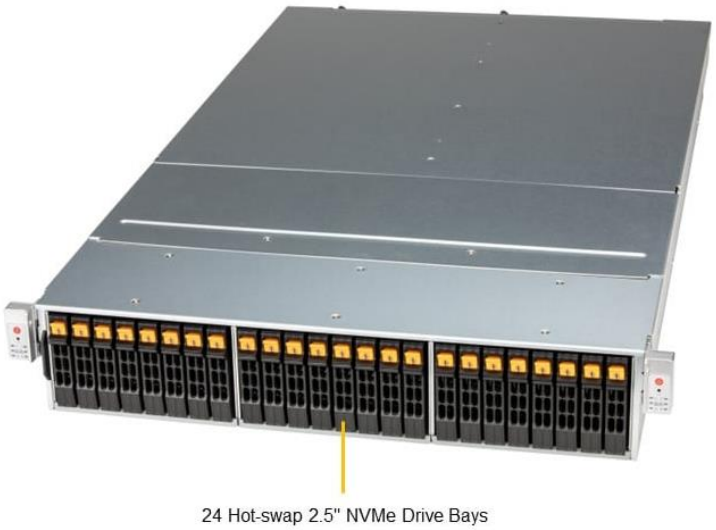


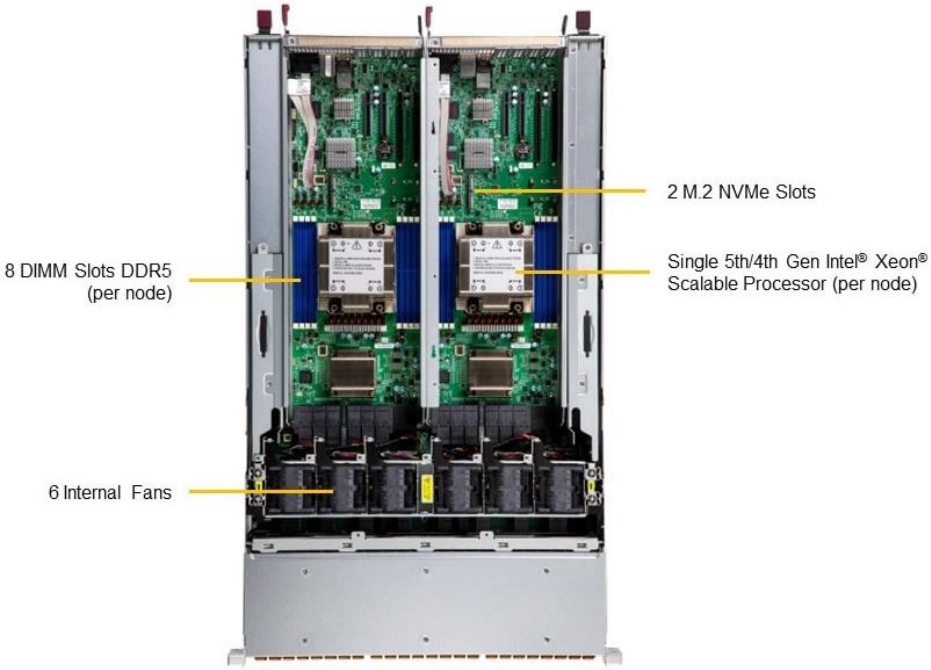
By combining efficient checksum calculation in x86 CPU's AVX (Advanced Vector Extensions) technology with its lockless data path, xiRAID achieves near-raw drive performance with minimal system resource utilization. The I/O handling parallelization and lockless datapath minimize RAID performance overhead, delivering speeds close to raw hardware capabilities.

xiRAID supports multiple RAID levels, including RAID 0, 1, 5, 6, RAID 7.3 (3 drives parity), all nested RAIDs (10, 50, 60, 70), and N+M. This flexibility allows customers to select their preferred protection level, based on specific application requirements.

Supermicro SuperServer

The Supermicro SuperServer comprises two independent server nodes sharing 24 NVMe PCIe Gen5 drives in a 2U form factor. This Storage Bridge Bay (SBB) server is ideal for implementing high-availability NVMe drives.





