Marc Staimer, President & CDS Dragon Slayer Consulting

W h i t e P a p e r

Hyperparallel Seismic Processing

How Pavilion and NVMe-oF Redefine Oil & Gas Storage Performance

Introduction

Oil and Gas organizations are well known for using incredibly demanding high-performance applications. Applications include Halliburton Landmark Seispace, Paradigm Echos, Schulmberger Seismic services, and more. Most of these applications run in what is referred to as high-performance computing systems (HPC). HPC aggregates multiple servers computing power to deliver much greater performance than can be derived from a single server in order to solve compute-intensive problems such as seismic processing and visualization. Servers are commonly referred to as nodes. The operating system running on the nodes are generally Linux or Windows. The HPC nodes are then clustered via a high-speed interconnect to aggregate their performance.

The critical Oil and Gas applications are typically 2D and 3D seismic processing. Seismic data volumes have increased exponentially over the past few years because of greatly improved multi-sensor arrays and sophisticated acquisition techniques. The seismic processing applications have continually advanced their analytic techniques to derive more accurate actionable information. More recently, Oil and Gas organizations have been adding AI and machine learning to refine their ability to accurately analyze, interpret, and make decisions on both this higher resolution data and historical oil field data.

The primary HPC performance bottleneck is the storage. HPC seismic applications generally read and process huge amounts of data concurrently. Throughput is the primary requirement and demands multiple GB/s to TB/s. Traditional SAN storage simply wasn't designed for those consistent rates and cannot keep up. It slows the seismic processing applications to unacceptable levels. This is why Oil and Gas HPC systems have primarily solved the storage throughput problem either by using direct-attached storage (DAS) with software-defined storage (SDS) within the HPC server nodes running a parallel file system, or via a clustered parallel file system storage (CPFSS) such as GPFS (now known as IBM Spectrum Scale), open source Lustre, Panasas, Quobyte, and others.

Both of these methodologies are struggling to keep up with the paradigm shift in data volumes and resolution. Problems in performance, performance density, and capacity density are becoming untenable.

This paper examines why these common Oil and Gas HPC storage methodologies fail to provide the necessary storage performance, performance density, and capacity density required with today's modern seismic processing. It then details how a Pavilion Data disaggregated NVMe-oF storage array uniquely solves all three of the problems cost-effectively.

Table of Contents

Introduction 2

Current Oil and Gas HPC Storage Methodologies 4

DAS with SDS 4

Clustered Parallel File System Storage (CPFSS) 6

The Pavilion Solution 7

Pavilion HFA Specs 7

How Pavilion Solves Oil and Gas High-Performance Requirements 7

How Pavilion Solves Oil and Gas Storage Performance Density Requirements 7

How Pavilion Solves Oil and Gas High Data Protection Requirements 8

How the Pavilion HFA Solves Oil and Gas High Scalability Requirements 8

How the Pavilion HFA Solves Oil and Gas Reasonable Cost Requirements 8

How the Pavilion HFA Stacks up to DDN Storage 8

Conclusion 9

Current Oil and Gas HPC Storage Methodologies

As stated above, the two most prevalent storage solutions or methodologies for Oil and Gas HPC applications are DAS in the individual HPC clustered server nodes with an SDS-based shared parallel file system and clustered parallel file system storage. Each has pros and cons, but neither is currently keeping up with the advancements in 2D and especially 3D seismic processing. The issues are performance, performance density[1], capacity density, and total cost of ownership (TCO). A deeper look at them shows why.

DAS with SDS

DAS leverages the internal media storage within each node of the HPC cluster. That media can be storage class memory (SCM) drives, NVMe flash drives, SATA/SAS solid-state drives (SSDs), or hard disk drives (HDDs). In an HPC cluster, each node is part of a clustered parallel file system. Every HPC server node has access to its own and every other node's data.

There are some compelling perceptions of DAS that makes it attractive to the Oil and Gas HPC applications. The first is that the storage drives are physically closer to the compute. That generally enables lower latencies than shared external storage because of the shorter physical distance and speed of light latency. Lower latency should produce faster access to the first byte of a read and faster writes. Also, faster transactions, which is far less important with Oil and Gas applications.

There is the additional perception that DAS or SDS is easier to implement, manage, and upgrade. And finally, the most compelling perception is the price of DAS drives. It is thought that DAS drives cost three times less than the same drive in a shared storage system. All of these perceptions are flawed.

