Reference Architecture

Data Center
Splunk SmartStore and Intel® Architecture

High-Performance Splunk SmartStore Using Intel® Technologies

Achieve both scalability and high performance by combining MinIO and Intel® technology to optimize economics and operational efficiency.

Krishna Srinivas,
Engineer, MinIO
Jonathan Symonds,
Chief Marketing Officer, MinIO
Murali Madhanagopal,
Enterprise Architect, Cloud & Enterprise Solutions Group, Intel
Brien Porter,
Solutions Lead, Cloud & Enterprise Solutions Group, Intel

Table of Contents

Executive Summary

The Splunk SmartStore is an ideal choice for large, sophisticated Splunk customers. By offering a primary storage tier based on object storage, enterprises can enjoy exceptional scale, allowing them to discover more from their data and get the most from their Splunk investments.

Choosing the right SmartStore architecture is important. By combining MinIO with Intel® hardware (CPU, drives and network), enterprises can combine scale with performance, creating a model that co-optimizes economics and operational efficiency—keeping the Splunk team in the data, not waiting for it.

This reference architecture covers specific benchmark work performed by Intel and MinIO to validate performance at scale on Intel's hardware portfolio. The paper details the architecture, testing environment and results along with a discussion of how the solution would scale and perform in Splunk SmartStore environments.

Key Takeaways

The combination of Minio fast software and Intel hardware results in fast performance. The results below are done on a 10 GbE network. Throughput can be further enhanced using a 100 GbE network and adding a second network that is specifically dedicated to internode communication (see Table 3).

Upload Performance (MB/s)	**3.9 GB/s**
Download Performance (MB/s)	**4.6 GB/s**
Sparse Search with 0% in Cache	**419 seconds**
Dense Search with 0% in Cache	**460 seconds**

Why Disaggregation Matters: Splunk SmartStore

Disaggregation of storage and compute is one of the primary tools of the elite hyperscalers and has been adopted by innovative, data-driven organizations like Splunk. Specifically for Splunk, disaggregation offered the opportunity to glean more insight from more data. It expands the capabilities of the platform by separating compute (in the form of indexers) from storage (in the form of S3-compatible object storage). The Splunk SmartStore's disaggregated architecture features a distributed scale-out model that intelligently tiers data, delivering excellent economics without sacrificing performance.

The efficiency of the SmartStore model depends on the scalability and availability of the remote data. These requirements make MinIO an ideal SmartStore endpoint. Optimized for Intel® 2nd Generation Intel® Xeon® Scalable processors with Intel® Advanced Vector Extensions 512 (Intel® AVX-512) and Single Instruction Multiple Data (SIMD) instruction sets, MinIO is purpose-built for the high-performance, petascale architectures that Splunk demands.

With the optimization of each component, the overall solution benefits disproportionately from reduced costs, improved performance and greater scalability.

Flexible accommodation and reduced overhead through disaggregation can be achieved as follows:

- **Independent scalability/resource reduction.** Compute infrastructure can be sized to support the needs of high-performance analytics tools, and the storage can be sized independently to support the storage needs of the associated workload/enterprise goals.
- **Container support.** Legacy, appliance-based object stores do not support containerization, blended storage or dense compute. These are required in a modern microservices architecture to deliver elasticity and multi-tenancy.
- **Independent tiering.** With an architecture where compute and storage are separated, the architect gains the ability to create tiers of data with different performance and cost profiles.
- **Security.** By adopting storage containing best-of-breed security approaches that encrypt data in flight and at rest, enterprises can enjoy peace of mind knowing that their data is secure.

MinIO Overview

MinIO object storage is performant, cloud-native, security-enabled and resilient while offering excellent economics. Of all the SmartStore alternatives, only MinIO is deployed inside Splunk's own product suite from the highly successful Data Stream Processor (DSP) to the Data Fabric Search (DFS) product.

MinIO follows a fundamentally different architecture compared to other storage systems. Because MinIO is purpose-built to serve only objects, a single-layer architecture achieves all of the necessary functionality without compromise. The advantage of this design is an object server that is both high-performance and lightweight.

MinIO is a multi-tenant and multi-user system and is designed to scale seamlessly to any size. The tenants are fully isolated from each other with their own instances of MinIO clusters. Each tenant in turn may have multiple users with varying levels of access privileges. Each tenant's cluster operates independently of other tenants' clusters. Each cluster is a collection of fully symmetric and distributed server sets that participate equally in serving the objects. Standard HTTP load balancers or round-robin DNS may be employed.

Within a cluster, racks of homogenous servers are grouped into zones. Zones are the basic unit of expansion and they bring the concept of rack-awareness and failure-domains. A zone can be as small as four servers and as large as multiple racks. A cluster is scaled by adding one or more zones at a time and there is no rebalancing penalty for scaling. Zones also allow heterogeneous expansion of the cluster.