Each node supports:

1. Single Socket E (LGA-4677) 5th/4th Generation Intel® Xeon® Scalable processor, up to 350W TDP
2. 2 PCIe 5.0 x16 HHHL slots
3. 2 PCIe 5.0 x8 HHHL slots
4. 8 DIMMs per node with 1DPC, up to 2TB memory capacity with 8 DIMMs of 256GB 3DS RDIMM DDR5-5600 ECC memory per node
5. Dedicated 1GbE Private Ethernet connection for node-to-node communication (Heartbeat)
6. Redundant Titanium 2000W Power Supplies

High Availability

The Supermicro SuperServer with Xinnor's xiRAID supports high availability between the two server nodes. Multiple RAID groups can be created on each server node using xiRAID. If one server node fails, xiRAID Classic, implemented with PaceMaker and Corosync, automatically fails over its RAID groups to the other node. When the failed node is restored to operation, its RAID groups fail back from the second node. In the fail-over process, the applications running on the RAID groups created by xiRAID are also migrated to the surviving node, assuring service continuity.

This solution can survive multiple drive failures and a complete node failure without data loss and minimal downtime.

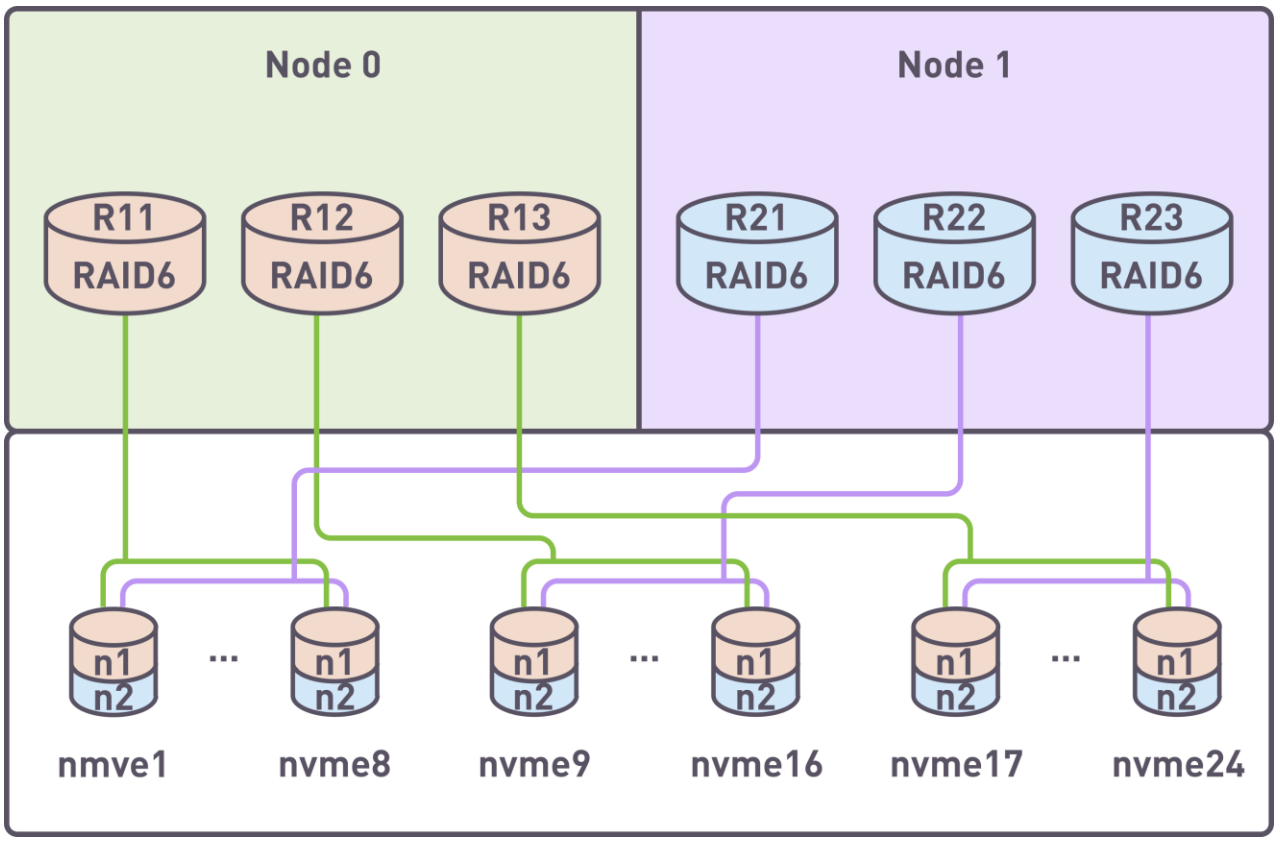
PaceMaker and Corosync Setup

Pacemaker and Corosync HA cluster setup requires the following configurations:

