SAMSUNG

Pioneering the future of disaggregated storage solution with VMware vSAN and Samsung JBOF reference system

White Paper

This document provides empirical support for effectiveness of disaggregated storage architecture through a range of experiments conducted using the combination of VMware vSAN and Samsung JBOF reference system

Introduction

Disaggregated storage in the AI era: Embracing NVMe-oF

In the era of artificial intelligence (AI), the demand for high-capacity storage has become paramount, and a notable trend in meeting this demand is the move towards disaggregation. Disaggregated storage architecture stands out as a promising approach, offering a range of significant benefits that cater to the evolving needs of modern data centers and cloud environments.

Central to this innovative architecture are two foundational elements, processor node and storage node. The processor node has computation resources and acts as the brain of the storage system, orchestrating I/O operations, safeguarding data protection, and managing volume. The storage node has multiple storage devices, and is specifically designed to maximize storage capacity and data access performance.

NVMe-oF (NVMe over Fabrics) has emerged as a game-changer as the demands for faster storage communication grew. NVMe-oF is engineered to capitalize on the inherent advantages of NVMe SSDs, and deliver high performance comparable to local NVMe drives even across network fabrics. This not only results in ultra-low and stable latencies but also a scalable storage performance, fulfilling the stringent demands of modern applications and workloads.

Building on this foundation, a disaggregated storage architecture has following benefits:

- **Flexible scalability:** The flexibility afforded by the disaggregated storage architecture allows data centers and cloud environments to adapt rapidly to changing workloads and requirements. Computation and storage resources can be precisely tailored to meet the specific application needs, providing a more agile infrastructure. For example, for larger capacity requirements, storage nodes can be added independently while more number of diskless processor nodes can be deployed to fuel further computation resources. This flexible scalability leads to better resource utilization which reduces cost. Also, this separation results in an increased storage bandwidth and utilization, as processor nodes can access storage nodes in a more efficient and parallelized manner.

- **Enhanced availability:** Disaggregated storage architecture improves system resilience by decoupling the processor node and storage node. When a processor node failure occurs, storage nodes can easily reconnect to other alternative processor node resources, thus reducing the downtime and enhancing availability.

Leveraging VMware vSAN & Samsung JBOF reference system: A deep dive into disaggregated storage benefits

In this document, we aim to show the advantages of disaggregated storage architecture through practical experimentation using VMware vSAN, a prominent and widely adopted enterprise storage solution in the data center and cloud ecosystem, and Samsung JBOF reference system, a leading solution that harnesses the potential of NVMe-oF storage nodes.

Our goal is to demonstrate the key benefits of disaggregated storage architecture and illustrate the easy manageability and intelligent observability of Samsung's technologies. Through a series of experiments and performance evaluations, we will provide empirical evidence of the enhanced storage bandwidth, availability, utilization, and flexibility that disaggregated storage architecture can deliver in the AI-driven landscape of modern computing.

Solution Overview

What is VMware vSAN?

vSAN is a software-defined storage solution developed by VMware. It virtualizes local storage resources from multiple hosts in a cluster to create a shared storage pool. This technology simplifies storage management, enhances scalability, and improves performance while reducing the need for traditional SAN or NAS storage solutions.

vSAN ESA (Express Storage Architecture) is a specialized configuration of vSAN optimized for high-performance NVMe storage devices. It leverages the benefits of NVMe, such as low latency and high throughput, to deliver exceptional storage performance for demanding workloads.

In recent developments, VMware introduced vSAN Max as a groundbreaking solution. As shown in Figure 1, by adopting the disaggregated approach, it allows the separation of vSphere and vSAN. This separation allows for compute and storage resources to scale independently from each other, and represents a shift toward disaggregated storage architectures that offer supreme levels of flexibility, scalability, and resource optimization.

Even though vSAN Max disaggregates compute resources for VM workloads away from nodes providing storage processing, the storage processing resources remain converged with the storage devices. The disaggregated storage trend is expected to continue to evolve and more disaggregated vSAN configuration will emerge in the near future. The fundamental changes for future vSAN design would be the transition from local storage devices residing in host to storage devices using a storage node. This shift further extends the flexibility and efficiency of disaggregated storage and responds to the ever-increasing demands of modern data centers and cloud environments.

What is Samsung JBOF reference system?

It is an intelligent, fabric-attached JBOF for disaggregated storage that enables the customer servers to transparently access and manage NVMe SSDs through NVMe-oF, which supports both of RDMA and TCP. The servers can manage and access the SSDs as if they are locally attached with it. The disaggregated storage architecture enables the independent scaling of storage nodes from processor nodes, allowing data centers to effectively handle the rapidly increasing data and quickly recover from node failures.

