Solution Brief

Data Center | Splunk

Modernizing Splunk Platforms with VAST Data Universal Storage

VAST Data Universal Storage can reduce Splunk data size by 2.5x, with minimal impact on average Splunk search run time and indexer ingest rate, while using real-world, high cardinality production data[1]

Solution Benefits

- **Reduce data storage needs.** VAST Data Universal Storage can significantly reduce storage needs and organization's storage costs, with minimal impact to search run time and indexer data ingest rates.
- **Improve operational efficiency.** Petabytes of data can reside in less than half of a data center rack.
- **Scale compute and storage independently.** Allows organizations to accommodate massive storage growth demands without additional compute infrastructure.

Authors

Victor Colvard
Security Systems Engineer
Intel IT

Murali Madhanagopal
Solutions Architect
Data Center Platforms Group

Frank Ober
Solutions Architect
Data Center Platforms Group

Elaine Rainbolt
Industry Engagement Manager
Intel IT

Merritte Stidston
Technical Solution Specialist
Data Center Platforms Group

Executive Summary

Splunk Enterprise is a software platform widely used for monitoring, searching, analyzing, and visualizing data. It captures, indexes, and correlates massive amounts of data in a searchable container and produces graphs, alerts, dashboards, and visualizations. In early 2018, Intel IT built a Cyber Intelligence Platform (CIP) based on Splunk and Apache Kafka. This platform ingests data from hundreds of sources and security tools, providing context-rich visibility and a common work surface, decreasing the time required to identify and respond to sophisticated cyber threats.

Often, Splunk is deployed on high-performance, converged infrastructure. This approach can be costly if the Splunk indexers (servers) are designed with SSDs for storing both "hot" and "cold" data. A converged infrastructure can also lead to the addition of Splunk indexers simply to increase data storage capacity. An obvious alternative is to disaggregate compute from storage. However, disaggregated infrastructure can cause negative performance impacts to both Splunk search run times and data ingest rates.

VAST Data Universal Storage offers a unique solution for disaggregating Splunk platforms. VAST Data uses advanced data reduction algorithms and Intel® Optane™ SSDs to reduce Splunk data storage requirements without sacrificing performance. We tested a VAST Data all-flash storage enclosure with Intel IT's high cardinality, production data and reduced Splunk data size by 2.5x.[1] Average Splunk search run time degraded by only three percent, while the Splunk indexers showed a mere 10 percent reduction in data ingest rate.[1] VAST Data Universal Storage can reduce Splunk cold storage capacity requirements and enable independent compute and storage scaling.

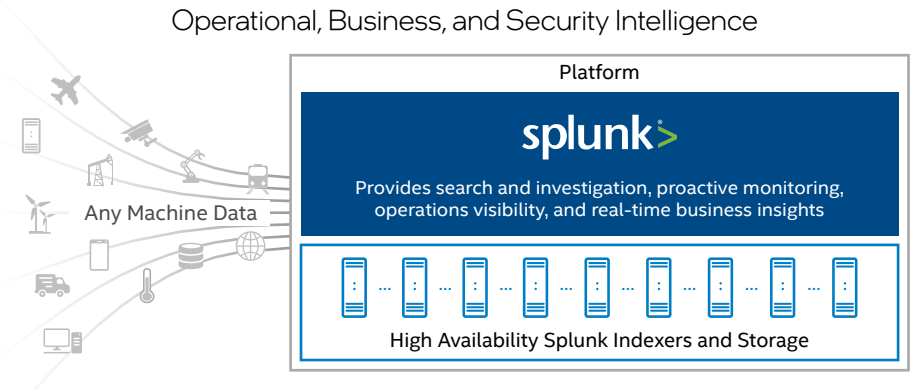


Figure 1. Massive stores of machine- and user-generated data can bring value to organizations across a broad range of applications and industries.

