NetApp StorageGRID with Splunk SmartStore

NetApp Solutions

NetApp
August 05, 2022

Table of Contents

NetApp StorageGRID with Splunk SmartStore

TR-4869: NetApp StorageGRID with Splunk SmartStore

Karthikeyan Nagalingam, Bobby Oommen, Joseph Kandatilparambil

Splunk Enterprise is the market-leading Security Information and Event Management (SIEM) solution that drives outcomes across the Security, IT, and DevOps teams. Data volumes continue to grow at exponential rates, creating massive opportunities for enterprises that can leverage this vast resource. Splunk Enterprise continues to gain adoption across a wider variety of use cases. As the use cases grow, so does the amount of data that Splunk Enterprise ingests and processes. The traditional architecture of Splunk Enterprise is a distributed scale-out design providing excellent data access and availability. However, enterprises using this architecture are faced with growing costs associated with scaling to meet the rapidly growing volume of data.

Splunk SmartStore with NetApp StorageGRID solves this challenge by delivering a new deployment model in which compute and storage is decoupled. This solution also unlocks unmatched scale and elasticity for Splunk Enterprise environments by allowing customers to scale across single and multiple sites, all while reducing costs by allowing compute and storage to scale independently and adding intelligent tiering to cost-effective cloud- based S3 object storage.

The solution optimizes the amount of data in local storage while maintaining search performance, allowing compute and storage to be scaled on demand. SmartStore automatically evaluates data access patterns to determine which data needs to be accessible for real-time analytics and which data should reside in lower-cost S3 object storage.

This technical report outlines the benefit NetApp provides to a Splunk SmartStore solution while demonstrating a framework for designing and sizing Splunk SmartStore in your environment. The result is a simple, scalable, and resilient solution that delivers a compelling TCO. StorageGRID provides the scalable and cost-effective S3 protocol/API-based object storage, also known as remote storage, allowing organizations to scale their Splunk solution at a lower cost while increasing resiliency.

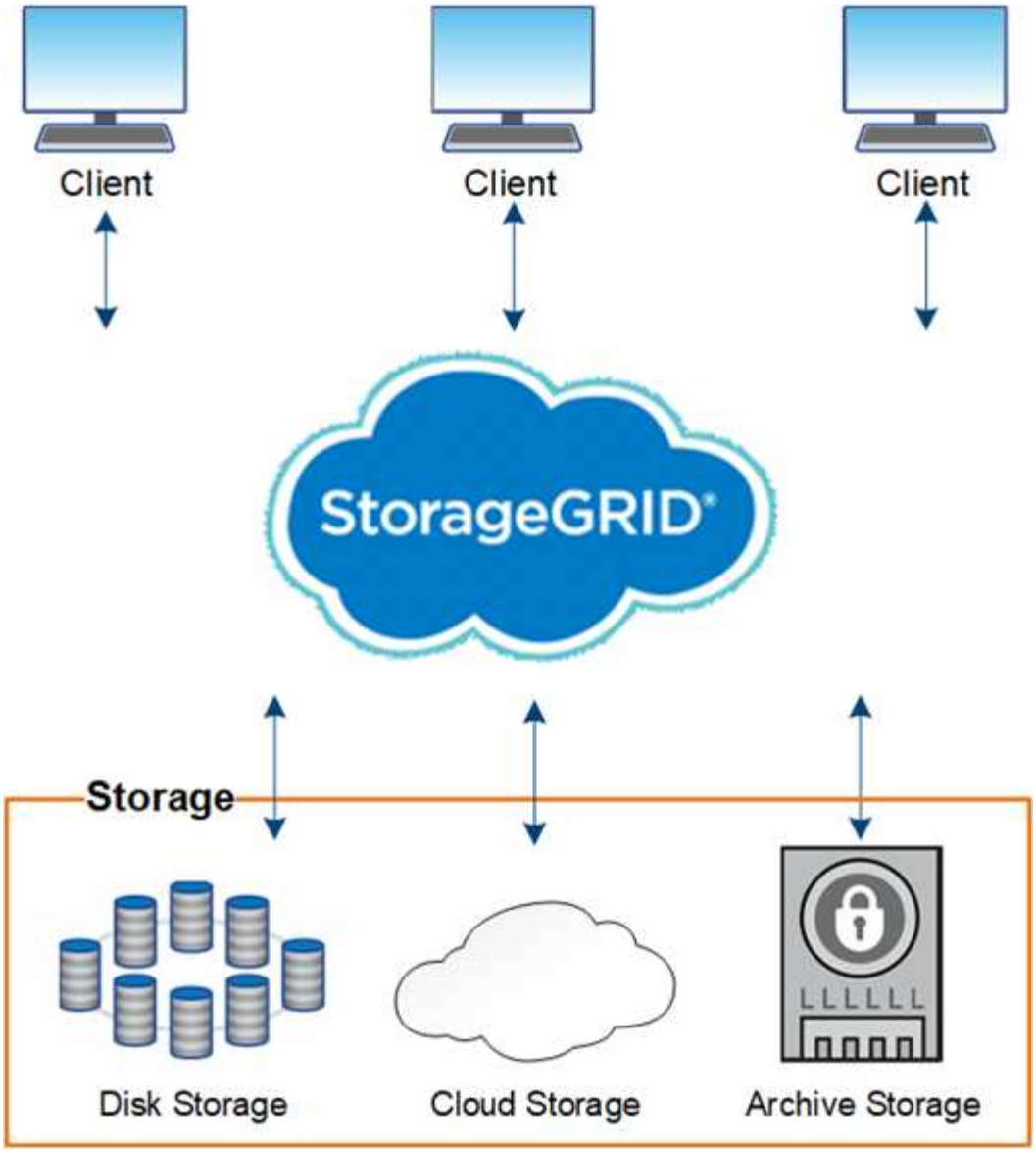
Splunk SmartStore refers to object storage as remote stores or remote storage tiers.

About NetApp StorageGRID

NetApp StorageGRID is a software-defined object storage solution for large archives, media repositories, and web data stores. With StorageGRID, NetApp leverages two decades of experience in delivering industry-leading innovation and data management solutions while helping organizations manage and maximize the value of their information both on-premises and in public, private, or hybrid cloud deployments.

StorageGRID provides secure, durable storage for unstructured data at scale. Integrated, metadata-driven lifecycle management policies optimize where your data lives throughout its life. Content is placed in the right location, at the right time, and on the right storage tier to reduce cost. The single namespace allows the data to be accessed via a single call regardless of geographical location of the StorageGRID storage. Customers can deploy and manage multiple StorageGRID instances between datacenters and in the cloud infrastructure.

A StorageGRID system is composed of globally distributed, redundant, heterogeneous nodes that can be integrated with both existing and next-generation client applications.



IDC MarketScape recently named NetApp as a leader in the latest report, IDC MarketScape: Worldwide Object-Based Storage 2019 Vendor Assessment. With nearly 20 years of production deployments in the most demanding industries, StorageGRID is a recognized leader in unstructured data.

With StorageGRID, you can achieve the following:

- Deploy multiple StorageGRID instances to access data from any location between data centers and the cloud through a single namespace that easily scales to hundreds of petabytes.
- Provide flexibility to deploy and centrally manage across infrastructures.
- Provide unmatched durability with fifteen-nines of durability leveraging layered Erasure Coding (EC).
- Enable more hybrid multi-cloud capabilities with validated integrations into Amazon S3 Glacier and Azure Blob.
- Meet regulatory obligations and facilitate compliance through tamper-proof data retention, without proprietary APIs or vendor lock-in.

For more information about how StorageGRID can help you solve your most complex unstructured data management problems, see the NetApp StorageGRID homepage.

About Splunk Enterprise

Splunk Enterprise is a platform for turning data into doing. Data generated by various sources such as log files, websites, devices, sensors, and applications are sent to and parsed by the Splunk Indexers, allowing you to derive rich insights from the data. It might identify data breaches, point out customer and product trends, find opportunities to optimize infrastructure, or create actionable insights across a wide variety of use cases.