Inside each zone is a collection of distributed erasure-sets. Each erasure-set contains up to a maximum of 16 drives. There is no fixed limit to the number of erasure-sets within a zone or the number of zones. Objects are striped across all the drives within an erasure-set with erasure-code and bitrot protection. An erasure-set is the fundamental unit of data protection and high-availability within the data-center. Every operation in MinIO is atomic, transactional and strictly consistent. A distributed quorum lock is acquired only at the time of namespace commit at an object-level granularity. The entire cluster is designed to be heavily concurrent and resilient.

When a new object enters the system, the endpoint URL and the bucket DNS name determines the physical location of the cluster. Within a cluster, a zone with the maximum amount of free drive space is chosen and further within it, an erasure-set is chosen using the deterministic hashing algorithm. This architecture allows applications to scale geographically using the most proven practices used by the Internet applications.

Additional features of the MinIO architecture include:

- **Erasure coding.** MinIO protects data with per-object, inline erasure coding, which is written in assembly code to deliver the highest performance possible. MinIO's implementation ensures that objects can be read or new objects written even if up to half the drives are lost or unavailable. This delivers resiliency in a fraction of the disk space that traditional replication would require. MinIO's erasure code implementation delivers economic, performance and resiliency benefits over both traditional approaches and other SmartStore solutions.
- **Bitrot protection.** Silent data corruption, or bitrot, is a serious problem caused by the corruption of disk drives without the user's knowledge. MinIO's optimized implementation of the HighwayHash algorithm ensures that Splunk's indexers will never read corrupted data—it captures and heals corrupted objects on the fly. This capability is critical for the types of investigative use cases that Splunk generally runs.
- **Encryption.** It is one thing to encrypt data in flight, it is another to protect data at rest. MinIO supports multiple sophisticated server-side encryption schemes to protect data – wherever it may be. Server-side and client-side encryption are supported using AES-256-GCM, ChaCha20-Poly1305 and AES-CBC. Encrypted objects are tamper-proofed with AEAD server-side encryption. Given the exceptionally low overhead, auto-encryption can be turned on for every application and instance.
- **Continuous replication and Lambda compute.** To protect against data center failures, Splunk can be configured for multi-site indexer clustering. Multi-site indexer clustering with SmartStore requires support for cross-site replication between physical object stores. Using server-side bucket replication, MinIO can offer strict consistency guarantees—even in the face of Splunk's highly dynamic datasets.
- **Performance.** MinIO is designed for high-performance workloads and is a natural fit for Splunk SmartStore. With read/write speeds in excess of 183 GB/s and 171 GB/s respectively on Non-Volatile Memory Express (NVMe), MinIO can help teams using Splunk spend less time waiting and more time learning.[1]
- **Simplicity.** MinIO can be installed and configured within minutes simply by downloading and executing a single binary. The number of configuration options and variations

are kept to a minimum, resulting in near-zero system administration tasks and fewer paths to failures. Upgrading MinIO is done with a single command, which is non-disruptive and incurs zero downtime, lowering total cost of ownership (TCO).

- **Software-defined.** MinIO is software defined and runs on a broad range of standard hardware, meaning the enterprise can optimize its cost/performance tradeoff at the hardware level. In addition, MinIO is 100 percent open source under the Apache v2 and GNU Affero AGPL v3 licenses. This means that MinIO's customers are free from vendor lock-in, free to inspect, to innovate, to modify and to redistribute.
- **Cloud-native.** Built from scratch over the last four years, MinIO offers the highest level of interoperability with modern, cloud-native technologies such as Docker (more than 400 million pulls), Kubernetes and other microservices. This cloud orientation, coupled with MinIO's software-defined approach and open source licensing, provide enterprise-class certainty that is measured in decades.
- **No metadata datastore.** MinIO eliminates the need for a separate metadata datastore by writing data and metadata together. In addition to its scalability, this metadata design has the advantage that, in case of damage to an object, the damage can be healed/corrected for the individual object.

Better Together: Intel® Technology Accelerates MinIO Performance

MinIO takes advantage of Intel technologies to create a high-performance, high-bandwidth, and durable object storage solution. Here are examples of how MinIO uses Intel® hardware.

- **Data durability and performance boost.** MinIO protects the integrity of data with erasure coding and bitrot protection checksums. These performance-critical erasure code algorithms have been accelerated using SIMD instructions on Intel architecture using Intel® Advanced Vector Extensions 2 (Intel® AVX2) and Intel AVX-512. Offloading these calculations has a positive impact on overall system performance.