Figure 2 depicts an architecture of Samsung JBOF reference system. It shows high performance, up to 400GbE. To help save the OpEx (Operating Expenditure) through high reliability, availability, and serviceability, Samsung JBOF firmware offers various management features for the server and SSDs such as

- SSD anomaly detection with Samsung Extended SMART
- Non-disruptive firmware upgrade (NDU)
- I/O path failure detection / auto recovery
- SSD power management
- Automated subsystem management

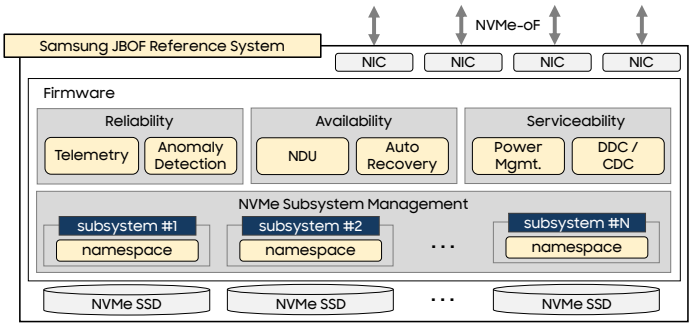

Figure 2. Samsung JBOF reference system architecture

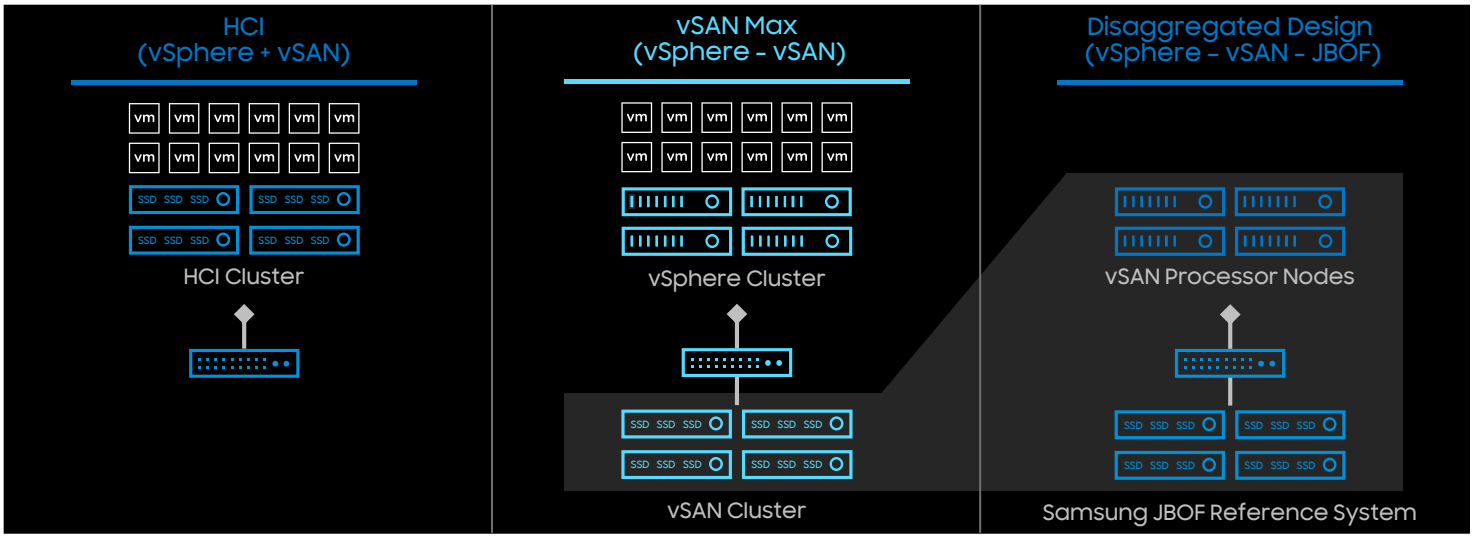

Figure 1. Evolution of vSAN architecture: From HCI to disaggregated design

SAMSUNG

Evaluation Environment

Overall test configuration

To ensure accurate and meaningful results, our test environment was meticulously set up with state-of-the-art hardware and optimal networking. This section delves into the comprehensive configuration employed in our testing phase. As shown in Figure 3, a total of 6 vSAN processor nodes were employed, each running on a DELL R750 server (diskless), optimized for high performance and reliability in our test environment. We utilized 3 Samsung JBOF reference systems that each hosts 32 NVMe SSDs. This setup ensured high-speed access to data and maximized the performance potential of the NVMe drives. For network connection, a high-speed 600GbE connection was established for the data path between vSAN processor nodes and Samsung JBOF reference systems, ensuring rapid data transfers and minimized any potential bottlenecks that could hamper performance. For traffic between the vSAN processor nodes, we implemented a 100GbE connection, guaranteeing seamless and efficient communication between the nodes.