About Splunk SmartStore

Splunk SmartStore expands on the benefits of the Splunk architecture while simplifying its ability to scale cost-effectively. The decoupling of compute and storage resources results in indexer nodes optimized for I/O with significantly reduced storage needs because they only store a subset of data as cache. You do not have to add extra compute or storage when only one of those resources is necessary, which allows you to realize significant cost savings. You can use cost-effective and easily scalable S3-based object storage, which further simplifies the environment, reduces costs, and allows you to maintain a more massive data set.

Splunk SmartStore delivers significant value to organizations, including the following:

- Lowering storage cost by moving warm data to cost-optimized S3 object storage
- Scaling seamlessly by decoupling storage and compute
- Simplifying business continuity by leveraging resilient cloud-native storage

Next: Benefits of this solution.

Benefits of this solution

Previous: Introduction.

- **Performance.** The combination of Splunk SmartStore and NetApp StorageGRID provides fast migration of data between hot buckets and warm buckets using object storage. StorageGRID turbocharges the migration process by providing fast performance for large object workloads.
- **Multisite ready.** The StorageGRID distributed architecture allows Splunk SmartStore to extend deployments across single and multiple sites through a single global namespace where data can be accessed from any site regardless of where the data lives.
- **Improved scalability.** Scale storage resources independently from compute resources to meet evolving needs and demands in your Splunk environment, thereby providing improved TCO.
- **Capacity.** Meet rapidly growing volumes in Splunk deployment with StorageGRID by scaling a single namespace to over 560PB.
- **Data availability.** Optimize data availability, performance, geo-distribution, retention, protection, and storage costs with metadata-driven policies that can adjust dynamically as the business value of your data evolves.

Increase performance with the SmartStore cache, which is a component of the indexer that handles the transfer of bucket copies between local (hot) and remote (warm) storage. Splunk sizing for this solution is based on the guidelines provided by Splunk. The solution allows adding compute, hot storage, or S3 resources to meet the growing demand in terms of the number of users or ingest rate across single and multisite deployments.

Next: Intelligent tiering and cost savings.

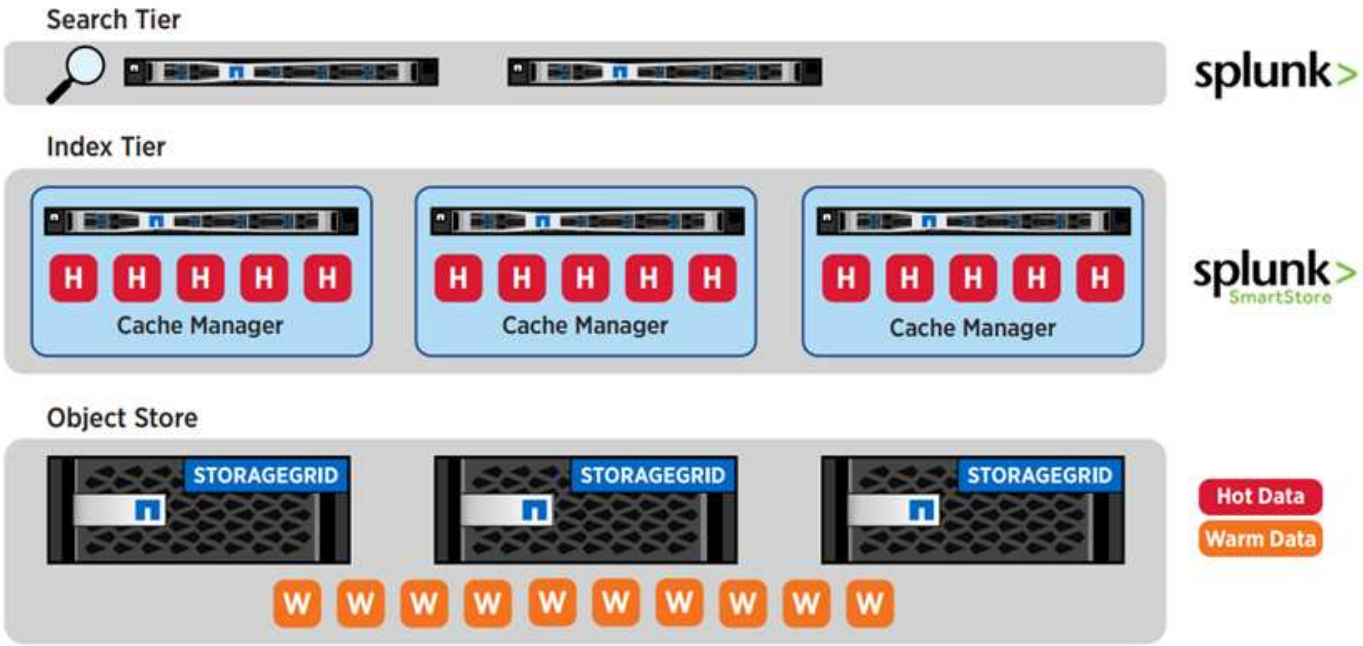
Intelligent tiering and cost savings

Previous: Benefits of this solution.

As customers realize the power and ease of using Splunk data analytics, they naturally want to index an ever-growing amount of data. As the amount of data grows, so does the compute and storage infrastructure required to service it. Since older data is referenced less frequently, committing the same amount of compute resources and consuming expensive primary storage becomes increasingly inefficient. To operate at scale, customers benefit from moving warm data to a more cost-effective tier, freeing compute and primary storage for hot data.

Splunk SmartStore with StorageGRID offers organizations a scalable, performant, and cost-effective solution. Because SmartStore is data-aware, it automatically evaluates data-access patterns to determine which data needs to be accessible for real-time analytics (hot data) and which data should reside in lower-cost long-term storage (warm data). SmartStore uses the industry-standard AWS S3 API dynamically and intelligently, placing data in S3 storage provided by StorageGRID. The flexible scale-out architecture of StorageGRID allows the warm data tier to grow cost effectively as needed. The node-based architecture of StorageGRID makes sure that performance and cost requirements are met optimally.

The following image illustrates the Splunk and StorageGRID tiering.



The industry-leading combination of Splunk SmartStore with NetApp StorageGRID delivers the benefits of decoupled architecture through a full-stack solution.

Next: Solution overview.

Solution overview

Previous: Intelligent tiering and cost savings.

NetApp StorageGRID

NetApp StorageGRID is a high-performance and cost-effective object storage platform. It offers intelligent,

policy-driven global data management using a distributed, node-based grid architecture. It simplifies the management of petabytes of unstructured data and billions of objects through its ubiquitous global object namespace combined with sophisticated data management features. Single-call object access extends across sites and simplifies high availability architectures while ensuring continual object access regardless of site or infrastructure outages.

Multitenancy allows multiple cloud and enterprise unstructured data applications to be securely serviced within the same grid, increasing the ROI and use cases for StorageGRID. Multiple service levels can be created with metadata-driven object lifecycle policies, optimizing durability, protection, performance, and locality across multiple geographies. Users can adjust policies and realign the data landscape non-disruptively as their requirements change.

SmartStore leverages StorageGRID as the remote storage tier and allows customers to deploy multiple geographically distributed sites for robust availability and durability, presented as a single object namespace. This allows Splunk SmartStore to take advantage of the StorageGRID high performance, dense capacity, and ability to scale to hundreds of nodes across multiple physical sites using a single URL to interact with the objects. This single URL also allows storage expansion, upgrades, and repairs to be nondisruptive, even beyond a single site. The StorageGRID unique data management policy engine provides optimized levels of performance and durability and adherence to data locality requirements.

Splunk Enterprise

Splunk, a leader in the collection and analysis of machine-generated data, helps simplify and modernize IT thru its operational analytics capabilities. It also expands into business analytics, security, and IoT use cases. Storage is a critical enabler for a successful Splunk software deployment.

Machine-generated-data is the fastest-growing type of big data. The format is unpredictable and comes from many different sources, often at high rates and in great volumes. These workload characteristics are often referred to as digital exhaust. Splunk SmartStore helps to make sense of this data and provides smart data tiering for optimized placement of hot and warm data on the most cost-effective storage tier.

Splunk SmartStore

Splunk SmartStore is an indexer capability that uses object storage (also referred to as remote storage or remote storage tiers) such as StorageGRID to store warm data using the S3 protocol.

As a deployment’s data volume increases, demand for storage typically outpaces demand for computer resources. SmartStore allows you to manage your indexer storage and compute resources cost-effectively by scaling compute and storage separately.