Leveraging Intel® Advanced Vector Extensions 512 (Intel® AVX-512) instructions can boost the performance of erasure coding in the MinIO server by up to 5x on 2nd Generation Intel® Xeon® Scalable processors for certain parity combinations, as compared to Intel® Advanced Vector Extensions 2 (Intel® AVX2).[2]

- **Enhanced storage performance.** SATA-based and NVMe-based Intel® 3D NAND SSDs, along with Intel® Optane™ SSDs, provide performance, stability, efficiency and low-power consumption. This reference architecture uses NVMe-based Intel 3D NAND SSDs to provide MinIO with a fast storage layer. In turn, MinIO is able to translate the performance of the NVMe SSD into significantly faster object storage read and write throughput[3].
- **Linear scaling with Intel® processors.** Testing shows that MinIO servers scale performance linearly. As MinIO clusters increase in size, so does the performance of the cluster. The Intel® Xeon® Scalable processor family provides a wide range of performance options that deliver energy efficiency, high performance for intensive storage-as-a-service (STaaS) workloads and lower performance options for less demanding STaaS workloads.
- **Increased server bandwidth.** Intel® Ethernet Network Adapters provide flexible bandwidth options. They are available with 10, 25 and 40 GbE ports to support whatever network infrastructure may be deployed in your data center racks. This reference architecture uses the Intel 40 GbE network adapters to provide fast, low-latency networking between the clients and MinIO, and between MinIO servers. MinIO can take full advantage of Intel 40 GbE networking to provide high bandwidth between the MinIO nodes and clients.

In particular, the results in the MinIO Splunk SmartStore solution took advantage of these two Intel® technologies:

- **2nd Generation Intel Xeon Scalable processors:** These are optimized for disaggregated big data analytics workloads. The processors incorporate architecture improvements and enhancements for compute-intensive and data-intensive workloads, making them well suited to the work of ingesting and analyzing massive quantities of real-time and near-real-time machine data. With Intel AVX-512 technology, MinIO can take advantage of the wider SIMD registers of 64 bytes as well as the increased total number of registers (32) to more efficiently perform calculations on the CPU for erasure coding and hashing techniques. Using Intel technology results in increased throughput and decreased latency for performance-critical workloads while at the same time freeing up CPU cores for other application tasks.[4]
- **Intel 3D NAND SSDs.** These Peripheral Component Interconnect Express (PCIe)/NVMe-based SSDs deliver scalable, cost-effective performance and low latency for Splunk Enterprise indexers. The SSDs also offer outstanding quality, reliability, advanced manageability and serviceability to minimize service disruptions.

Benchmark Architecture

The benchmarks were conducted in Intel's labs and used the following architecture:

Overview of Testing Environment

The testing environment combined hardware from Intel with software from MinIO and Splunk (see Figure 1).