- Using the Supermicro SuperServer SSG-221E-DN2R24R's internal Ethernet network interconnect as the cluster heartbeat link. A second heartbeat can be configured on an external node-to-node dedicated Ethernet link (not used in this solution).
- Configuring STONITH properly in the cluster. SuperServer SSG-221E-DN2R24R nodes feature BMC IPMI, compatible with the `fence_ipmilan` Pacemaker fencing agent. Each SBB node's BMC IPMI must be network-accessible from the other node.
- Synchronizing system time on both nodes using properly configured `chrony` or `xntpd` daemons.
- xiRAID Classic Pacemaker agent requires additional steps for configuration. These steps are detailed in the "Integrating xiRAID Classic Into a Pacemaker Cluster" document.

Performance Results on Read Intensive Drives

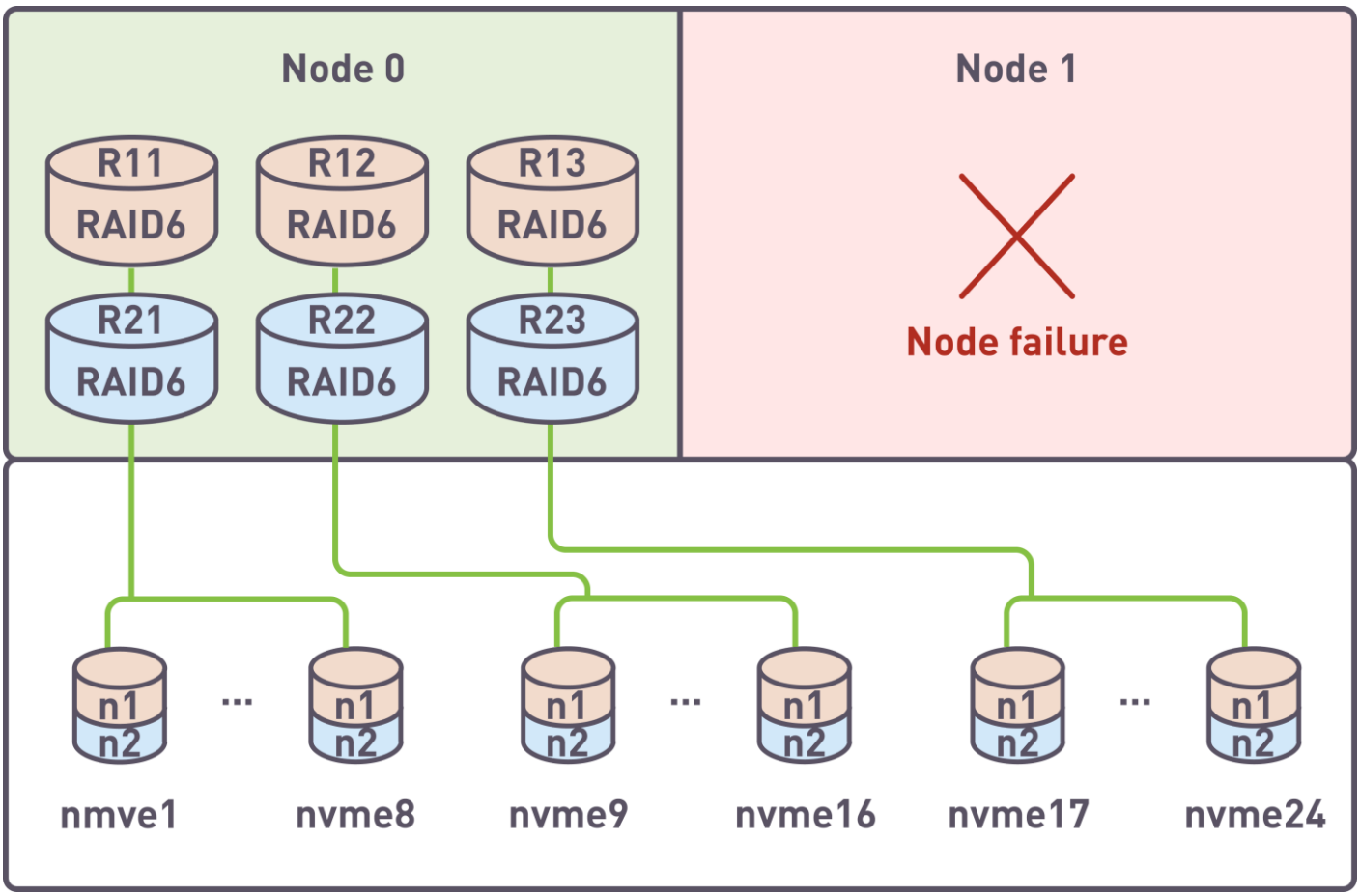
The system was equipped with 24 KIOXIA CM7-R KCMYXRUG1T92 drives. To fully utilize the drive PCIe bus connections, each drive was split into two namespaces. The first namespaces were used for RAIDs active on Node0, while the second namespaces were used for RAIDs active on Node1. To fully utilize the drive performance, we created a symmetrical configuration of relatively small RAIDs to be close to most real-life configurations. As a result, 3 RAIDs of 8 namespaces each were created at each node. The following picture and table showcase the namespace mapping.