SmartStore introduces a remote storage tier, using the S3 protocol, and a cache manager. These features allow data to reside either locally on indexers or remote storage. The cache manager, which resides on the indexer, manages data movement between the indexer and the remote storage tier. Data is stored in buckets (hot and warm) along with bucket metadata.

With SmartStore, you can reduce the indexer storage footprint to a minimum and choose I/O-optimized compute resources because most data resides on the remote storage tier. The indexer maintains a local cache, representing the minimal amount of data necessary to return the results requested and predicted. The local cache contains hot buckets, copies of warm buckets participating in active or recent searches, and bucket metadata.

Splunk SmartStore with StorageGRID enables customers to incrementally scale the environment with high-performance and cost-effective remote storage while providing a high degree of elasticity to the overall solution. This allows customers to add any components (hot storage and/or warm S3 storage) in any given quantity at any given time, whether they need more indexers, change data retention, or to increase the ingest

rate without any disruption.

Next: Flexible StorageGRID features for Splunk SmartStore.

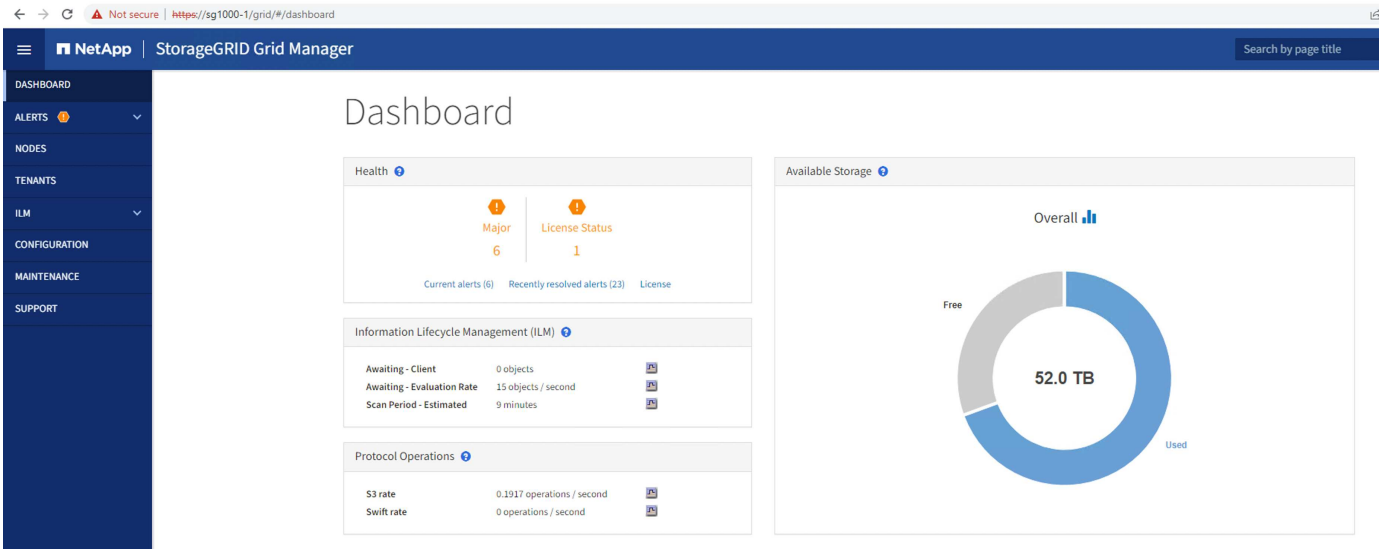
Flexible StorageGRID features for Splunk SmartStore

Previous: Solution overview.

StorageGRID has a wide variety of features that users can leverage and customize for their ever-changing environment. From deploying to scaling your Splunk SmartStore, your environment demands rapid adoption to changes and should be nondisruptive to Splunk. The StorageGRID flexible data management policies (ILM) and traffic classifiers (QoS) let you plan and adapt to your environment.

Simple management with Grid Manager

Grid Manager is the browser-based graphical interface that allows you to configure, manage, and monitor your StorageGRID system across globally distributed locations in a single pane of glass, as shown in the following image.



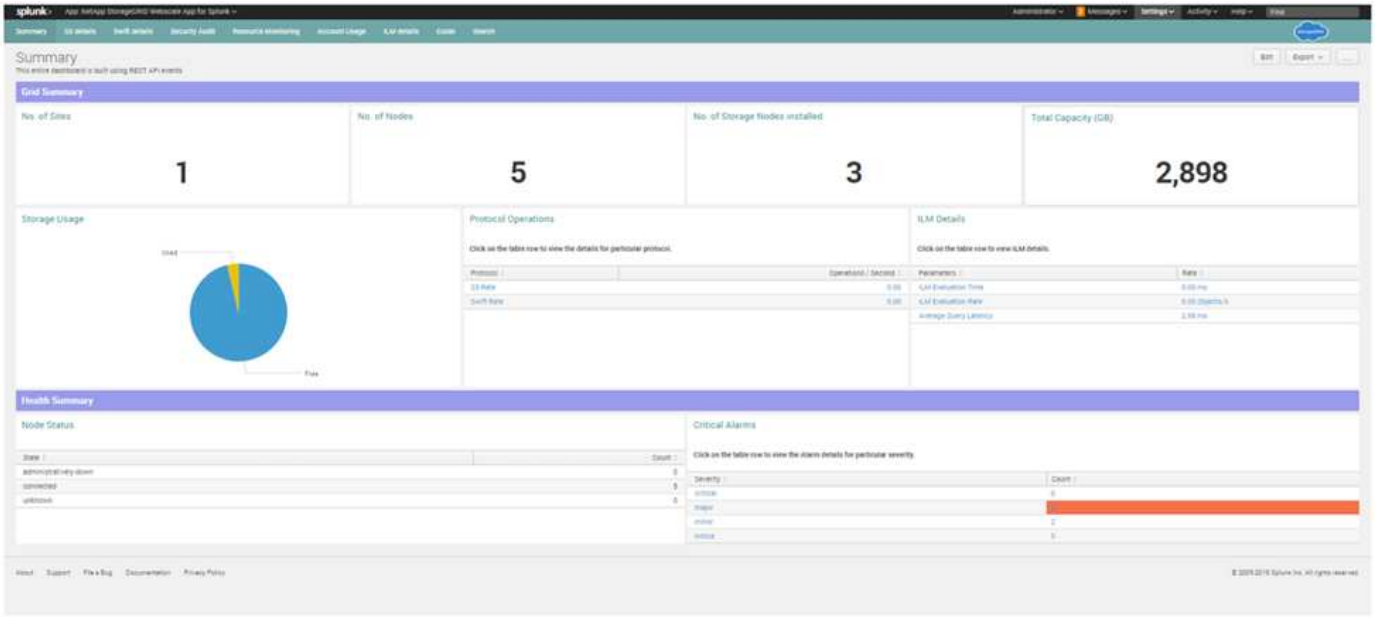
Perform the following tasks with the Grid Manager interface:

- Manage globally distributed, petabyte-scale repositories of objects such as images, video, and records.
- Monitor grid nodes and services to ensure object availability.
- Manage the placement of object data over time using information lifecycle management (ILM) rules. These rules govern what happens to an object's data after it is ingested, how it is protected from loss, where object data is stored, and for how long.
- Monitor transactions, performance, and operations within the system.

NetApp StorageGRID App for Splunk

The NetApp StorageGRID App for Splunk is an application specific for Splunk Enterprise. This app works in conjunction with the NetApp StorageGRID Add-on for Splunk. It provides visibility into StorageGRID health, account usage information, security audit details, resource usage and monitoring, and so on.

The following image shows the StorageGRID App for Splunk.



ILM policies

StorageGRID has flexible data management policies that include keeping multiple copies of your objects and using EC (erasure coding) schemes like 2+1 and 4+2 (and many others) to store your objects depending on specific performance and data protection requirements. As workloads and requirements change over time, it's common that ILM policies must change over time as well. Modifying ILM policies is a core feature, allowing StorageGRID customers to adapt to their ever-changing environment quickly and easily.