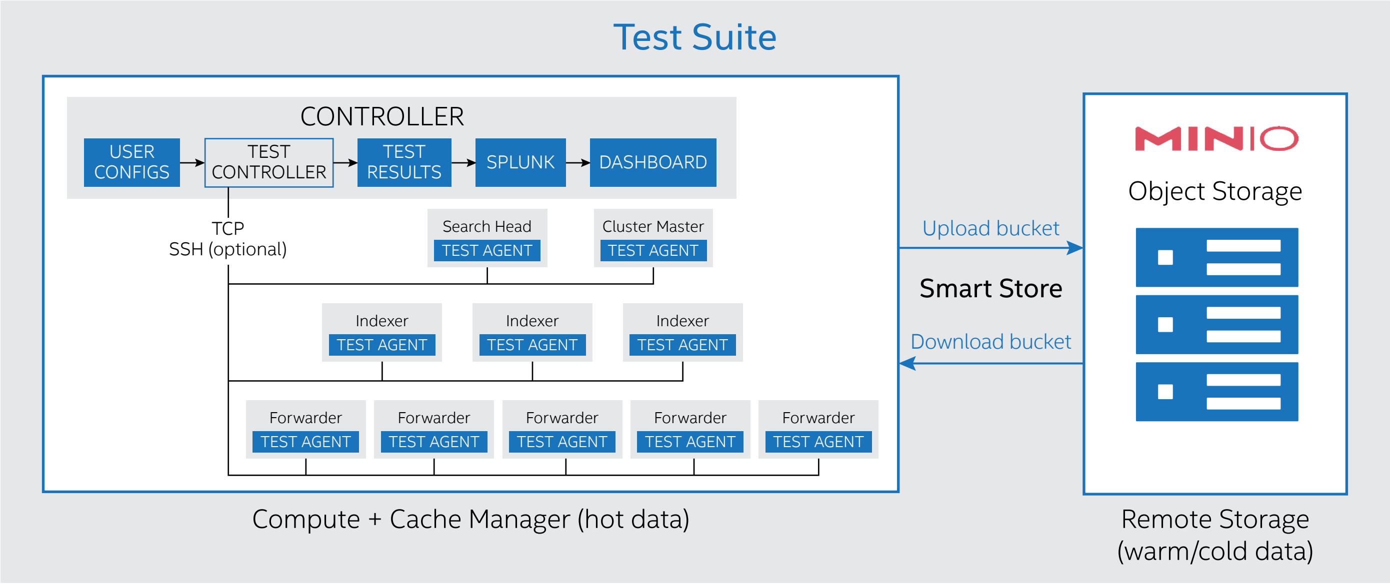

Figure 1. Testing environment.

Table 1 provides the hardware bill of materials that was used in the test.

Note that while the testing configuration used the Intel® Xeon® Gold 6248 processor, MinIO customers can also use the Intel Xeon Gold 6248R processor, which offers the same performance for a better price point.[9]

Index Tier (10 Nodes)

Component	Description	QTY
CPU	Intel® Xeon® Gold 6248 (20 cores), 2.5 GHz	2
Memory	RDIMM 32 GB DDR4-2900	12
Caching Tier	Intel® Optane™ P4800X 375 GB PCIe x4 U.2	2
Capacity Tier	Intel® SSD P4510 8TB 2.5" Peripheral Component Interconnect Express (PCIe) Non-Volatile Memory Express (NVMe)	1
Boot Device	Intel® SSD S4510 960 GB 2.5" SATA	1
Network Interface Card (NIC)	Intel® Ethernet Network Connection OCP X527-DA4 (10 GbE)	1

Search Tier (3 Nodes)

Component	Description	QTY
CPU	Intel Xeon Gold 6248 (20 cores), 2.5 GHz	2
Memory	RDIMM 32 GB DDR4-2900	12
Caching Tier	Intel Optane P4800X 375 GB PCIe x4 U.2	2
Capacity Tier	Intel SSD P4510 8TB 2.5" PCIe NVMe	1
Boot Device	Intel SSD S4510 960 GB 2.5" SATA	1
NIC	Intel Ethernet Network Connection OCP X527-DA4 (10 GbE)	1

Table 1. Hardware Bill of Materials

MinIO Object Storage (8 nodes)

Component	Description	QTY
CPU	Intel Xeon Gold 6248 (20 cores), 2.5 GHz	2
Memory	RDIMM 32 GB DDR4-2900	12
Caching Tier	Intel Optane P4800X 375 GB PCIe x4 U.2	2
Capacity Tier	Intel SSD P4510 8TB 2.5" PCIe NVMe	6
Boot Device	Intel SSD S4510 960 GB 2.5" SATA	1
NIC	Intel Ethernet Network Connection OCP X527-DA4 (10 GbE)	1

Data Generation Tier (10 nodes)

Component	Description	QTY
CPU	Intel® Xeon® E5-2699 v4 (22 cores), 2.2 GHz	2
Memory	RDIMM 32GB DDR4-2666	12
Capacity Tier	Intel SSD P4510 4 TB 2.5" NVMe U.2	2
Boot Device	Intel SSD S4510 480 GB 2.5" SATA	1
NIC	Intel Ethernet Connection OCP X527-DA4 (10 GbE)	1

Management Nodes (two nodes)
1. Cluster Master
2. Management Controller

Component	Description	QTY
CPU	Intel Xeon E5-2699 v4 (22 cores), 2.2 GHz	2
Memory	RDIMM 32GB DDR4-2666	6
Capacity Tier	Intel SSD P4510 4 TB 2.5" NVMe U.2	1
Boot Device	Intel SSD S4510 480 GB 2.5" SATA	1
NIC	Intel Ethernet Network OCP X527-DA4 (10 GbE)	1

Table 2 provides details about software components, versions and configurations that were used in the test.

Component	Description
Splunk	Splunk 8.0.0 Build 1357bef0a7f6 Search Heads: 3 Indexers: 10 Masters: 1 Forwarder Machines: 10 Forwarders Per Machines: 10
MinIO	MinIO Server 2020-04-15T19:42:18Z MinIO Sidekick v0.1.8
Operating System	CentOS 7.6.1810

Table 2. Software Stack

The network was 10 Gbps from the indexers to the switch and from the switch to the MinIO servers (see Figure 2). The bandwidth available to the switches was 80 Gbps.