Business Challenge: Putting Dark Data to Work

Most companies collect more data than they can integrate, store, or use to derive insights for decision making; this underutilized data is often called dark data. A Mordor Intelligence study reveals that the dark data analytics market is expected to realize 21.7 percent compounded annual growth from 2021 through 2026.[2] Whether this data pertains to sales, manufacturing logistics, security events, or any other key area, the inability to effectively utilize data can create gaping holes in business awareness. Even if the data is ingested into an analytics tool, many organizations are not prepared to process large volumes of data in a timely manner.

This is where Splunk Enterprise can play a transformational role. With Splunk, organizations bring data together, ask questions, find answers, and take action. The power of Splunk Enterprise gives users the ability to analyze massive data sets in near real time. The data can come from many sources, including applications, devices, networks, OS, IoT sensors, and web traffic. Splunk enables organizations to collect data, develop data models, generate dashboards, and quickly deliver business insights. The resulting insights can help identify security threats, optimize application performance, understand customer behavior, uncover supply chain issues, and address many more business challenges and opportunities.

A More Efficient Way to Store Splunk Data

Splunk is powerful, especially when hundreds of users are simultaneously searching over massive data sets. Over time, though, Splunk platforms can grow to multi-petabyte data lakes. And often, enterprises build their Splunk platforms with compute and storage in a single converged system. In such cases, hot data is kept in local, high-performance storage to keep search times from spanning into multiple minutes. As the total amount of searchable data grows, storage capacity is often added via more compute nodes.

This is where VAST Data Universal Storage solutions can play an important role. VAST Data's storage solutions allow IT organizations to design disaggregated Splunk platforms, so compute and storage can scale independently. VAST Data also incorporates high-speed networking and evolves Network File System (NFS) and other storage protocols. Compute resources (VAST Data protocol servers) can be scaled separately from the VAST Data all-flash storage enclosures, built with Intel® Optane™ SSDs for low-latency writes and 44 QLC 3D NAND SSDs for high-density reads. In addition, VAST's data deduplication technology and data compression algorithms can significantly reduce the amount of Splunk data storage capacity required. To learn more, see VAST's "Breaking Data Reduction Tradeoffs with Global Compression" brief.

Continued Modernization of Intel's CIP

Advanced cyber threats continue to increase in frequency and sophistication, threatening computing environments and impacting businesses' ability to grow. Our Cyber Intelligence Platform (CIP) has significantly improved the efficiency and effectiveness of our Information Security organization in preventing, detecting, and responding to potential threats.

But CIP needs to continue to evolve. In the three years since we built our CIP, we have seen significant increases in:

- The number of Splunk users
- The number of data sources ingested
- The types of data ingested
- The number of consumers of Splunk data
- The amount of data ingested

Our CIP currently ingests over 20 TB of data per day and stores many petabytes of data in 126 Splunk indexers. Our IT team is now evaluating a variety of new products and technologies for the next generation of our CIP's architecture. Two of the most high-value options are:

- Adding container technology to support multiple Splunk indexer instances per server
- Disaggregating the platform's compute capacity from its storage capacity, allowing each to scale independently

Unfortunately, there are two potential drawbacks with compute/storage disaggregation: increased average Splunk search run time and lower Splunk indexer data ingest rates. We tested VAST Data Universal Storage to determine if it can deliver a reduction in required storage while simultaneously maintaining Splunk performance as well as enabling independent compute and storage scaling.

VAST Data at a Glance

VAST Data Universal Storage resets flash storage economics by making high-performance, remarkably dense storage affordable for applications from petabyte-scale databases to the largest data archives. Universal Storage blends flash storage technologies with an exabyte-scale file and object storage architecture.

Universal Storage combines low-cost Intel® QLC NAND SSDs and Intel Optane SSDs with stateless, containerized storage services, all connected over a low-latency NVMe fabric. This fabric is well-suited to petabyte-class storage and future scale-out. The media and fabric form the core of VAST Data's Disaggregated Shared Everything scale-out architecture. VAST then applies its own algorithms to the Disaggregated Shared Everything architecture to achieve high levels of storage efficiency, resilience, and scale. In short, Universal Storage strives to end the hard disk-centric era in data centers and eliminate the complexity of storage tiering necessitated by mechanical media trade-offs. Learn more about overall storage costs and TCO at VAST Data TCO Calculator.