Performance

StorageGRID scales performance by adding more nodes, which can be VMs or bare metal or purpose-built appliances like the SG5712, SG5760, SG6060, or SGF6024. In our tests, we exceeded the SmartStore key performance requirements with a minimum-sized three-node grid using the SG6060 appliance. As customers scale their Splunk infrastructure with additional indexers, they can add more storage nodes to increase performance and capacity.

Load Balancer and endpoint configuration

Admin nodes in StorageGRID provide the Grid Manager UI (user interface) and REST API endpoint to view, configure, and manage your StorageGRID system, as well as audit logs to track system activity. To provide a highly available S3 endpoint for Splunk SmartStore remote storage, we implemented the StorageGRID load balancer, which runs as a service on admin nodes and gateway nodes. In addition, the load balancer also manages local traffic and talks to the GSLB (Global Server Load Balancing) to help with disaster recovery.

To further enhance endpoint configuration, StorageGRID provides traffic classification policies built into the admin node, lets you monitor your workload traffic, and apply various quality-of-service (QoS) limits to your workloads. Traffic classification policies are applied to endpoints on the StorageGRID Load Balancer service for gateway nodes and admin nodes. These policies can assist with traffic limiting and monitoring.

Next: Splunk architecture.

Splunk architecture

Previous: Flexible StorageGRID features for Splunk SmartStore.

Key definitions

The next two tables list the Splunk and NetApp components used in the distributed Splunk deployment.

This table lists the Splunk hardware components for the distributed Splunk Enterprise configuration.

Splunk component	Task
Indexer	Repository for Splunk Enterprise data
Universal forwarder	Responsible for ingesting data and forwarding data to the indexers
Search head	The user front end used to search data in indexers
Cluster master	Manages the Splunk installation of indexers and search heads
Monitoring Console	Centralized monitoring tool used across the entire deployment
License master	License master handles Splunk Enterprise licensing
Deployment server	Updates configurations and distributes apps to processing component
Storage component	Task
NetApp AFF	All-flash storage used to manage hot tier data. Also known as local storage.
NetApp StorageGRID	S3 object storage used to manage warm tier data. Used by SmartStore to move data between the hot and warm tier. Also known as remote storage.

This table lists the components in the Splunk storage architecture.

Splunk component	Task	Responsible component
SmartStore	Provides indexers with the ability to tier data from local storage to object storage.	Splunk
Hot	The landing spot where universal forwarders place newly written data. Storage is writable, and data is searchable. This data tier is typically composed of SSDs or fast HDDs.	ONTAP
Cache Manager	Manages local cache of indexed data, fetches warm data from remote storage when a search occurs, and evicts least frequently used data from the cache.	SmartStore

Splunk component	Task	Responsible component
Warm	Data is rolled logically to the bucket, renamed to the warm tier first from the hot tier. Data within this tier is protected and, like the hot tier, can be composed of larger capacity SSDs or HDDs. Both incremental and full backups are supported using common data protection solutions.	StorageGRID

Splunk distributed deployments

To support larger environments in which data originates on many machines, you need to process large volumes of data. If many users need to search the data, you can scale the deployment by distributing Splunk Enterprise instances across multiple machines. This is known as a distributed deployment.

In a typical distributed deployment, each Splunk Enterprise instance performs a specialized task and resides on one of three processing tiers corresponding to the main processing functions.

The following table lists the Splunk Enterprise processing tiers.

Tier	Component	Description
Data input	Forwarder	A forwarder consumes data and then forwards the data to a group of indexers.
Indexing	Indexer	An indexer indexes incoming data that it usually receives from a group of forwarders. The indexer transforms the data into events and stores the events in an index. The indexer also searches the indexed data in response to search requests from a search head.
Search management	Search head	A search head serves as a central resource for searching. The search heads in a cluster are interchangeable and have access to the same searches, dashboards, knowledge objects, and so on, from any member of the search head cluster.

The following table lists the important components used in a distributed Splunk Enterprise environment.

Component	Description	Responsibility
Index cluster master	Coordinates activities and updates of an indexer cluster	Index management

Component	Description	Responsibility
Index cluster	Group of Splunk Enterprise indexers that are configured to replicate data with each other	Indexing
Search head deployer	Handles deployment and updates to the cluster master	Search head management
Search head cluster	Group of search heads that serves as a central resource for searching	Search management
Load Balancers	Used by clustered components to handle increasing demand by search heads, indexers and S3 target to distribute load across clustered components.	Load Management for clustered components

See the following benefits of Splunk Enterprise distributed deployments:

- Access diverse or dispersed data sources
- Provide functionality to handle the data needs for enterprises of any size and complexity
- Achieve high availability and ensure disaster recovery with data replication and multisite deployment

Splunk SmartStore

SmartStore is an indexer capability that enables remote object stores such as Amazon S3 to store indexed data. As a deployment's data volume increases, demand for storage typically outpaces demand for compute resources. SmartStore allows you to manage your indexer storage and compute resources cost-effectively by scaling those resources separately.

SmartStore introduces a remote storage tier and a cache manager. These features allow data to reside either locally on indexers or on the remote storage tier. The cache manager manages data movement between the indexer and the remote storage tier, which is configured on the indexer.

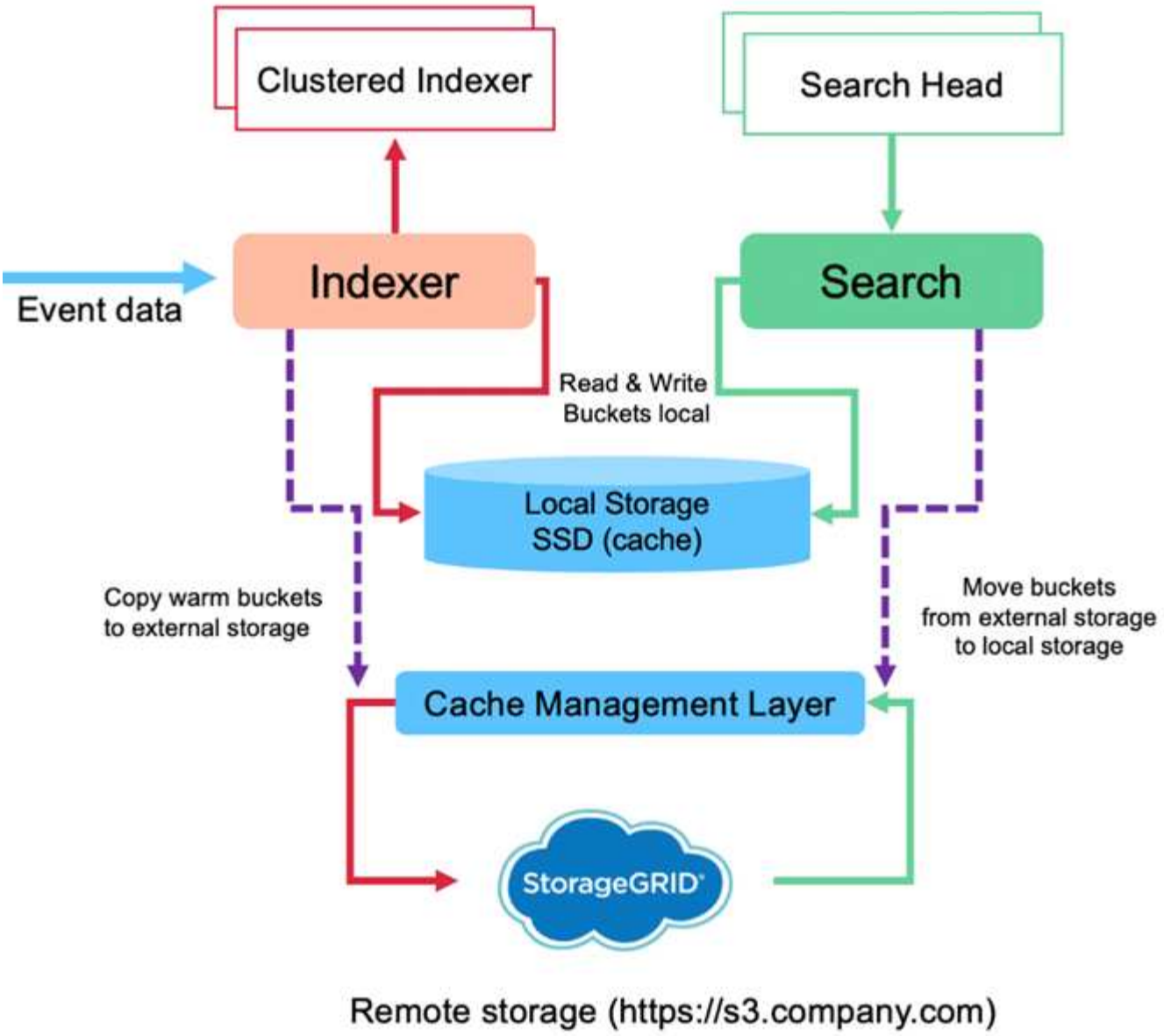
With SmartStore, you can reduce the indexer storage footprint to a minimum and choose I/O-optimized compute resources. Most data resides on the remote storage. The indexer maintains a local cache that contains a minimal amount of data: hot buckets, copies of warm buckets participating in active or recent searches, and bucket metadata.