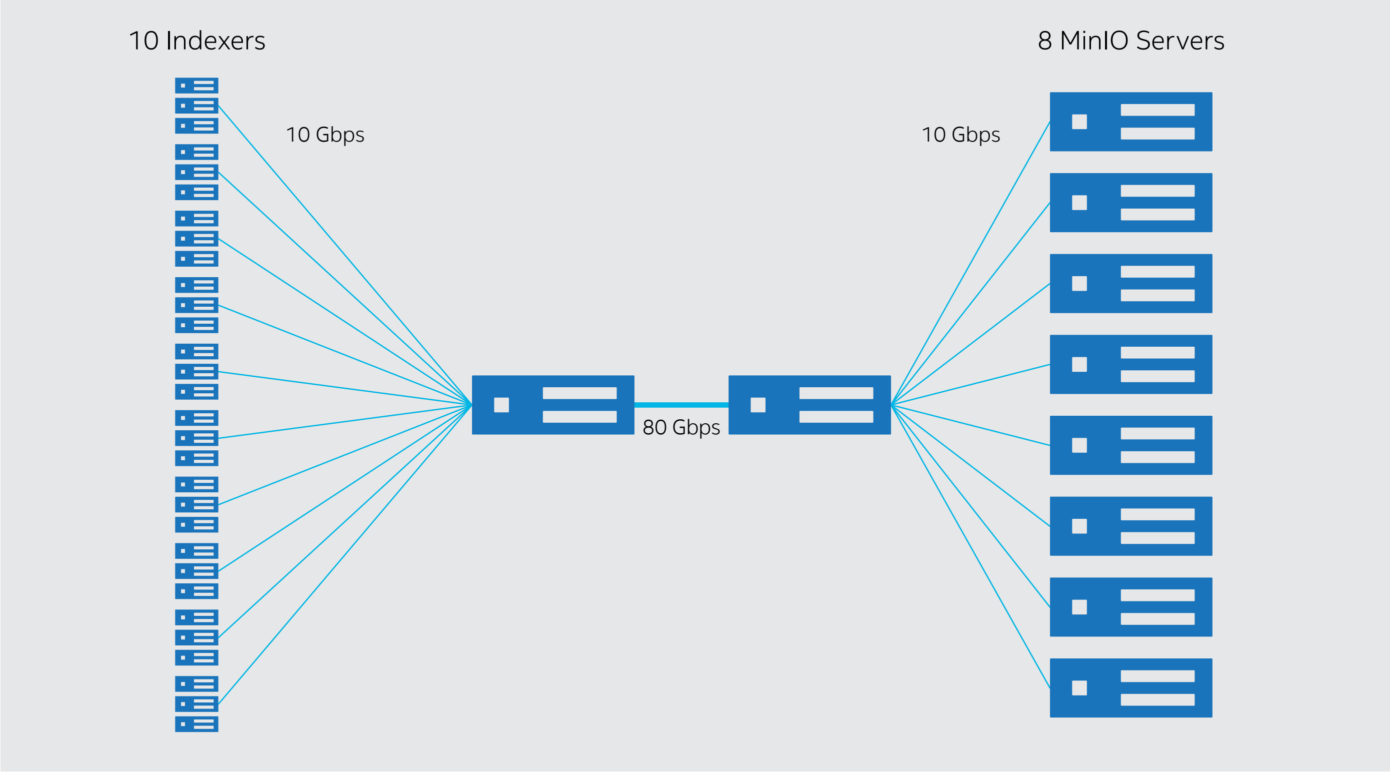

Figure 2. Network layout.