Because of its wide range of enterprise-oriented features, including object storage for the cloud, VAST Data's enclosures offer a compelling alternative to conventional storage appliances. VAST Data Universal Storage can also provide container storage servers of varying types, such as being a proxy server to S3 in the cloud. For more information, see VAST Data Universal Storage Explained.

Testing VAST Data Universal Storage

To test VAST Data Universal Storage performance, Intel IT developed a test with a cybersecurity use case, searching across real-world, high cardinality, production security data. High cardinality data is not very common or unique—making it more complex—and typically has a negative impact on average Splunk search run time. For the PoC, our team utilized data from four production data sources: DNS data, firewall data, endpoint detection data, and NetFlow data.

PoC Equipment and Configurations

The PoC setup included four main components: ten Splunk indexing nodes, a VAST Data protocol server and VAST Data all-flash storage enclosure, and the network protocol used for transferring data from the Splunk indexers to the VAST solution.[3]

Test #1: Only Splunk Indexers

A typical Splunk indexer node will contain a high amount of local (SSD or HDD) storage, which stores both hot and cold data. In Test #1, and as shown in Figure 2, we utilized ten Splunk indexers. Each indexer is a two-socket server with 2nd Generation Intel® Xeon® Scalable processors.[3] For Test #1, each indexer used only one 8 TB Intel® SSD DC P4510. These drives use TLC 3D NAND media and a PCIe 3.1 x4 connection.

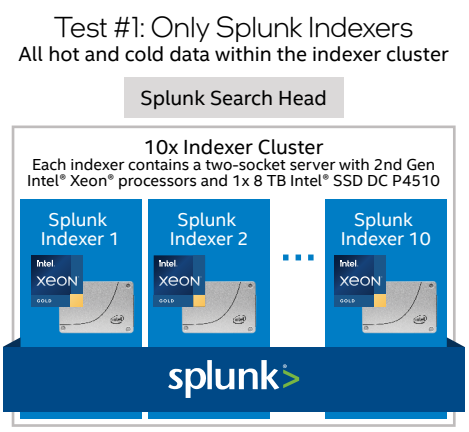


Figure 2. Our first test configuration placed all hot and cold data within the indexer cluster, reflecting a conventional hyperconverged design approach.

Test #2: Splunk Indexers plus VAST Data Universal Storage Solution

In Test #2, we wanted to evaluate the potential for disaggregating storage from compute. For example, a single replicated NVMe SSD could be allocated as a hot data store, with the remaining data pushed to remote cold storage. The intention of this configuration is to equip a node with sufficient high-performance hot storage for real-time search workloads—for example, a single work shift or the current day's data on a local NVMe storage device.

We started with the same ten Splunk indexers utilized in Test #1 with dual-socket 2nd Generation Intel Xeon processors using only one 8 TB Intel SSD DC P4510. However, as shown in Figure 3, we added three other components: a VAST Data protocol server, a VAST Data all-flash storage enclosure, and the network protocol used for transferring data from the Splunk indexers to the VAST solution.[3]

The VAST Data protocol server contains four load-balanced compute nodes that were networked to the indexers. The compute nodes support NFSv3, NFSv3/RDMA, SMB, and S3 protocols. For Intel's PoC, we used NFSv3 for connectivity.

The VAST protocol server passes data through state-of-the-art NVMe fabric switches to **one VAST all-flash storage enclosure**. This ultra-dense, all-flash storage enclosure contains 12 Intel® Optane™ SSD P4800X drives and 44 Intel® SSD D5-P4326 QLC 3D NAND drives, yielding 675 TB of capacity. The Intel Optane SSDs handle writes, while economical QLC SSDs handle reads and provide the primary capacity where data rests.