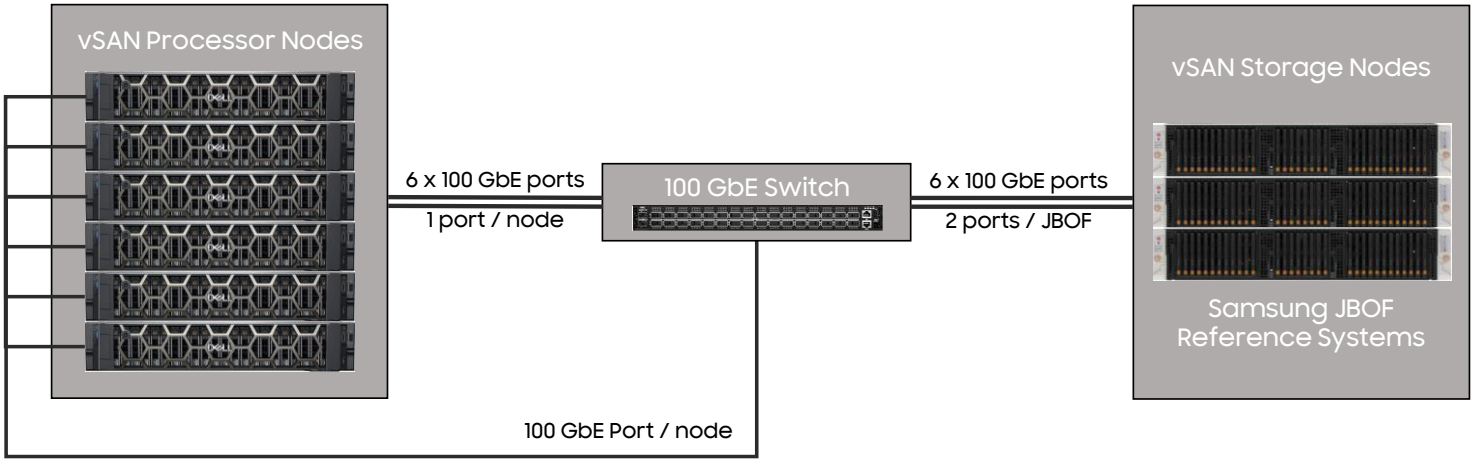

Figure 3. Overall test configuration: 6 vSAN processor nodes with 3 Samsung JBOF reference systems

After configuring our test environment, we captured key visuals to showcase the successful integration and setup. Figure 4 confirms the seamless claiming of the NVMe RDMA disks, which highlights the integration of Samsung JBOF reference system with the vSAN setup.