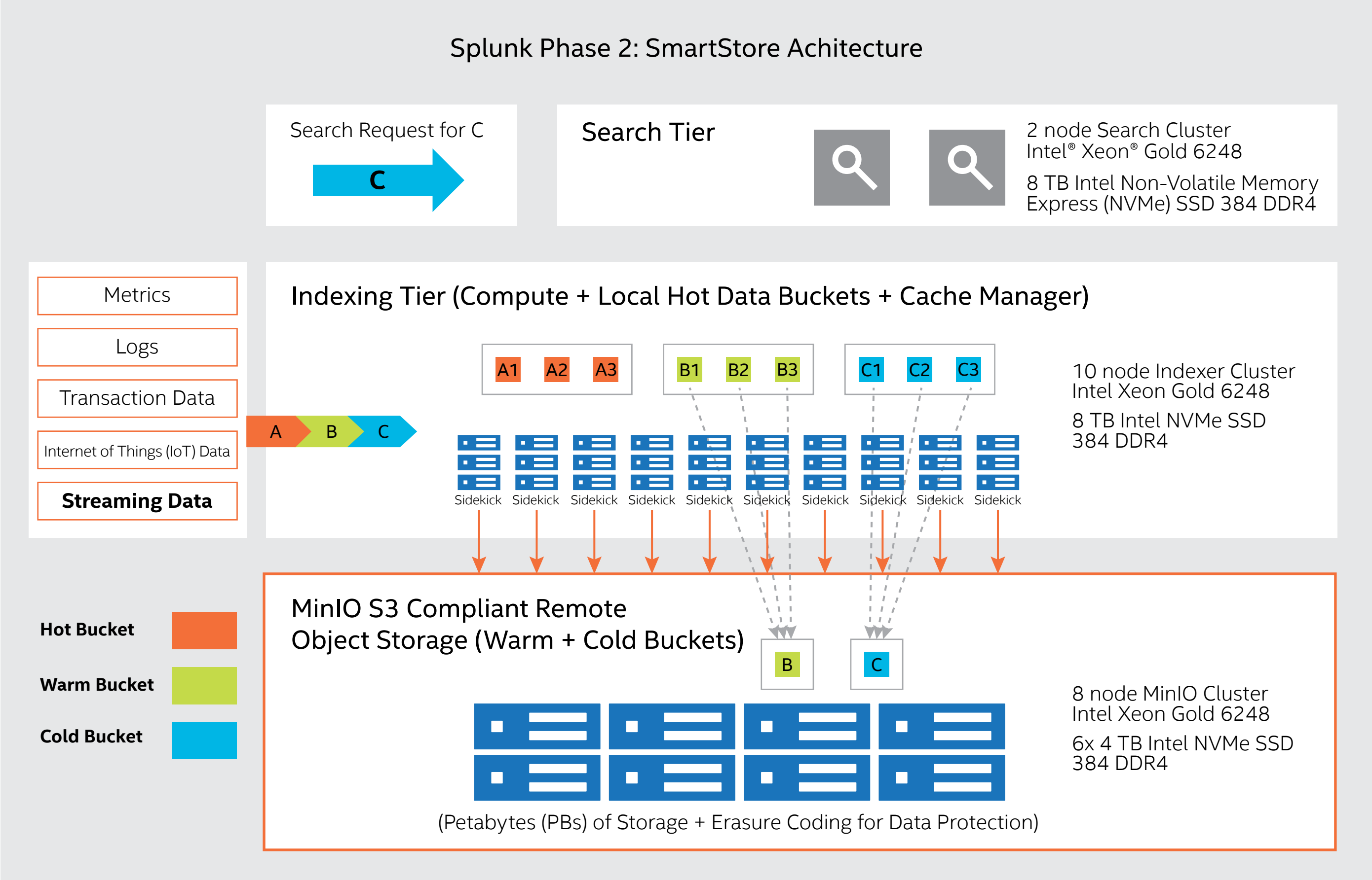

Figure 3. Solution architecture.

Benchmarking Object Storage Performance Using Warp

Before the evaluation, each hardware component was tested using Warp, an open-source S3 benchmarking tool.

The following parameters were used:

```
$ warp get --duration=20m --warp-client=10.105.185.
{33...38},10.105.185.{41...44} --host=localhost:9000
--access-key=minio --secret-key=minio123 --obj.size=100m
--concurrent 64 --objects 5000
```

Here is the Warp output:

```
Operation: PUT. Concurrency: 640. Hosts: 1.

* Average: 4097.94 MiB/s, 42.97 obj/s (18m17.767s,
starting 18:22:26 PDT)

Aggregated Throughput, split into 1097 x 1s time
segments:

 * Fastest: 4161.0MiB/s, 43.63 obj/s (1s, starting
18:22:43 PDT)

 * 50% Median: 4097.8MiB/s, 42.97 obj/s (1s, starting
18:33:18 PDT)

 * Slowest: 4031.9MiB/s, 42.28 obj/s (1s, starting
18:22:36 PDT)

--------------------

Operation: GET. Concurrency: 640. Hosts: 1.

* Average: 4747.38 MiB/s, 49.78 obj/s (18m36.538s,
starting 18:42:23 PDT)

Aggregated Throughput, split into 1116 x 1s time
segments:

 * Fastest: 5395.1MiB/s, 56.57 obj/s (1s, starting
18:57:36 PDT)

 * 50% Median: 4749.1MiB/s, 49.80 obj/s (1s, starting
18:49:12 PDT)
```

The throughput figures from Warp are close to the throughput numbers seen by the indexers, indicating that everything is as expected from a performance perspective.

> The takeaway from this work is that MinIO's object store should be able to deliver sustained aggregated upload throughput of up to 4,097.94 MB/s and download throughput of up to 4,747.38 MB/s given the hardware provided.

Benchmarking MinIO/Splunk SmartStore Performance

Per the test plan, all benchmarking was completed before the test suite began.