Splunk SmartStore data flow

When data incoming from various sources reaches the indexers, data is indexed and saved locally in a hot bucket. The indexer also replicates the hot bucket data to target indexers. So far, the data flow is identical to the data flow for non-SmartStore indexes.

When the hot bucket rolls to warm, the data flow diverges. The source indexer copies the warm bucket to the remote object store (remote storage tier) while leaving the existing copy in its cache, because searches tend to run across recently indexed data. However, the target indexers delete their copies because the remote store provides high availability without maintaining multiple local copies. The master copy of the bucket now resides in the remote store.

The following image shows the Splunk SmartStore data flow.



The cache manager on the indexer is central to the SmartStore data flow. It fetches copies of buckets from the remote store as necessary to handle search requests. It also evicts older or less searched copies of buckets from the cache, because the likelihood of their participating in searches decreases over time.

The cache manager's job is to optimize the use of the available cache while ensuring that searches have immediate access to the buckets they need.

Software requirements

The table below lists the software components that are required to implement the solution. The software components that are used in any implementation of the solution might vary based on customer requirements.

Product family	Product name	Product version	Operating system
NetApp StorageGRID	StorageGRID object storage	11.6	n/a

Product family	Product name	Product version	Operating system
CentOS	CentOS	8.1	CentOS 7.x
Splunk Enterprise	Splunk Enterprise with SmartStore	8.0.3	CentOS 7.x

Single and multisite requirements

In an Enterprise Splunk environment (medium and large deployments) where data originates on many machines and where many users need to search the data, you can scale your deployment by distributing Splunk Enterprise instances across single and multiple sites.

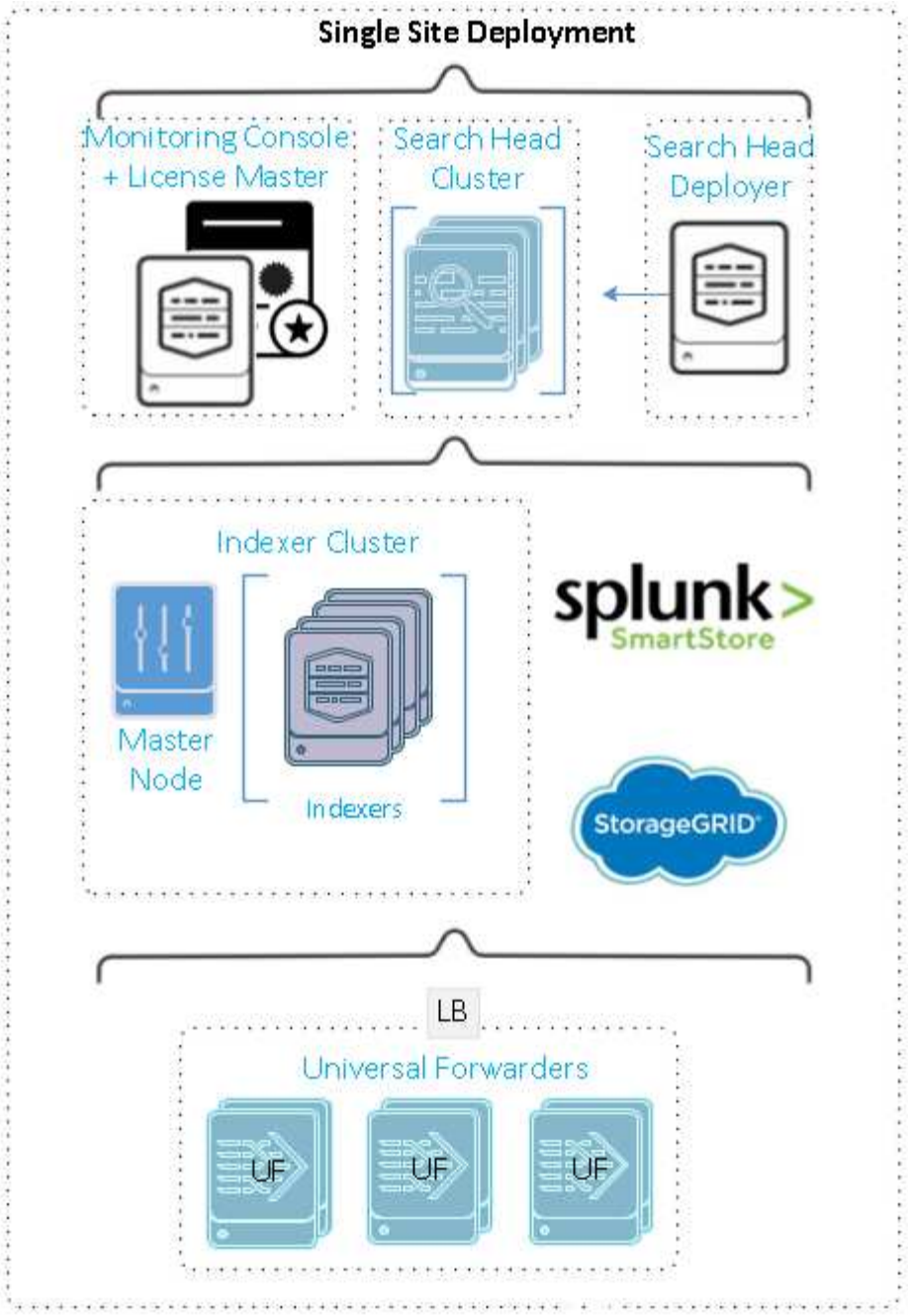
See the following benefits of Splunk Enterprise distributed deployments:

- Access diverse or dispersed data sources
- Provide functionality to handle the data needs for enterprises of any size and complexity
- Achieve high availability and ensure disaster recovery with data replication and multisite deployment