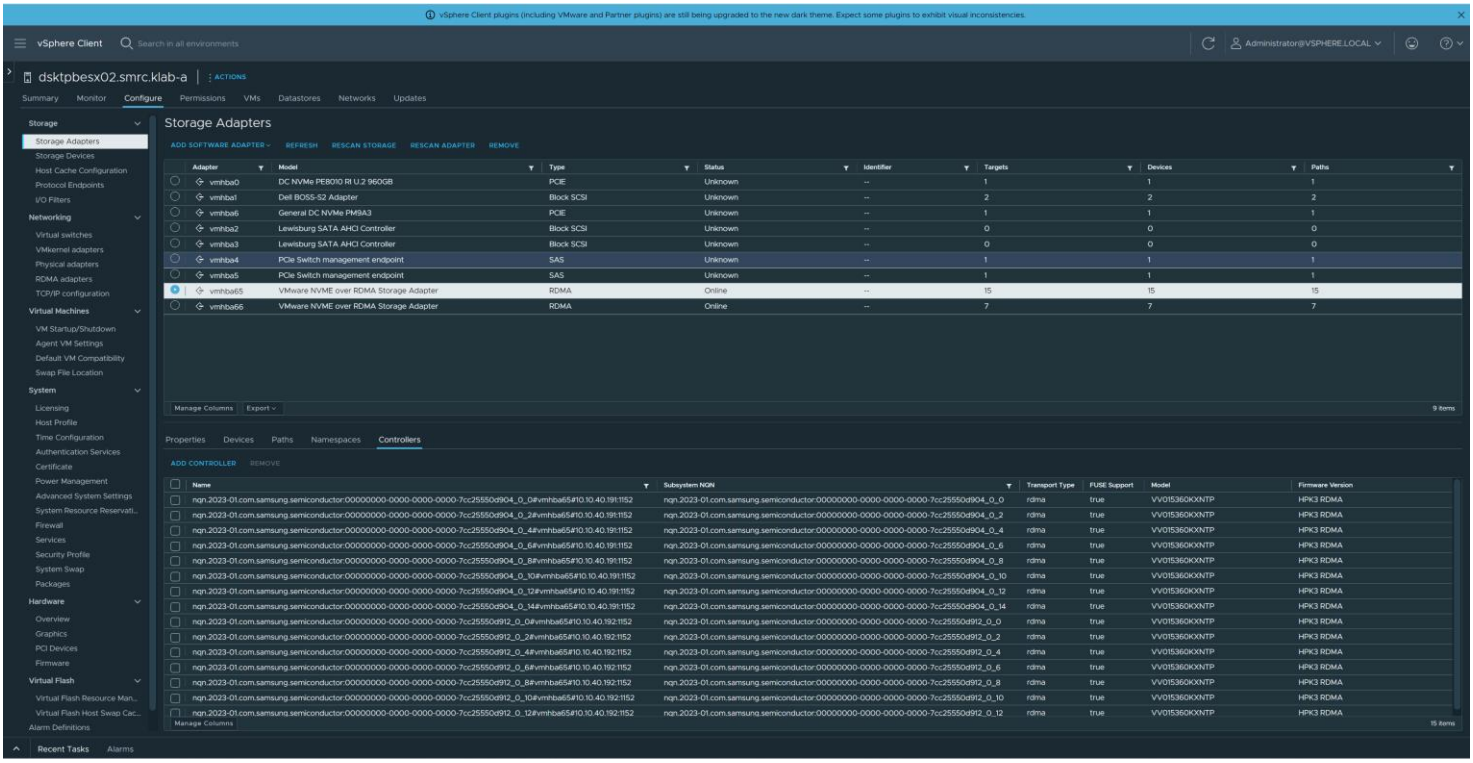

Figure 4. Configuration complete: VMware vSAN and Samsung JBOF integration snapshot

SAMSUNG

Hardware and software specification

We selected our hardware and software for the vSAN configuration to ensure optimal performance and reliability. We deployed VMware vSphere 8.0 U2 version, running on a Dell EMC PowerEdge R750 server. Table 1 shows a detailed specification we used for vSAN processor node.

	Specification
Server Platform	Dell EMC PowerEdge R750
CPU	Intel(R) Xeon(R) Gold 6342 CPU @ 2.80GHz (24C/48T) * 2EA
Main Memory	Samsung DDR4-3200 ECC/REG 32GB * 16EA (Total 512GB)
NIC	Mellanox ConnectX-5 100GbE Dual Port + Mellanox ConnectX-6 100GbE Single Port (Up to 300Gb/s bandwidth)
Operating System	VMware vSphere 8.0 U2

Table 1. Specification for VMware vSAN processor node

Table 2 shows the vSAN cluster configuration used for the test. We utilized RAID-5 with compression enabled, and each VMDK size is set to a consistent size of 50GB.

	vSAN configuration
RAID	RAID-5
Compression	Enabled
Per VMDK Size	50 GB

Table 2. vSAN cluster configuration

For the storage node, we employed Samsung JBOF reference system housed in the SMC server. We ran our Samsung JBOF firmware on a SMC ASG-2115S-NE332R server which has 32 Samsung PM1743 E3.S 16TB SSDs. Table 3 shows a comprehensive overview of the hardware and firmware specifications.

	Specification
Server Platform	SMC ASG-2115S-NE332R
CPU	AMD EPYC (Genoa) 9334 CPU @ 2.70GHz (32C/64T) * 1EA
Main Memory	Samsung DDR5-4800 32GB * 4EA (Total 128GB)
NIC	Mellanox ConnectX-6 100GbE Dual Port * 2EA (Up to 400Gb/s bandwidth)
Storage	Samsung PM1743 16TB NVMe 5.0 SSD (E3.S) * 32EA
Operating System	Ubuntu 22.04.1 LTS with Samsung JBOF Firmware

Table 3. Specification for Samsung JBOF reference system

Benchmark parameters: A detailed overview

Table 4 outlines the specific parameters employed for the compute scalability test, and Table 5 does the same for the capacity scalability test.