Step 1: Ingestion Performance

The indexing rate for each indexer was set to 5 TB/indexer/day, running indexing on the 10-node-cluster for one day with measurements taken on the indexing rate/indexer and aggregate indexing rate. This resulted in an overall capacity of 33 TB per day. This produced results on how much data is written to MinIO and how much was retained in the indexer cache.

Next, all hot buckets were rolled to warm and pushed to remote storage.

Step 2: Upload Performance (Steady/Peak):

The steady-state bucket upload throughput to MinIO was then measured. Indexing was performed for one day and stored locally on the indexer. The SmartStore was then added to the indexes.conf file and all buckets were uploaded at the same time, measuring maximum bucket upload throughput.

Step 3: Download Performance

All data was evicted from all of the indexers followed by the execution of searches that require all data to be downloaded. This allowed for the measurement of bucket download throughput.

Step 4: Search Performance

Both dense and sparse searches were executed when the data was present in cache as well as when the data was not present. Additionally, search performance was measured at 25 percent, 50 percent, and 75 percent data in cache.

Step 5: Scale Testing

To test scale, the cache size on the SmartStore was configured to be the size of one day's worth of indexer data. Data was ingested for a second day at 5 TB/day index rate and upload throughput was measured. Searches were then executed to measure cache hits/misses as well as download throughput.

Configuration Details

Sidekick is a sidecar load-balancer that runs on the indexer nodes. Every Sidekick instance is configured to load balance the local indexer's traffic across the MinIO nodes. The Sidekick instances are run as:

```
$ sidekick --address :9000 http://minio{1...9}:9000
```

Sidekick can be downloaded from github.com/minio/sidekick.

Splunk SmartStore cache manager is configured to communicate with Sidekick running on the same server. Sidekick handles distributing and load balancing the cache manager's requests across the MinIO servers.

Configuring the Splunk SmartStore to connect with MinIO is straightforward. First, configure MinIO and provide the secret keys. Second, point MinIO to the Splunk instance.
The configuration is provided below:

```
[volume:s3]
storageType = remote
path = s3://smartstore/
remote.s3.access_key = minio
remote.s3.secret_key = minio123
remote.s3.supports_versioning = false
remote.s3.Client endpoint device = http://
localhost:9000
remote.s3.list_objects_version = v2

[INDEXNAME - TBD]
remotePath = volume:s3/$_index_name
homePath = $SPLUNK_DB/minio/db
coldPath = $SPLUNK_DB/minio/colddb
thawedPath = $SPLUNK_DB/minio/thaweddb
```

It should be noted that in this case remote.s3.endpoint is set to http://localhost:9000 so that the cache manager's requests go through the Sidekick load balancer.

Test Suite Description

Intel ran a number of tests designed to evaluate system performance. The performance numbers gathered with this suite are meant to serve as comparison points between different environments, workloads or Splunk configurations. For example, this kit could be used to gauge how much of a performance boost can be achieved by moving from 12-core machines to 24-core machines, or by doubling the number of indexers.

The performance test creates a steady flow of data and a consistent scheduled search load. It exposes the following parameters in the YAML files and/or on the command line, though this is not a comprehensive list:

- **test_duration_s.** The duration of the test specified in seconds.
- **indexer_gb_per_day.** Data volume per indexer, specified in GB/day.
- **indexers.** Number of indexers. If this is three or greater, then indexer clustering will be used.
- **search_heads.** Number of search heads. If this is three or greater, then search head clustering will be used.
- **forwarder_machines.** Number of machines used to handle data generation, and to run Splunk forwarders.
- **forwarders_per_machine.** Number of universal forwarders used to run on each forwarder machine, where data generation occurs.
- **scheduled_searches.** Includes the following parameters:
  - **max_searches.** Number of scheduled searches to create.
  - **search_string.** Search string for every scheduled search. This search string can be modified to run dense or super-sparse searches, as follows (sparse is also an option, but was not used in testing):
    - Dense search (CPU-bound): "index=main every1 earliest=-1m | stats distinct_count(id)"
    - Super-sparse search (I/O-bound): "index=main every100k earliest=-1h | stats distinct_count(id)"
  - cron_schedule. Cron schedule of each scheduled search.

The following test metrics were used to measure the indexing and search performance:

- Search runtime
- Ingestion rate

Data Ingestion
Forwarders generate syslog data using a customized version of the publicly available Eventgen application (https://splunkbase.splunk.com/app/1924/). The customized version allows you to define generator parameters in the test's main YAML file instead of the Eventgen app's own .conf file. This adds options for density markers for syslog-style data, and allows an external process (in this case, the test) to start and stop Eventgen on demand.

Interpretation of Results
Table 3 shows the test results. The test configuration consisted of 10 indexers, 2 search heads, 8 MinIO nodes, and 48 3.6-TB NVMe SSDs.