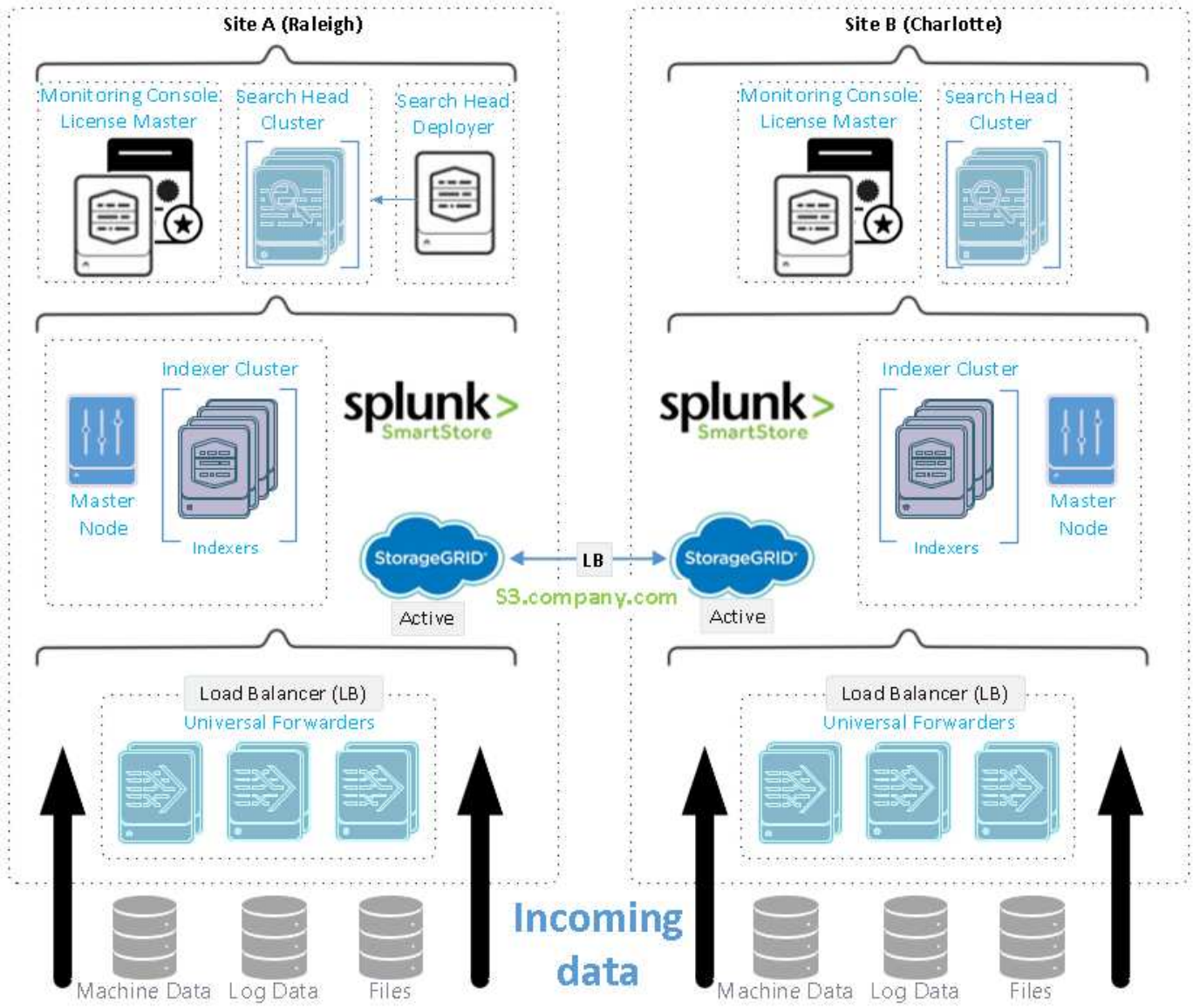
The following table lists the components used in a distributed Splunk Enterprise environment.

Component	Description	Responsibility
Index cluster master	Coordinates activities and updates of an indexer cluster	Index management
Index cluster	Group of Splunk Enterprise indexers that are configured to replicate each other’s data	Indexing
Search head deployer	Handles deployment and updates to the cluster master	Search head management
Search head cluster	Group of search heads that serves as a central resource for searching	Search management
Load balancers	Used by clustered components to handle increasing demand by search heads, indexers and S3 target to distribute load across clustered components.	Load management for clustered components

This figure depicts an example of a single-site distributed deployment.



This figure depicts an example of a multisite distributed deployment.



Hardware requirements

The following tables list the minimum number of hardware components that are required to implement the solution. The hardware components that are used in specific implementations of the solution might vary based on customer requirements.

> Regardless of whether you have deployed Splunk SmartStore and StorageGRID in a single site or in multiple sites, all systems are managed from the StorageGRID GRID Manager in a single pane of glass. See the section “Simple Management with Grid Manager” for more details.

This table lists the hardware used for a single site.

Hardware	Quantity	Disk	Usable capacity	Note
StorageGRID SG1000	1	n/a	n/a	Admin node and load balancer
StorageGRID SG6060	4	x48, 8TB (NL-SAS HDD)	1PB	Remote storage

This table lists the hardware used for a multisite configuration (per site).

Hardware	Quantity	Disk	Usable capacity	Note
StorageGRID SG1000	2	n/a	n/a	Admin node and Load balancer
StorageGRID SG6060	4	x48, 8TB (NL-SAS HDD)	1PB	Remote storage

NetApp StorageGRID Load Balancer: SG1000

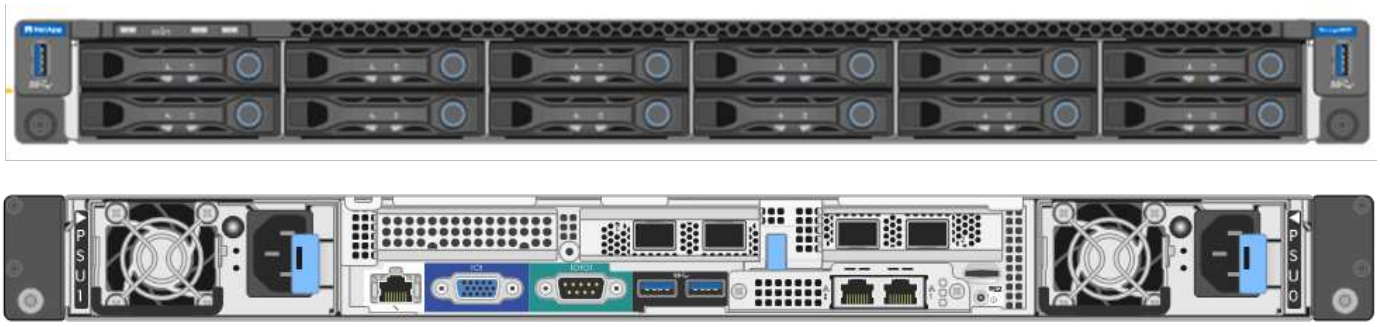
Object storage requires the use of a load balancer to present the cloud storage namespace. StorageGRID supports third- party load balancers from leading vendors like F5 and Citrix, but many customers choose the enterprise-grade StorageGRID balancer for simplicity, resiliency, and high performance. The StorageGRID load balancer is available as a VM, container, or purpose-built appliance.

The StorageGRID SG1000 facilitates the use of high availability (HA) groups and intelligent load balancing for S3 data-path connections. No other on-prem object storage system provides a customized load balancer.

The SG1000 appliance provides the following features:

- A load balancer and, optionally, admin node functions for a StorageGRID system
- The StorageGRID Appliance Installer to simplify node deployment and configuration
- Simplified configuration of S3 endpoints and SSL
- Dedicated bandwidth (versus sharing a third-party load balancer with other applications)
- Up to 4 x 100Gbps aggregate Ethernet bandwidth

The following image shows the SG1000 Gateway Services appliance.



SG6060

The StorageGRID SG6060 appliance includes a compute controller (SG6060) and a storage controller shelf (E-Series E2860) that contains two storage controllers and 60 drives. This appliance provides the following features:

- Scale up to 400PB in a single namespace.
- Up to 4x 25Gbps aggregate Ethernet bandwidth.
- Includes the StorageGRID Appliance Installer to simplify node deployment and configuration.
- Each SG6060 appliance can have one or two additional expansion shelves for a total of 180 drives.

- Two E-Series E2800 controllers (duplex configuration) to provide storage controller failover support.
- Five-drawer drive shelf that holds sixty 3.5-inch drives (two solid-state drives, and 58 NL-SAS drives).

The following image shows the SG6060 appliance.

Splunk design

The following table lists the Splunk configuration for a single site.

Splunk component	Task	Quantity	Cores	Memory	OS
Universal forwarder	Responsible for ingesting data and forwarding data to the indexers	4	16 Cores	32GB RAM	CentOS 8.1

Splunk component	Task	Quantity	Cores	Memory	OS
Indexer	Manages the user data	10	16 Cores	32GB RAM	CentOS 8.1
Search head	User front end searches data in indexers	3	16 Cores	32GB RAM	CentOS 8.1
Search head deployer	Handles updates for search head clusters	1	16 Cores	32GB RAM	CentOS 8.1
Cluster master	Manages the Splunk installation and indexers	1	16 Cores	32GB RAM	CentOS 8.1
Monitoring Console and license master	Performs centralized monitoring of the entire Splunk deployment and manages Splunk licenses	1	16 Cores	32GB RAM	CentOS 8.1

The following tables describe the Splunk configuration for multisite configurations.

This table lists the Splunk configuration for a multisite configuration (site A).