For the compute scalability test, 4KB random workloads were utilized to saturate the CPU resources. During the test, we progressively increased the number of vSAN processor nodes from 4 to 5 and then to 6, to observe the impact on performance and scalability.

For the capacity scalability evaluation, 128KB sequential workloads were selected to show the maximum throughput of Samsung JBOF reference system. Here, the number of JBOF systems was incrementally adjusted from 1 to 2 and finally to 3, allowing us to gauge the system's capacity performance as more JBOFs were added.

	Compute scalability (3 JBOFs)
4 vSAN processor nodes	15 VM per node (total 60 VM) 4 VMDK + 4 Thread per VM OIO = 240 per node
5 vSAN processor nodes	12 VM per node (total 60 VM) 4 VMDK + 4 Thread per VM OIO = 192 per node
6 vSAN processor nodes	10 VM per node (total 60 VM) 4 VMDK + 4 Thread per VM OIO = 160 per node

Table 4. Parameters for compute scalability test

	Capacity scalability (6 vSAN processor nodes)	
1 JBOF	8 VM per node (total 48 VM) 4 VMDK + 4 Thread per VM OIO = 128 per node	5 SSD per host 200 Gbits (200 GbE * 1)
2 JBOFs		10 SSD per host 400 Gbits (200 GbE * 2)
3 JBOFs		15 SSD per host 600 Gbits (200 GbE * 3)

Table 5. Parameters for capacity scalability test

- NumJob is assign to each VMDK (VMware Virtual Machine Disk) (e.g. 4 VMDK is 4 NumJobs).
- Thread is the queue depth (e.g. 4 threads = 4 QD).
- OIO (Outstanding IO) is the concurrent I/O in parallel on each processor node (e.g. OIO = # of VM x # of VMDK x # of Thread).

Key Benefits of Disaggregated Storage Architecture

Following the environment setup detailed earlier, we conducted a Proof of Concept (PoC) study, and our findings confirm advantages in the following areas:

- Flexible scalability
- Storage availability
- Manageability and observability

In the subsequent sections, we will delve into each of these areas, shedding light on our observations and the compelling benefits offered by the disaggregated storage architecture.

Adapting on demand: The promise of flexible scalability

One of the key advantages of disaggregated storage architecture is the ability to independently scale compute and storage resources. Specifically, when there's a surge in computing demands, organizations can dynamically ramp up compute resources by adding diskless servers without necessarily expanding storage. Conversely, as data storage needs grow, storage capacity can be augmented without incurring additional compute overhead. This capability not only provides vSAN with heightened scalability but also drives cost-efficiency.

Use case for compute resources demand

To validate the benefits of this scalability, we conducted a series of experiments. Initially, we created a CPU bottleneck on the vSAN processor nodes by running multiple VMs executing a random read workload, and scaled up the number of vSAN processor nodes from 4 to 5 and then to 6.

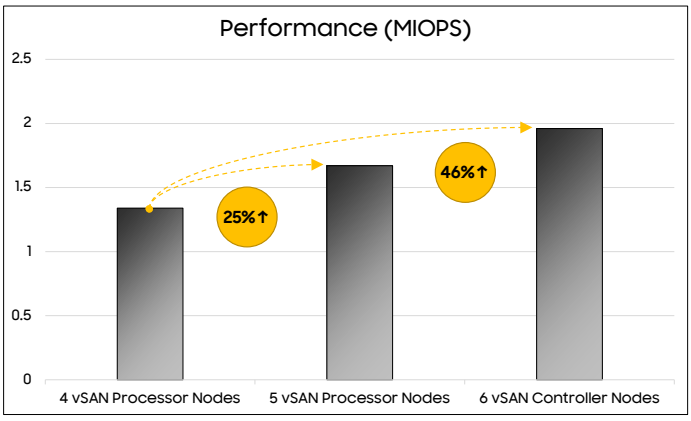

Figure 5. Performance scaling with vSAN processor node expansion

As shown in Figure 5, there was a tangible and proportional improvement in performance as the number of hosts increased. With 4 vSAN processor nodes, the system delivered 1.34MIOPS. When we added a fifth node, the performance showed 1.67MIOPS which is 25% enhancement. With 6 vSAN processor nodes, we saw a further rise in performance, reaching 1.96MIOPS, a total increase of 46% from 4 vSAN processor node configuration.