Take the first perception on performance. Data located on the local drives of an HPC node will indeed be relatively fast. However, what happens when the data being sought is located on the drives of a different node? Then the performance takes a rapid nosedive. The data must come from the storage in the node with the data, through the node's storage controller, PCIe, memory, CPU, I/O, interconnect to the HPC node requesting it. All of which adds latency to the performance and affects throughput, depending on several factors including:

1) Demands on the HPC node with the data
2) Interconnect traffic and available bandwidth
3) Memory utilization
4) CPU utilization
5) PCIe utilization at the time

> *(Any HPC server node accessing the data on another node will not have the same or similar performance of accessing data on itself. All these aspects cause stubborn inconsistent seismic application performance variations).*

Then there is the issue of HPC server node storage capacity, which is regularly constrained. Most HPC server nodes are limited to no more than a few dozen NVMe SCM or flash drives. NVMe drives are restricted by the number of available x86 PCIe v3 lanes in each server node. The server nodes can handle more SATA/SAS SSDs and HDDs, depending on the drive slot real estate in the server node. However those drives are markedly slower than NVMe SSDs. Capacity limitations and swelling data make it increasingly likely that requested data will be on a different node than the

[1] Performance density is the amount of performance measured in throughput and/or IOPS per rack unit (RU). Performance density has become increasingly important as data center real estate has become costlier, see the Pavilion performance density white paper https://paviliondata.io/performancedensitywhitepaper/

.

one requesting it. This leads to poor performance density and rapid expansion of rack space, floor tiles, NICs, cables, transceivers, switches, power, cooling, ops, management, troubleshooting hot fixes, patching, upgrades, maintenance, personnel training, and more, increasing total cost.

HPC implementations attempt to moderate performance issues are by manually localizing the data as much as possible. Although localization takes a lot of effort and time ongoing while complicating scalability, data management, and data protection. Another attempt at mitigating this problem is the use of very high bandwidth Interconnect such as 100Gb/s Ethernet or Infiniband and remote direct memory access (RDMA) to minimize latencies and increase throughput.

The perception of simpler management also clashes with real-world reality. Drive failures (and there are always drive failures) significantly lower storage and HPC node performance. RAID rebuilds are very memory and CPU intensive sucking up cycles that are not available for HPC processing during rebuilds, also using RAID can also be costly, as it lowers overall storage capacity. Drive rebuild times have also crept up significantly as well. Mirroring is an alternative, albeit costly one. Replacing a drive typically requires a disruption by opening up the server node and taking it offline. Not very simple. Scaling capacity or performance becomes an issue because it's coarse grain. Adding capacity or performance requires adding additional HPC server nodes increasing storage, processing, and networking that may not be needed. This adds more resource consumption and complexity to the cluster and increases TCO.

The most inaccurate and persistent DAS perception is that it's the lowest HPC storage cost. DAS frequently costs much more than expected. In addition to the wasteful data center consumption from very poor storage performance density previously discussed, there are the high data protection costs. The most common way to protect data from drive or HPC node failures, while limiting performance degradation, is to implement multi-copy mirroring. Multi-copy mirroring makes minimally two or more data copies and places them on different drives in different HPC nodes. Except each copy consumes 100% more storage capacity. Creating a consistent copy of the application and its data requires the application be paused during the copy, increasing the application completion time. The cost savings from utilizing server drives over shared storage drives just vanished. If more than two copies are made, storage costs become much higher than shared storage.

There are several generally overlooked significant problems caused by HPC DAS. These include:

- **SKU Sprawl**. Manufacturer HPC server node configurations or SKUs change minimally once a year and often multiple times per year. The server bought last year is just as likely to not exist when another one is required. New HPC server nodes will have a different configuration, such as different CPUs, memory, SCM, NVMe SSDs, SSDs, HDDs, and other components. Lots of SKUs over time create excessive numbers of combinations, spares, management complexity, troubleshooting frustration, while likely becoming operationally overwhelming and costly.
- **Excessive Data Movement**. Poor storage performance density leaves few cycles for data protection. DAS-based SDS steals CPU cycles and memory from the HPC server node. Disk and data protection (RAID, snapshots, and multi-copy mirroring) are storage, CPU, and memory-intensive taking away additional resources from the HPC server nodes. Multi-copy mirroring makes asynchronous copies of the data and increasing bandwidth consumption since it moves these copies across the interconnect to other nodes.
- **Minimal Storage Services**. One of the ways the performance hits are being addressed is through the elimination of storage software services such as RAID, snapshots, thin provisioning, and others, because they consume too many resources. Multi-copy mirroring and RAID does not protect the data from human errors such as accidental deletions, malware such as ransomware, and disgruntled employees. That requires other forms of data protection such as snapshots, time-stamped replication, or backup.