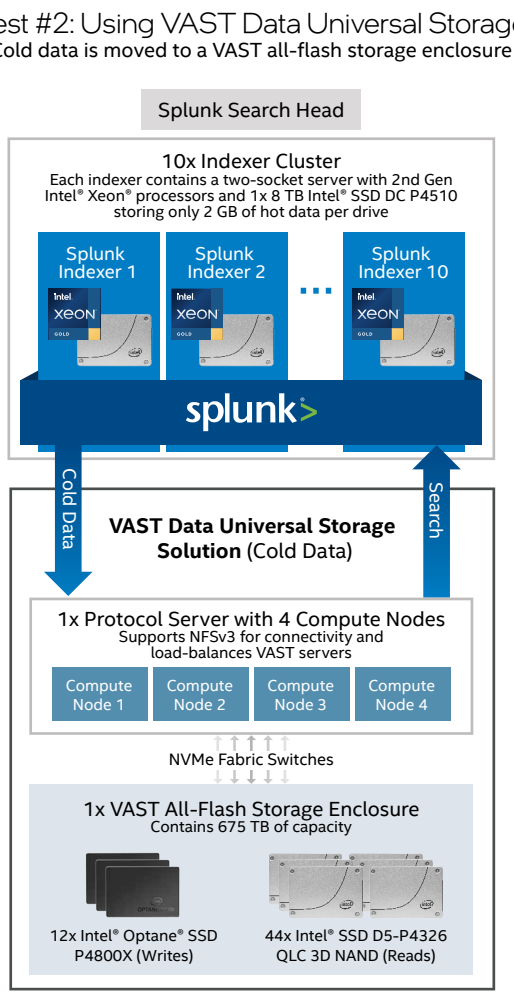
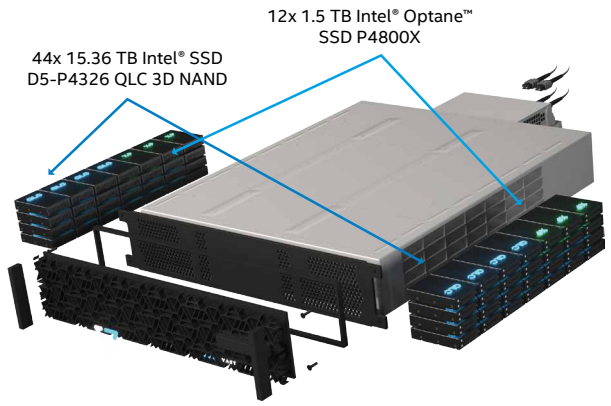


Figure 3. In Test #2, we utilized the same ten indexer nodes to ingest the production data while transferring the cold data over NFS to the VAST Data Universal Storage solution.

A Look Inside the VAST Data All-Flash Storage Enclosure

The VAST Data all-flash storage enclosure packs 56 drive bays and 4x 100 Gb Ethernet or InfiniBand connectivity into a dense 2U rack-mounted form factor.



Intel Optane media's low latency and high endurance make it uniquely qualified to work closely with VAST Data node CPUs on data blocks. Working together, the components build very wide data stripes that subsequently can be efficiently written to QLC drives. This process extends the QLC media's useful life, employs highly efficient erasure coding for data protection, and implements data reduction algorithms.

PoC Test Run Time and Search Query

Test #1 was performed using ten Splunk indexers configured with 8 TB Intel SSD P4510 drives. Each indexer was configured with a 2 GB hot volume and 8 TB cold data volume, both residing on the Intel SSD P4510 local drive. The indexers ingested an average of 5.5 TB of data per server over a 24-hour period.[3] During each test, thirty dense searches ran every minute.[4] Each of the thirty concurrent searches scanned data that was between ten and eleven minutes old. This ensured that the data being scanned by each search was residing on cold volumes on the local SSDs.

Test #2 was performed with the same data sources, utilizing the same ten Splunk indexers. Each of the ten indexers ingested an average of 5 TB of data over a 24-hour period. Each indexer used the same 2 GB hot volume on the Intel SSD P4510 local drive. But in Test #2, the cold volume was configured on an NFS mount hosted by the VAST Data appliance. Over 99 percent of the indexed data was transferred using NFS from the ten Splunk indexers to the VAST Data cluster as cold data. Like Test #1, thirty dense searches ran every minute. Each of the concurrent searches scanned data that was between ten and eleven minutes old. This ensured that the data being scanned by each search was residing on cold volumes in the VAST Data storage enclosure.