Use case for capacity resources demand

Next, we ran a sequential read workload to observe the performance impact on the storage node. In this experiment, we scaled up the number of Samsung JBOF reference systems from 1 to 2 and then 3.

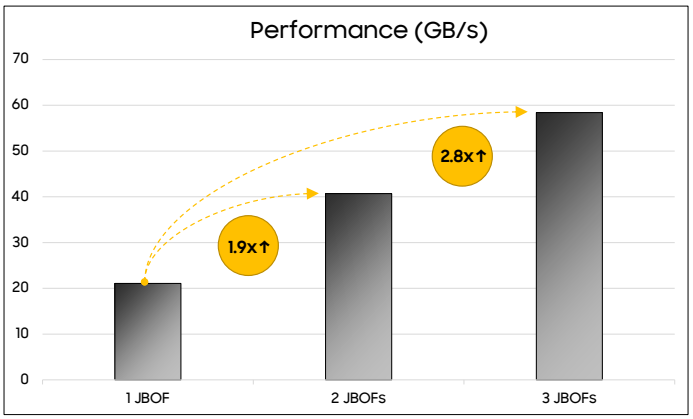

Figure 6. Performance evolution with Samsung JBOF reference system addition

Figure 6 shows how performance scales with the addition of storage nodes. With just a single JBOF system, the system achieved a throughput of 21.1GB/s. Doubling the storage by adding another JBOF system doubled the throughput, reaching 40.7GB/s. And, with three Samsung JBOF reference systems, the performance surged to 58.4GB/s.

An additional benefit is its capability to support exceptionally dense vSAN clusters. It permits the connection of over 24 drives to each diskless vSAN processor node, facilitating the easy addition of single or multiple drives to the vSAN cluster. In contrast, without the JBOF system, integrating multiple, or sometimes even a single drive can be challenging. In certain cases, it might necessitate the inclusion of a new HCI node to the vSAN cluster.

※ Note: Actual performance may vary and is subject to change based on the evaluation environment, including specific hardware and software configurations.

SAMSUNG

Use case for on demand resource adaptation

The ability to independently scale processor nodes without attached storage devices or leverage cost-effective storage nodes during infrastructure expansion or in response to failures reduces costs significantly. As shown in Figure 7, this disaggregated approach offers adaptability for specific needs: high performance, balanced, or high capacity. To boost performance, simply add more vSAN processor nodes; for increased capacity, expand with JBOF units. This flexible system allows for targeted scaling, minimizing costs and eliminating the need for expensive storage components.

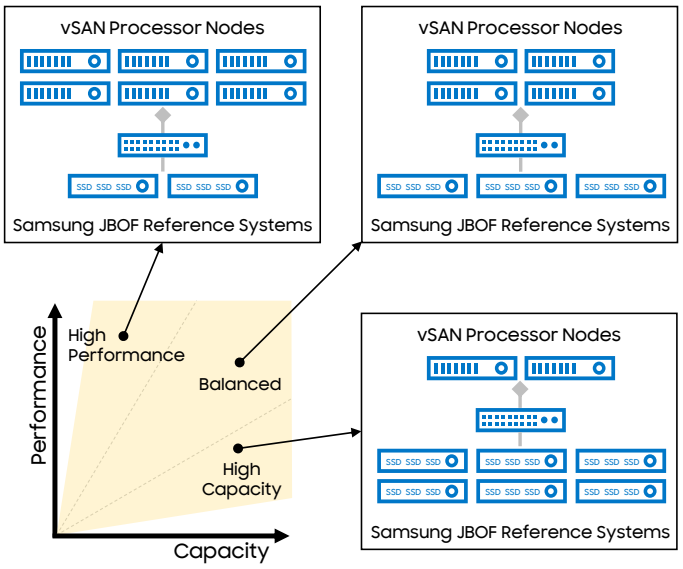

Figure 7. Adapt on demand for performance and capacity

Operational efficiency: The quest for resync time reduction

Disaggregated storage architecture greatly cuts down the costs that come with vSAN processor node failures. In an aggregated vSAN HCI cluster, or a disaggregated vSAN Max cluster, the storage cluster went through a costly resync operation whenever a node failed, since the node processing storage also contained storage devices. Moreover, this resync operation time increased with the growth in storage capacity. During a resync operation, the quality of service (e.g. performance) decreases significantly. More importantly, the storage cluster would be exposed to the risk of losing data if another node failure occurs during the resync process.