Clustered Parallel File System Storage (CPFSS)

CPFSS extends the parallel file system to an external shared storage system. The storage system appears as if it is just other HPC server nodes with lots of I/O ports and lots of data storage. A block storage system sits behind those nodes. All of the HPC server nodes draw their data in parallel from the nodes fronting the storage.

A good way to understand CPFSS is as a 2-sided unbalanced HPC cluster. One side of the cluster is the compute. The other side is the storage. All of the compute server nodes read and write data concurrently from the storage side. The storage side has far fewer HPC server nodes. There are several good aspects to this methodology which makes it so popular. All HPC cluster nodes have access to all the data concurrently. The HPC server nodes see the HPC storage nodes like all other nodes in the cluster. Simple. Configurations are not complex or hard to implement. And the data is fully protected with snapshots, clones, and RAID.

However, there are problems with this type of system. It increases storage system bottlenecking reducing performance, the data protection reduces performance, there is additional latency because of multiple systems, it has meager storage performance density, and very high cost. A closer look shows why.

Reduced Performance

Most CPFSS are limited to two active-active storage controllers. The throughput cannot exceed the throughput of those two controllers. This bottleneck typically means adding additional CPFSS systems to meet requirements, complicating provisioning, access, operations, management, and troubleshooting. Simplifying access is generally resolved by duplicating data on multiple systems. This requires a lot more capacity in each CPFSS system and significantly more cost.

Data protection such as snapshots are CPU intensive affecting CPFSS controller performance and causing difficult compromises between RPO (recovery point objective or the amount of data that can be lost based on the time gap between snapshot events) and system performance.

Moreover, there is always additional latency between requesting HPC server nodes, HPC storage nodes, and storage controllers. Latency affects the beginning of every job and concurrent jobs.

Modest Performance Density

CPFSS systems consume multiple rack units. Storage capacity density can be okay, but performance density is limited. Take DDN's latest fastest block system the SFA18K active-active dual controller storage system. The SFA18K has throughput of up to 90GB/s and 3.2 million 4K IOPS, with (16) 40/100Gbps InfiniBand ports in 4 RU and uses the SAS interface with non-NVMe SSDs. That's their densest performance configuration. It equates into 22.5 GB/s throughput, 800K 4K IOPS, and 4 InfiniBand connections per RU. It's a 50% increase over their last fastest system.

Keep in mind that the "up to" performance is in perfect conditions and unlikely in most circumstances. Nothing else is running. No data protection is occurring, no RAID rebuilds, no encryption, and so forth. Remember, it's still limited by the throughput of two active-active controllers and the SAS interface.

Performance and Capacity Scalability

CPFSS tends to scale capacity into multiple PBs well and fairly linearly. Performance scalability is a completely different story. Most CPFSS such as DDN's SFA18K are limited by a dual controller architecture. The throughput and IOPS are constrained by the limits of the two storage controllers. It does not matter how many drives are added, performance cannot exceed that of the two storage controllers. This hinders performance scalability.

Very High Cost

CPFSS vendors recognize that there are not very many storage systems that can handle Oil and Gas throughput requirements and they charge a premium for their storage system. In a best-case scenario, the storage drives alone in the system cost three to six times of the of same drives in a generic server. Pavilion has a better way to balance performance and cost.

The Pavilion Solution

Pavilion took a hard look at Oil and Gas storage requirements and realized that the storage needed a fundamental rethink. Oil and Gas demands extremely high throughput storage performance, excellent storage performance density, capacity density, and first-rate data protection; it also needs a reasonably low total cost of ownership. This is what Pavilion is delivering with no compromises in its disaggregated NVMe-oF-based storage array, the Hyperparallel Flash Array (HFA).