PoC Results in 2.5x Reduction in Data Size with Minimal Performance Impact

Average Splunk Search Run Time. In Test #1, with all hot and cold data stored in the ten indexer nodes, the average Splunk search run time was 25.1 seconds. In Test #2, with the cold data stored in the VAST Data all-flash storage enclosure, the average Splunk search run time was 25.9 seconds—a mere three percent slower than the "Only Splunk Indexers" configuration. Much of the reason for this near-performance parity comes from Intel Optane SSDs' very low latencies. Fast media responsiveness across the storage solution fabric all but compensates for the I/O delays normally incurred by accessing local NAND storage.

Average Splunk Indexer Data Ingest Rate. In Test #1, the Splunk indexers ingested an average of 5.5 TB per day. In Test #2, the Splunk indexers ingested an average of 5.0 TB per day. This slight degradation (10 percent) of indexer performance is due to time spent transferring the production data from the indexers over NFS to the VAST Data solution.

Splunk performance results with and without VAST Data Universal Storage are summarized in Table 1 and Figure 4.

Table 1. Splunk Performance Results With and Without VAST Data Universal Storage

	Avg. Splunk Search Run Time	Avg. Splunk Indexer Data Ingest Rate
Test #1: Only Splunk Indexers • 10 Indexers • **Data:** 10x Intel® SSD P4510 (8 TB) storing 2 GB of hot data ***and*** 8 TB of cold data	25.1 seconds	5.5 TB/Day
Test #2: Splunk Indexers plus VAST Data Universal Storage • 10 Indexers • **Hot Data:** 10x Intel P4510 SSDs (8 TB) storing *only* 2 GB hot data per drive • **Cold Data:** 1x VAST Universal Storage solution (12x Intel® Optane™ SSD P4800X and 44x Intel® SSD D5-P4326)	25.9 seconds	5.0 TB/Day

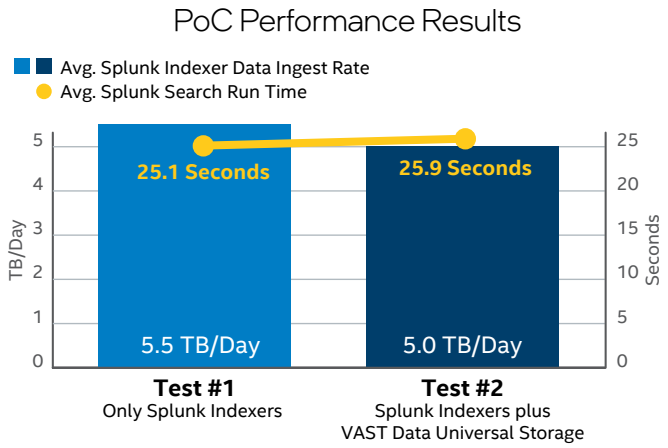


Figure 4. Transferring cold data to the VAST Data all-flash storage enclosure resulted in a slight decrease in average Splunk indexer ingest rate accompanied by a nominal increase in average Splunk search run time.

Data Reduction. Average Splunk search run time and indexer data ingest rate remained essentially constant. However, Test #2, which used Splunk indexers plus VAST Data Universal Storage, generated a **2.5x reduction in Splunk data size** (Figure 5). This reduction can provide significant benefits from a TCO perspective. For example, assume a conventional platform currently stores 10 PB of hot and cold data in SSDs on the indexers. Then, using NFS, the cold data is transferred to a VAST Data Universal Storage solution, which can reduce the size of the data by 2.5x to approximately 4 PB. This represents a **60 percent reduction of required storage capacity.**