As shown in Figure 8, with our new approach, however, this long resync operation can be avoided when a vSAN processor node fails. This is achieved by utilizing the disaggregated storage architecture and vSAN's reconnection policy. For instance, in the event of a vSAN processor node failure, other active vSAN processor nodes can connect to the storage node that belonged to the failed vSAN processor node. In this way, all the data is pulled by other vSAN processor nodes and the vSAN cluster will continue to work normally without any service quality impact. This not only saves a lot of money but also avoids extra problems that can happen if there's another failure during a resync.

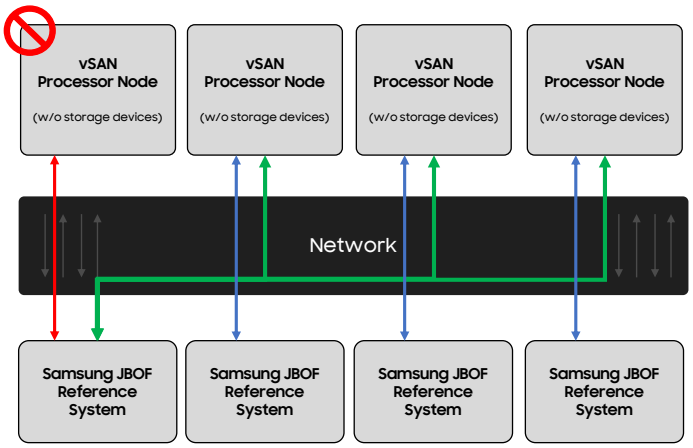

Figure 8. Zero downtime for processor node failure

Use case for resync time reduction

Table 6 shows the estimated resync time calculation. The formula MAX((BytesToSync / AvgWriteBytesPerSec) * 5) calculates the resynchronization time following a failure. Here, "BytesToSync" signifies the data amount needing resynchronization, while "AvgWriteBytesPerSec" depicts the average disk write speed, of which only 20% is designated for data resync. The operation's duration is essentially dictated by the slowest disk, as suggested by "MAX()".

For example, with a 512 TB storage node, conventional methods anticipate a resync time of roughly 2.42 days. As storage capacity increases, this resynchronization time can exacerbate, posing serious challenges. A prolonged resync time not only impacts operational efficiency but also increases the risk of additional failures. During such extended durations, any subsequent failure can escalate the likelihood of data loss. However, our disaggregated storage architecture can substantially mitigate this issue, drastically shortening resync durations, ensuring data recovery is swift, and minimizing operational interruptions.

* MAX((BytesToSync / AvgWriteBytesPerSec) * 5)
  - bytesToSync is the size of data that needs to be resynchronized.
  - AvgWriteBytesPerSec is the average WRITE bandwidth per a disk. 20% of total WRITE bandwidth is assigned to the data resync.
  - MAX means the total resync time is bound to the slowest disk.

* Example
  - 16TB * 32 SSDs = 512 TB per storage node
  - 100 GbE (12.5 GB/s) * 20% = 2.5 GB/s
  - 512 TB / 2.5 GB = 2.42 days

Table 6. Estimated resync time calculation

SAMSUNG

Use case for upgrade/maintenance

Historically, the task of upgrading or maintaining storage solutions has been fraught with complexities, often making it a cumbersome undertaking.

In conventional configurations, especially with tightly coupled compute and storage, depicted in Figure 9, upgrading a vSAN processor node typically required data movement. This operation not only consumes considerable time but also overburdens the system resources, potentially affecting service quality.

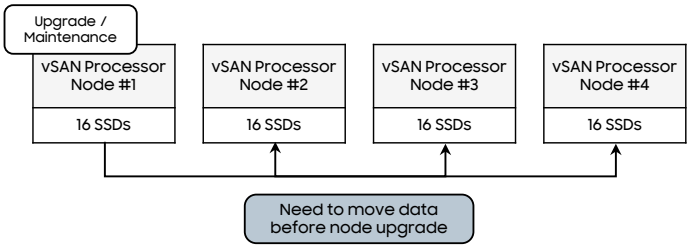

Figure 9. Data movement requirement with vSAN HCI

However, the introduction of the disaggregated architecture has revolutionized this process. As illustrated in Figure 10, in a disaggregated setup, other vSAN processor nodes can seamlessly manage the data of a vSAN processor node slated for an upgrade, facilitating the upgrade without any data migration. This refined methodology minimizes downtime, bolsters operational efficiency, and guarantees a more fluid and expedited upgrade experience.