Data Ingestation Test Results

Ingestion	Ingest Performance (MB/s)
Aggregate Throughput	170
Average/Indexer	17
Peak/Indexer	19.2
Upload	Upload Performance (MB/s)
Aggregate Throughput	3,925
Average/Indexer	390
Peak/Indexer	750
Download	Download Performance (MB/s)
Aggregate Throughput	4,608
Average/Indexer	460
Peak/Indexer	585
Sparse Search	Sparse Search Performance (MB/s)
100% in Cache	2.2
75% in Cache	125.1
50% in Cache	241.4
25% in Cache	399.4
0% in Cache	419.2
Dense Search	Dense Search Performance (MB/s)
100% in Cache	204.5
75% in Cache	285.2
50% in Cache	376
25% in Cache	432
0% in Cache	460.5

Table 3. Data Ingestion Test Results.[5]

As the results suggest, the system performance scales in a linear fashion. This is critical as Splunk SmartStores can rapidly grow to multiple petabytes (PBs). MinIO currently supports SmartStore in excess of 40 PBs for some of its largest customers.

To scale the solution, the architect must design a system that benefits from the hard-won knowledge of other hyperscalers—a system that is disaggregated, simple, secure and built from smaller components.

There are a number of different price/performance tradeoffs that will need to be determined by the architect—however, approaching the problem of infrastructure as if it were code is a key tenant in any design.

Upgrading for Additional Performance

While the performance of the hardware and software operated largely at their theoretical limits, there were architectural changes that would have enhanced the performance even more.

Specifically, it became apparent that the network interfaces on the MinIO servers were saturated at 10 Gbps when the SmartStore's cache manager downloaded the buckets from the MinIO server following indexer cache eviction. This was confirmed with the vnstat tool, which measures the network throughput at the interface. The vnstat command was run as vnstat -l -i <interface>.

Given the speed of the drives and the software, the network will continue to represent a bottleneck—even at 100 GbE (just six drives per server will saturate a 100 GbE network).[6]

> While 100 GbE will deliver outstanding performance (estimated to be 40-45 GBps), only by adding a second Network Interface Card (NIC) for internode traffic would one be able to move the bottleneck away from the network.

Splunk, MinIO and Intel: Powering an Analytical Revolution

Splunk SmartStore is an ideal choice for at-scale Splunk Enterprise customers. Designed correctly, it offers strengthened security, seamless scale and excellent economics.

Predicated on the concept of disaggregation, SmartStores built with Intel technology and MinIO offer the following benefits:

- The ability to combine exceptional performance on standard hardware
- Simplicity of management
- Outstanding data resiliency
- The ability to scale to dozens of PBs while retaining the performance of colocation.

These benefits are clearly demonstrated in the benchmark results. MinIO's Intel AVX-512-tuned software, paired with fast Intel Optane SSDs, easily saturated the network. This performance is a direct result of tight integration between the software and the hardware.

The benchmarks show that MinIO is saturating the 10 GbE network performance.[7] In this deployment, MinIO shares a single flat network for both the internode traffic (distributed erasure coding) and client-server traffic. This is a common scenario at many customer sites.

For NVMe SSDs, this work strongly suggests employing a 100 GbE network—potentially two. One would expect to see a 10x improvement, similar to what MinIO saw in its NVMe benchmark.[8]

---

[1] https://min.io/

[2] 2x Intel® Xeon® 6248 processor: 80 logical processors (40 cores, 80 threads); DRAM: 384 GB (32 GB x 12 DIMMs); Network: 2x Ethernet Connection X722 for 10GBASE-T; Storage: device-under-test: 4 x 8TB P4510 NVMe Intel® SSD; OS Software: CentOS Linux release 7.8.1127(Core) Test Software: MinIO Server version 2020-04-15T19:42:18Z. Benchmark Used: Reed Solomon benchmark. Baseline Config: Intel® Advanced Vector Extensions 2 (Intel® AVX2) Improved Configuration: Intel® AVX-512. Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors.

[3] See endnote 2 and Table 3

[4] See endnote 2 and Table 3.

[5] See endnote 2 (AVX-512 config).

[6] See MinIO's test results at https://min.io/resources/docs/MinIO-Throughput-Benchmarks-on-NVMe-SSD-8-Node.pdf.

[7] See endnote 2 and table 3.

[8] See endnote 6

[9] Cost and performance data as of September 28, 2020.

Xeon 6248: https://ark.intel.com/content/www/us/en/ark/products/192446/intel-xeon-gold-6248-processor-27-5m-cache-2-50-ghz.html

Xeon 6248R: https://ark.intel.com/content/www/us/en/ark/products/199351/intel-xeon-gold-6248r-processor-35-75m-cache-3-00-ghz.html