Splunk component	Task	Quantity	Cores	Memory	OS
Universal forwarder	Responsible for ingesting data and forwarding data to the indexers.	4	16 Cores	32GB RAM	CentOS 8.1
Indexer	Manages the user data	10	16 Cores	32GB RAM	CentOS 8.1
Search head	User front end searches data in indexers	3	16 Cores	32GB RAM	CentOS 8.1
Search head deployer	Handles updates for search head clusters	1	16 Cores	32GB RAM	CentOS 8.1
Cluster master	Manages the Splunk installation and indexers	1	16 Cores	32GB RAM	CentOS 8.1

Splunk component	Task	Quantity	Cores	Memory	OS
Monitoring Console and license master	Performs centralized monitoring of the entire Splunk deployment and manages Splunk licenses.	1	16 Cores	32GB RAM	CentOS 8.1

This table lists the Splunk configuration for a multisite configuration (site B).

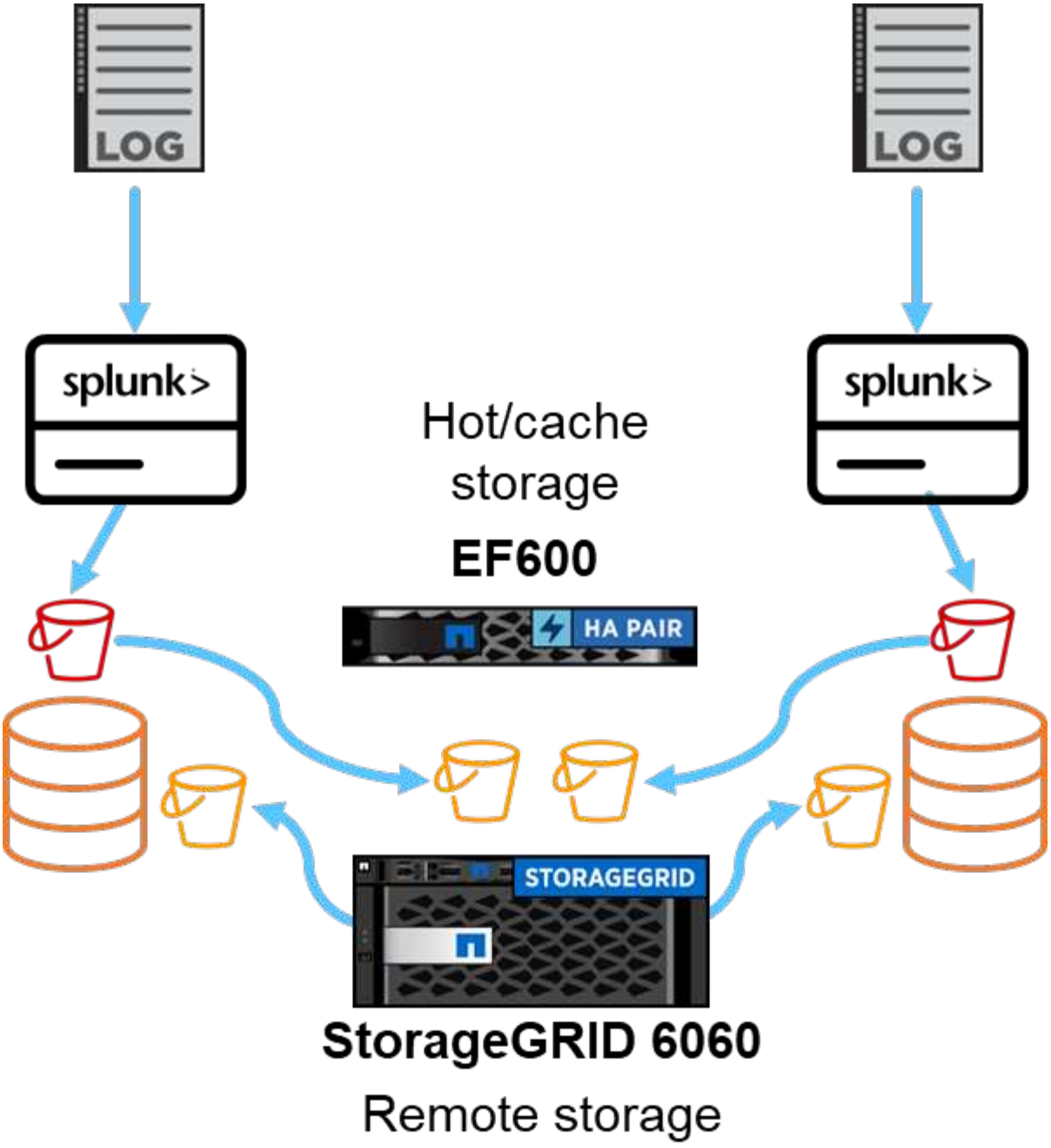
Splunk component	Task	Quantity	Cores	Memory	OS
Universal forwarder	Responsible for ingesting data and forwarding data to the indexers	4	16 Cores	32GB RAM	CentOS 8.1
Indexer	Manages the user data	10	16 Cores	32GB RAM	CentOS 8.1
Search head	User front end searches data in indexers	3	16 Cores	32GB RAM	CentOS 8.1
Cluster master	Manages the Splunk installation and indexers	1	16 Cores	32GB RAM	CentOS 8.1
Monitoring Console and license master	Performs centralized monitoring of the entire Splunk deployment and manages Splunk licenses	1	16 Cores	32GB RAM	CentOS 8.1

Next: Single-site SmartStore performance.

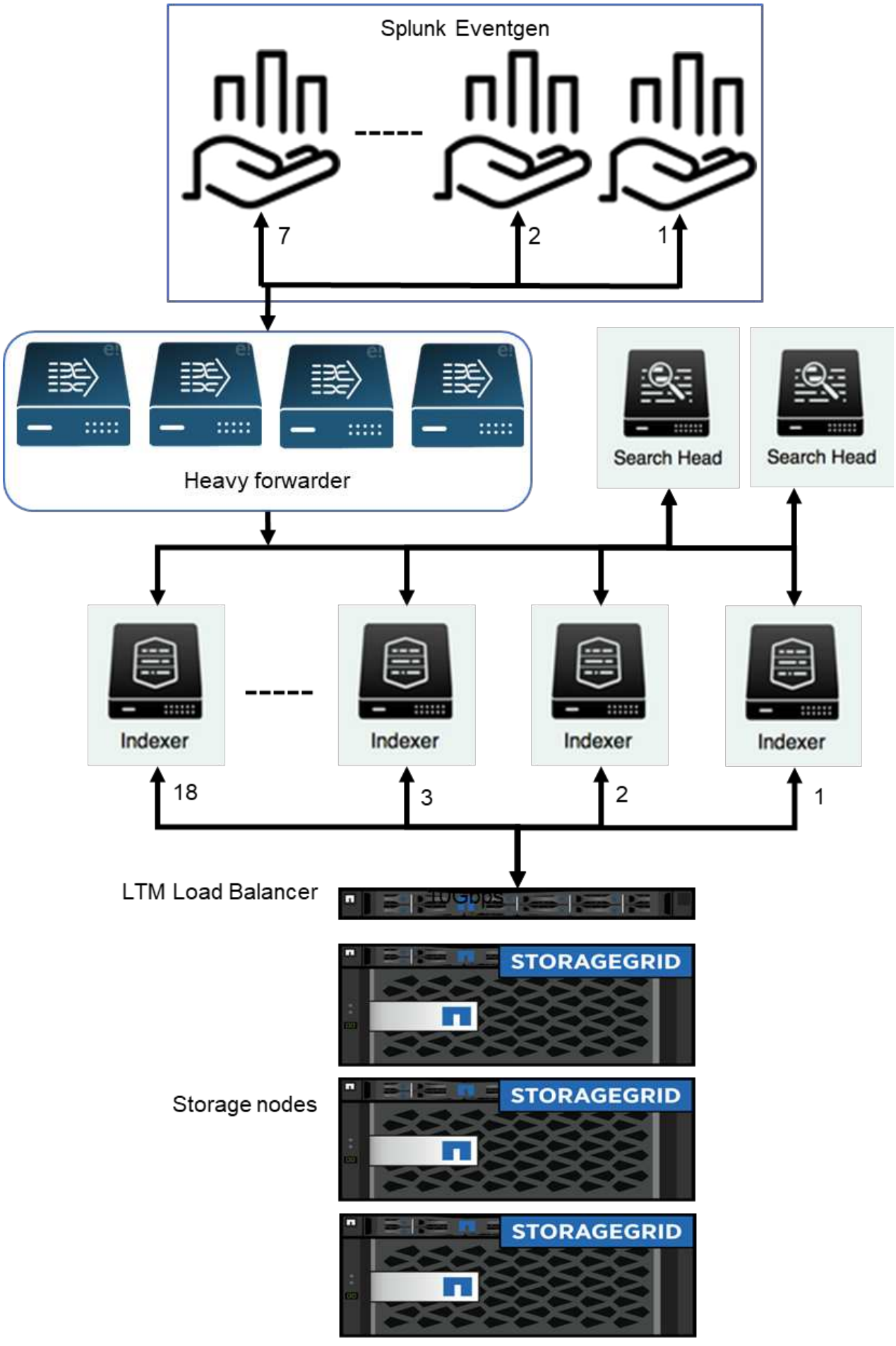
Single-site SmartStore performance

Previous: Splunk architecture.