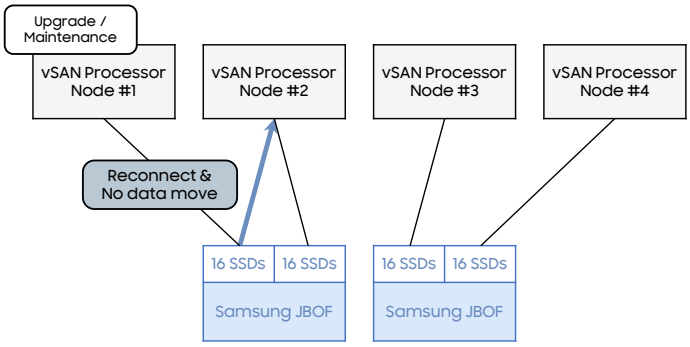

Figure 10. Effective upgrade without data movement with JBOF

Seamless operations: Easy management and insightful observability

Transitioning to disaggregated storage architecture not only alters the layout of storage but also reshapes the approach towards management and monitoring. Samsung's tailored features, in harmony with vSAN's capabilities, ensure that users can navigate, manage, and monitor their storage infrastructure with remarkable ease and precision

Use case for management

For easy management, there are several features which includes auto bring up/failover, CDC (Centralized Discovery Controller) /DDC (Direct Discovery Controller) functionalities, and NDU (Non-Disruptive Update).

The auto bring up/failover feature guarantees swift system recoveries post power disruptions and during transient failures, thereby enhancing system uptime and reliability. The CDC and DDC facilitate a streamlined discovery and connection process, centralizing the management of multiple storage nodes. This not only reduces operational complexities but also fosters efficient scalability. Finally, automatic and seamless firmware updates guarantee that the storage node consistently delivers peak performance. It also ensures that the system remains up-to-date without any operational halts.

Use case for monitoring

For insightful observability, Samsung JBOF reference system offers a wealth of telemetry information, including Samsung Extended SMART data. This detailed telemetry data provides a comprehensive view of the status and performance.

As a more advanced technology, it provides SSD failure prediction and degraded detection features using a wealth of telemetry data. This proactive approach helps anticipate potential issues with SSDs and ensures stable and reliable SSD usage, surpassing the reliability of simple SSD usage.

Conclusion

In an evolving landscape dominated by data-driven operations, the choice of storage architecture is pivotal. Our hands-on work with VMware vSAN and Samsung JBOF reference system illustrates that even in scenarios of vSAN processor node upgrades or failures, all the data is seamlessly handled by other vSAN processor nodes, ensuring the vSAN cluster continues to function without dip in service quality. This evidence underscores the compelling advantages of disaggregated storage architecture. While decoupling processor nodes from storage nodes might induce a slight initial CapEx increase, it simultaneously drives remarkable cost-efficiency. Notably, for workloads seeking to bolster storage without a proportional rise in compute resources, the architecture necessitates fewer compute-only vSAN processor nodes, optimizing resource allocation. The accrued long-term operational efficiencies, enhanced scalability, improved resiliency, and granular management features further validate the case for this approach.

In essence, as we navigate the future of storage, the confluence of VMware vSAN and Samsung JBOF reference system stands as a beacon, highlighting the potential of disaggregated storage. As organizations around the globe recognize and adopt this synergy, we anticipate disaggregated storage to set new industry standards, steering the future of storage solutions in the AI-dominated era.

Reference

- Samsung JBOF reference system: https://semiconductor.samsung.com/news-events/tech-blog/samsung-announces-innovations-to-enhance-memory-customer-experience-in-data-centric-era-at-fms-2023/
- Samsung PM1743 SSD: https://semiconductor.samsung.com/ssd/enterprise-ssd/pm1743/
- SMC storage A+ server: https://www.supermicro.com/en/products/system/datasheet/asg-2115s-ne332r
- vSAN: https://core.vmware.com/api/checkuseraccess?referer=/sites/default/files/vSAN%208%20TCO%20White%20Paper.pdf
- vSAN Max: https://core.vmware.com/blog/introducing-vsan-max
- vSphere 8.0 U2: https://docs.vmware.com/en/VMware-vSphere/8.0/rn/vsphere-esxi-802-release-notes/index.html

Acknowledgement

We express our sincere appreciation to SMRC (Samsung Memory Research Center) for their pivotal role in this collaborative effort. Their assistance in providing the experimental environment was essential for our research on the synergies between VMware vSAN and Samsung JBOF reference system. This project showcases how SMRC facilitates effective collaboration between Samsung memory products and customer/partner solutions, driving innovation in memory technology.

Samsung Electronics Co., Ltd.

129 Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677, Korea www.samsung.com 1995-21

SAMSUNG