Pavilion HFA Specs

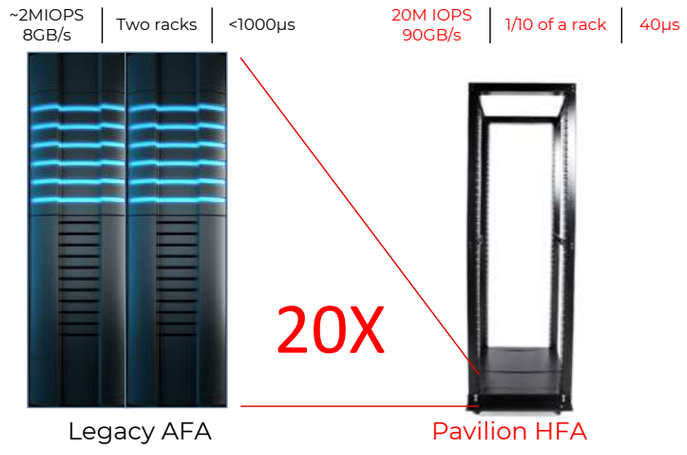
Each array is a compact 4 RUs with 20 active/active controllers, 90GB/s throughput, 20 million 4K IOPS, 40 Ethernet or InfiniBand ports, and 1.1PB of NVMe flash in 72 drive slots. That translates into a RU capacity density of 275TB and a performance density of 22.5GB/s throughput, 5 million 4K IOPS, and 10 Ethernet/InfiniBand ports.

How Pavilion Solves Oil and Gas High-Performance Requirements

The Pavilion HFA provides the best of DAS and CPFSS in a single system, enabling HPC applications to disaggregate their storage from compute. It utilizes extremely fast, low latency, high throughput NVMe flash drives and NVMe-oF. 40 100Gbps Ethernet/InfiniBand ports provide equivalent latency to that of internal NVMe flash drives in HPC server nodes, but with far greater capacity shared by ALL of the HPC server nodes. For NVMe-oF, Pavilion uses the NVMe/TCP and NVMe/RoCE v2 standard over Ethernet and NVMe/RDMA over InfiniBand. The shared Pavilion HFA storage appears to the HPC server nodes as internal DAS but has a much lower cost. Its capacity and performance are shared by ALL of the HPC nodes while it's managed, maintained, and operated as a single entity. NVMe-oF using InfiniBand is ideal for existing HPC customers, while NVMe-oF over Ethernet provides new HPC users with a lower cost solution than using InfiniBand. Pavilion HFA is flexible, letting HPC customers use NVMe-oF over Ethernet and InfiniBand simultaneously.

How Pavilion Solves Oil and Gas Storage Performance Density Requirements

Not only is the Pavilion HFA 20X smaller and cost 25X less than a legacy AFA configuration that provides the same performance, but it also more than doubles the HPC server node compute density per rack. It does this by eliminating requirements for internal storage in the HPC server nodes. This allows those nodes to be much smaller, instead of using servers designed for lots of DAS storage, HPC customers



can use smaller 1RU servers, micro servers, even blade servers. Complex I/O operations are prevalent in many critical Oil and Gas applications, and this is where the Pavilion HFA really shines. Providing 20M IOPS to these applications and 90GB/s of throughput, lets them analyze more data in less time, accelerating time to results and increasing revenue.

How Pavilion Solves Oil and Gas High Data Protection Requirements

The Pavilion HFA delivers first-rate data protection with RAID 6, data corruption detection, and almost instant space-saving snapshots. And it does so with a nominal impact on the performance. With up to 10 times the number of storage controllers of any other storage system, all 20 active-active storage controllers have access to all NVMe flash drives. That additional horsepower means there is more than enough CPU resources for snapshots, data corruption repairs, and RAID rebuilds with no noticeable impact on performance. Because the storage is shared, it does not require multi-copy mirroring to protect against HPC server node failures.

How the Pavilion HFA Solves Oil and Gas High Scalability Requirements

The Pavilion HFA scales up to 90 GB/s throughput and 20 million 4K IOPS with 40µs of latency and 1.1PB per 4RU. Additional 4RU systems add another 90 GB/s, 20 million 4K IOPS, and 1.1PB to CPFSS. Multiple units can be stacked to provide up to 900 GB/s throughput and 400M IOPS with 40µs of latency and 22PB per rack meeting the needs of the performance-sensitive Oil and Gas applications.

How the Pavilion HFA Solves Oil and Gas Reasonable Cost Requirements

A significant cost of shared storage systems are the drives. Pavilion recognized this and launched **OPEN**CHOICE, allowing customers to source their own NVMe flash drives from server suppliers, or repurposing HPC server drives, avoiding vendor lock-in. This allows the customer to repurpose the DAS server SSDs for their applications by installing them in the Pavilion HFA. Using low-cost server drives in a high-performance shared storage array cuts significant major costs from the storage system.