Node0			Node1		
RAID	**RAID parameters**	**Drives**	**RAID**	**RAID parameters**	**Drives**
R11	RAID level 6 Strip* size 128K	/dev/nvme1n1 /dev/nvme2n1 /dev/nvme3n1 /dev/nvme4n1 /dev/nvme5n1 /dev/nvme6n1 /dev/nvme7n1 /dev/nvme8n1	R21	RAID level 6 strip size 128K	/dev/nvme1n2 /dev/nvme2n2 /dev/nvme3n2 /dev/nvme4n2 /dev/nvme5n2 /dev/nvme6n2 /dev/nvme7n2 /dev/nvme8n2
R12	RAID level 6 strip size 128K	/dev/nvme9n1 /dev/nvme10n1 /dev/nvme11n1 /dev/nvme12n1 /dev/nvme13n1 /dev/nvme14n1 /dev/nvme15n1 /dev/nvme16n1	R22	RAID level 6 strip size 128K	/dev/nvme9n2 /dev/nvme10n2 /dev/nvme11n2 /dev/nvme12n2 /dev/nvme13n2 /dev/nvme14n2 /dev/nvme15n2 /dev/nvme16n2
R13	RAID level 6 strip size 128K	/dev/nvme17n1 /dev/nvme18n1 /dev/nvme19n1 /dev/nvme20n1 /dev/nvme21n1 /dev/nvme22n1 /dev/nvme23n1 /dev/nvme24n1	R23	RAID level 6 strip size 128K	/dev/nvme17n2 /dev/nvme18n2 /dev/nvme19n2 /dev/nvme20n2 /dev/nvme21n2 /dev/nvme22n2 /dev/nvme23n2 /dev/nvme24n2

The “strip size” (also known as “chunk size”) is the minimum amount of data that is written on one single drive within the RAID.

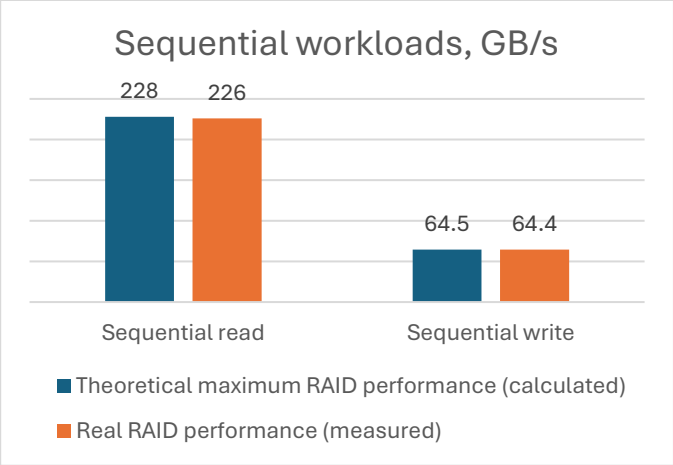
In case of failure of one of the two nodes, the RAID groups on the failed node are automatically migrated to the surviving one, as explained in the image below.