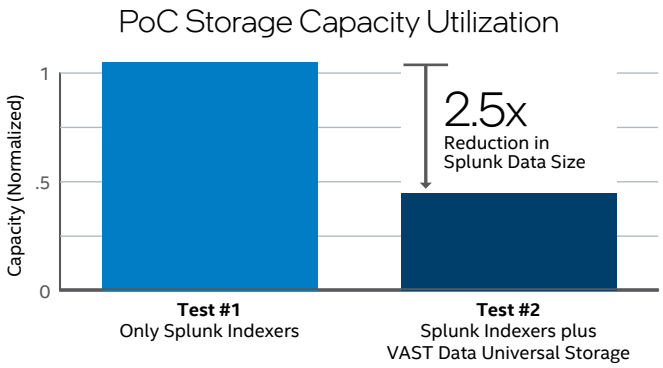


Figure 5. The VAST Data Universal Storage solution represents a 60 percent reduction in required storage capacity.

The PoC highlights the value of both VAST Data's advanced data reduction techniques and Intel Optane SSDs in the VAST Data all-flash storage enclosure. The minimal performance impact stems from this solution's disaggregation of compute from storage. Most of the compute burden normally associated with storage is handled by a VAST Data storage enclosure, leaving the indexers to their primary Splunk work. In addition, unlike in converged solutions, a configuration with the compute capacity disaggregated from storage capacity, allows IT managers to scale Splunk indexer node counts as needed without scaling storage concurrently.

Conclusion

Intel IT is always looking for new capabilities to effectively manage our data platforms while minimizing our total cost of ownership. We need agility to easily add new users, while adding and removing applications. We also need the ability to scale to support new data sources, as well as increased data—for both real-time and long-term use cases. Intel's Cyber Intelligence Platform (CIP), based on Splunk and Kafka, is one of these evolving platforms. Intel conducted a PoC of VAST Data Universal Storage as part of research into new industry-leading technologies for our next CIP architecture.

This PoC demonstrated the value of using VAST Data Universal Storage as a high-performance alternative to conventional converged infrastructure. Our PoC configuration with a VAST Data solution showed minimal impact on average Splunk search run time and average indexer data ingest rates, while generating a **2.5x reduction in Splunk data size.** In addition to this significant resource savings, we believe IT architects and engineers may also be able to reduce costs by disaggregating compute and storage, allowing each to scale independently as future demands dictate. This prevents underutilization of resources and allows more flexibility when upgrading the infrastructure components of a Splunk platform. Regardless of the data source—whether from cybersecurity, human resources, supply chain, or manufacturing—the VAST Data Universal Storage solution can help IT achieve hard-drive-class storage economics without sacrificing performance.

Learn More

You may also find the following resources useful:

- Transforming Intel's Security Posture with Innovations in Data Intelligence paper
- Intel Optane Data Center Solid State Drives
- VAST results with Splunk using synthetic data

To find the solution that is right for your organization contact your Intel representative.

[1] Splunk Enterprise with VAST Data proof of concept for disaggregated storage. **Workload:** Ten Splunk indexers configured as a cluster, with four ingest pipelines per indexer, and three Splunk search heads configured in a search head cluster. **IT production data with four data sources:** Endpoint Detection and Response (EDR), Domain Name System (DNS), Firewall, and NetFlow. **Server System Configuration:** Intel® Server System, 2x Intel® Xeon® Gold 6248 processor @ 2.5 GHz (20 cores), RAM 384 GB DDR4 @ 2900 MHz, 8 TB Intel® SSD P4510, OS = CentOS 7.5.1804, 3.10.0-1127.13.1.el7.x86_64. **Measurement Tools:** Splunk Enterprise version 8.0.6 collection tools. VAST Data Analytics collector version 3.2.1-sp1. See "PoC Equipment and Configurations" section for a description of non-VAST and VAST solution architectures. Test performed by Intel IT and Intel Data Platforms Group, February 2, 2021.

[2] Dark Analytics Market – Growth, Trends, COVID-19 Impact, and Forecasts (2021 – 2026), mordorintelligence.com/industry-reports/dark-analytics-market

[3] See endnote 1.

[4] For more information on Splunk searches visit: docs.splunk.com/Documentation/Splunk/8.0.6/Capacity/HowsearchtypesaffectSplunkEnterpriseperformance