Additional cost savings come from performance, capacity, and compute density gain benefits of storage disaggregation. Storage disaggregation means 2-3X fewer systems and associated racks, power, and cooling. That reduction can be as much as the savings from **OPEN**CHOICE. The 100% NVMe flash drives by themselves generally reduces operating expenses. It's because the drives have no moving parts, are not as fragile as HDDs, and consume much less power and cooling, so they last longer. All these cost savings equate into achieving an application TCO that is lower than HPC servers that use DAS or DAS-based SDS.

How the Pavilion HFA Stacks up to DDN Storage

DDN storage is the primary storage utilized in most Oil and Gas seismic process application implementations The Pavilion HFA has the best storage and compute density in the market and is far superior to the DDN SFA18K as this table shows.

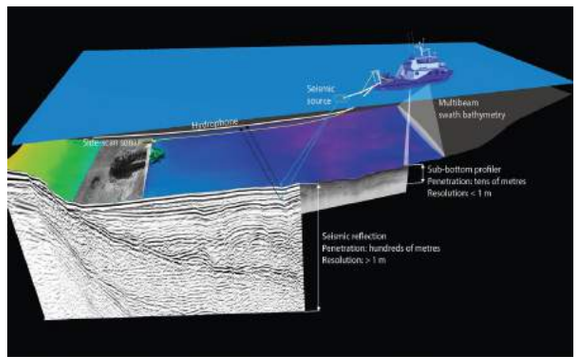
	DDN SFA18K	Pavilion HFA	Pavilion Advantage
Throughput (GB/S)	90	90	Equal
Sustained	????	90	Significant
IOPS (millions)	3.2	20	525%
Sustained	????	20	>525%
Latency	~300 µs[1]	40 µs	>8X Lower
InfiniBand EDR Ports	16	40	150% More
Ethernet 100Gb Ports	16	40	150% More
Storage Controllers	2	20	10X More

[1] DDN SFA18K only utilizes flash as a cache front end. There is no NVMe-oF meaning read latencies, if in the cache, are likely to start at 300 µs.

Oil and Gas organizations are well known for using incredibly demanding high-performance applications both upstream and downstream. Pavilion's HFA provides Oil and Gas customers with the storage performance, performance density, and capacity density these applications require.

Conclusion

Oil and Gas high-performance applications such as seismic processing, need extreme storage performance and scalability to keep up with the higher 3D resolution and its exponential data growth. Those needs translate into much higher throughput performance; much better throughput performance density; much better capacity density; top-line data protection; performance and capacity scalability; and a rational, affordable cost.



This is where the Pavilion HFA stands alone. It meets or exceeds each of these Oil and Gas high-performance application sustained requirements, not just peak. It is the only system architected to provide the best performance and cost of server direct attached storage with the convenience of Clustered Parallel File System Storage.

The Pavilion HFA makes Oil and Gas HPC applications soar!

For more detailed information about Pavilion's disaggregated storage and high-performance rack-scale systems please go to:

- Website: http://www.pavilion.io/
- Email: info@paviliondata.io
- Phone: (669) 263-6900

Paper sponsored by Pavilion. **About Dragon Slayer Consulting:** Marc Staimer, as President and CDS of the 21-year-old Dragon Slayer Consulting in Beaverton, OR, is well known for his in-depth and keen understanding of user problems, especially with storage, networking, applications, cloud services, data protection, and virtualization. Marc has published thousands of technology articles and tips from the user perspective for internationally renowned online trades including many of TechTarget's Searchxxx.com websites and Network Computing and GigaOM. Marc has additionally delivered hundreds of white papers, webinars, and seminars to many well-known industry giants such as: Brocade, Cisco, DELL, EMC, Emulex (Avago), HDS, HPE, LSI (Avago), Mellanox, NEC, NetApp, Oracle, QLogic, SanDisk, and Western Digital. He has additionally provided similar services to smaller, less well-known vendors/startups including: Asigra, Cloudtenna, Clustrix, Condusiv, DH2i, Diablo, FalconStor, Gridstore, Nexenta, Neuxpower, NetEx, NoviFlow, Pavilion Data, Permabit, Qumulo, SBDS, StorONE, Tegile, and many more. His speaking engagements are always well attended, often standing room only, because of the pragmatic, immediately useful information provided. Marc can be reached at marcstaimer@me.com, (503)-312-2167, in Beaverton OR, 97007.