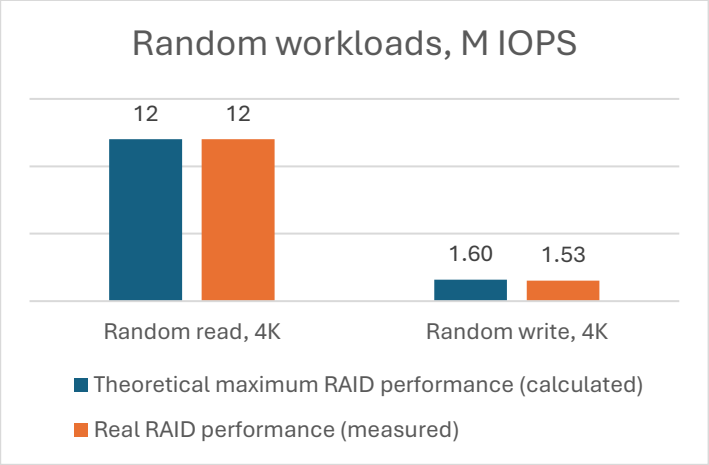


The following table reports the results obtained during simultaneous testing of both nodes' RAID groups or NVMes. Tests on RAID groups were conducted with 32 threads/64 queue depth per RAID group.

Expected RAID performance is based on RAID architecture:

- Sequential read performance should approximate the combined sequential read performance of all the drives within the RAID groups
- Sequential write performance should approximate the combined sequential write performance of all drives minus the parity drives
- Random read performance should match the combined random read performance of all the drives
- Random write performance in RAID 6 should be around 1/4 of the combined drives' random write performance, depending on the drives' random read/write performance ratio and CPU resources





Workload pattern	Total raw drives performance (measured)	Theoretical maximum RAID performance (calculated)	Real RAID performance (measured)	RAID Efficiency
Sequential read	228GB/s	228GB/s	226GB/s	99%
Sequential write	86GB/s	64.5GB/s	64.4GB/s	> 99%
Random read, 4K	12M IOPS	12M IOPS	12M IOPS	100%
Random write, 4K	6.5M IOPS	~1.6M IOPS	1.53M IOPS	~95%

Future testing with write-intensive drives is expected to demonstrate higher performance metrics.

Best Practices

When selecting the system architecture, it is recommended to allocate at least two CPU cores per NVMe drive installed in each system node. For example, if a system is configured with 24 NVMe drives, each node should have a CPU with at least 48 cores.

For memory requirements, the RAM capacity at each node should account for the operating system and Pacemaker needs. Additionally, it must include the memory required by the applications in scenarios where all cluster instances are running on a single node. Furthermore, it should accommodate the memory needed for xiRAID-managed RAIDs in cases where all system RAIDs are active on one node. The recommended value is 4 GB of RAM per RAID.

To get optimal performance, it is recommended that a tuned daemon be installed and configured at both nodes. It should use an `accelerator-performance` profile.

NVMe-based SBB architectures connect each NVMe drive to both SBB nodes using only two PCIe lanes. To fully utilize the performance of each drive, we recommend creating two NVMe namespaces per NVMe. The first namespace should be used in a RAID group on the first SBB node, and the second namespace should be used in another RAID group on the second SBB node.

Conclusion

Today, applications demand not only exceptional performance but also uncompromising resiliency. The combination of Supermicro SuperServer SSG-221E-DN2R24R and Xinnor's xiRAID Classic creates a highly efficient and robust Cluster-in-a-Box solution. This architecture provides unprecedented performance, utilizing the full potential of NVMe PCIe Gen5 drives while ensuring high availability and seamless failover capabilities.

The solution achieves remarkable efficiency, consistently operating at 95% to 100% of the raw hardware performance across various workloads. This optimization level ensures organizations maximize their investment in cutting-edge storage technologies while maintaining the reliability required for mission-critical operations. With resilience to multiple drive failures and a complete server node failure, the platform delivers uninterrupted performance and exceptional data protection.

By combining innovative hardware and software technologies, this joint solution addresses the evolving needs of modern storage systems, setting a new standard for performance, reliability, and scalability.

For more information on the SSG-221E-DN2R24R SBB Storage Server, see the Supermicro website: https://www.supermicro.com/en/products/system/storage/2u/ssg-221e-dn2r24r

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit www.supermicro.com

XINNOR

Xinnor is an Israeli-based software development company that specializes in creating innovative data storage solutions. Their main product is xiRAID, a software RAID engine that delivers exceptional performance. Visit www.xinnor.io