This section describes Splunk SmartStore performance on a NetApp StorageGRID controller. Splunk SmartStore moves warm data to remote storage, which in this case is StorageGRID object storage in the performance validation.



We used EF600 for hot/cache storage and StorageGRID 6060 for remote storage. We used the following architecture for the performance validation. We used two search heads, four heavy forwarders to forward the data to indexers, seven Splunk Event Generators (Eventgens) to generate the real-time data, and 18 indexers to store the data.


Splunk Eventgen
7
2
1
Heavy forwarder
Search Head
Search Head
Indexer
Indexer
Indexer
Indexer
18
3
2
1
LTM Load Balancer
10Gbps
STORAGEGRID
STORAGEGRID
STORAGEGRID
Storage nodes

Configuration

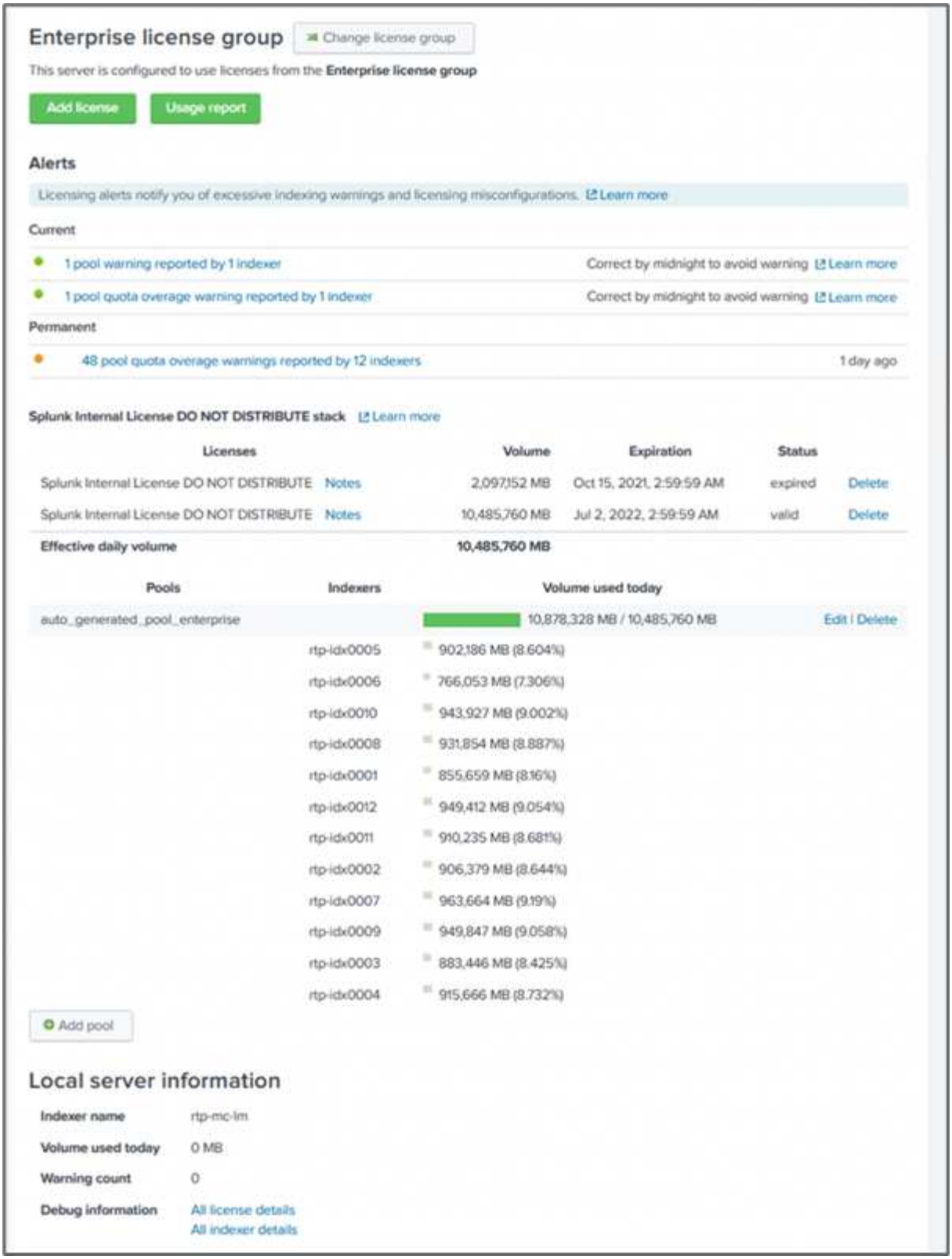
This table lists the hardware used for the SmartStorage performance validation.

Splunk component	Task	Quantity	Cores	Memory	OS
Heavy forwarder	Responsible for ingesting data and forwarding data to the indexers	4	16 cores	32GB RAM	SLED 15 SP2
Indexer	Manages the user data	18	16 cores	32GB RAM	SLED 15 SP2
Search head	User front-end searches data in indexers	2	16 cores	32GB RAM	SLED 15 SP2
Search head deployer	Handles updates for search head clusters	1	16 cores	32GB RAM	SLED 15 SP2
Cluster master	Manages the Splunk installation and indexers	1	16 cores	32GB RAM	SLED 15 SP2
Monitoring Console and license master	Performs centralized monitoring of the entire Splunk deployment and manages Splunk licenses	1	16 cores	32GB RAM	SLED 15 SP2

SmartStore remote store performance validation

In this performance validation, we configured the SmartStore cache in local storage on all the indexers for 10 days of data. We enabled the `maxDataSize=auto` (750MB bucket size) in Splunk cluster manager and pushed the changes to all the indexers. To measure the upload performance, we ingested 10TB per day for 10 days and rolled over all hot buckets to warm at the same time and captured the peak and average throughput per instance and deployment-wide from the SmartStore Monitoring Console dashboard.

This image shows the data ingested in one day.



We ran the following command from cluster master (the index name is `eventgen-test`). Then we captured the peak and average upload throughput per instance and deployment-wide through the SmartStore Monitoring Console dashboards.

```
for i in rtp-idx0001 rtp-idx0002 rtp-idx0003 rtp-idx0004 rtp-idx0005 rtp-
idx0006 rtp-idx0007 rtp-idx0008 rtp-idx0009 rtp-idx0010 rtp-idx0011 rtp-
idx0012 rtp-idx0013011 rtdx0014 rtp-idx0015 rtp-idx0016 rtp-idx0017 rtp-
idx0018 ; do  ssh $i "hostname;  date; /opt/splunk/bin/splunk _internal
call /data/indexes/eventgen-test/roll-hot-buckets -auth admin:12345678;
sleep 1  "; done
```

The cluster master has password-less authentication to all indexers (rtp-idx0001…rtp-idx0018).

To measure the download performance, we evicted all data from the cache by running the evict CLI twice by using the following command.

We ran the following command from cluster master and ran the search from the search head on top of 10 days of data from the remote store from StorageGRID. We then captured the peak and average upload throughput per instance and deployment-wide through the SmartStore Monitoring Console dashboards.

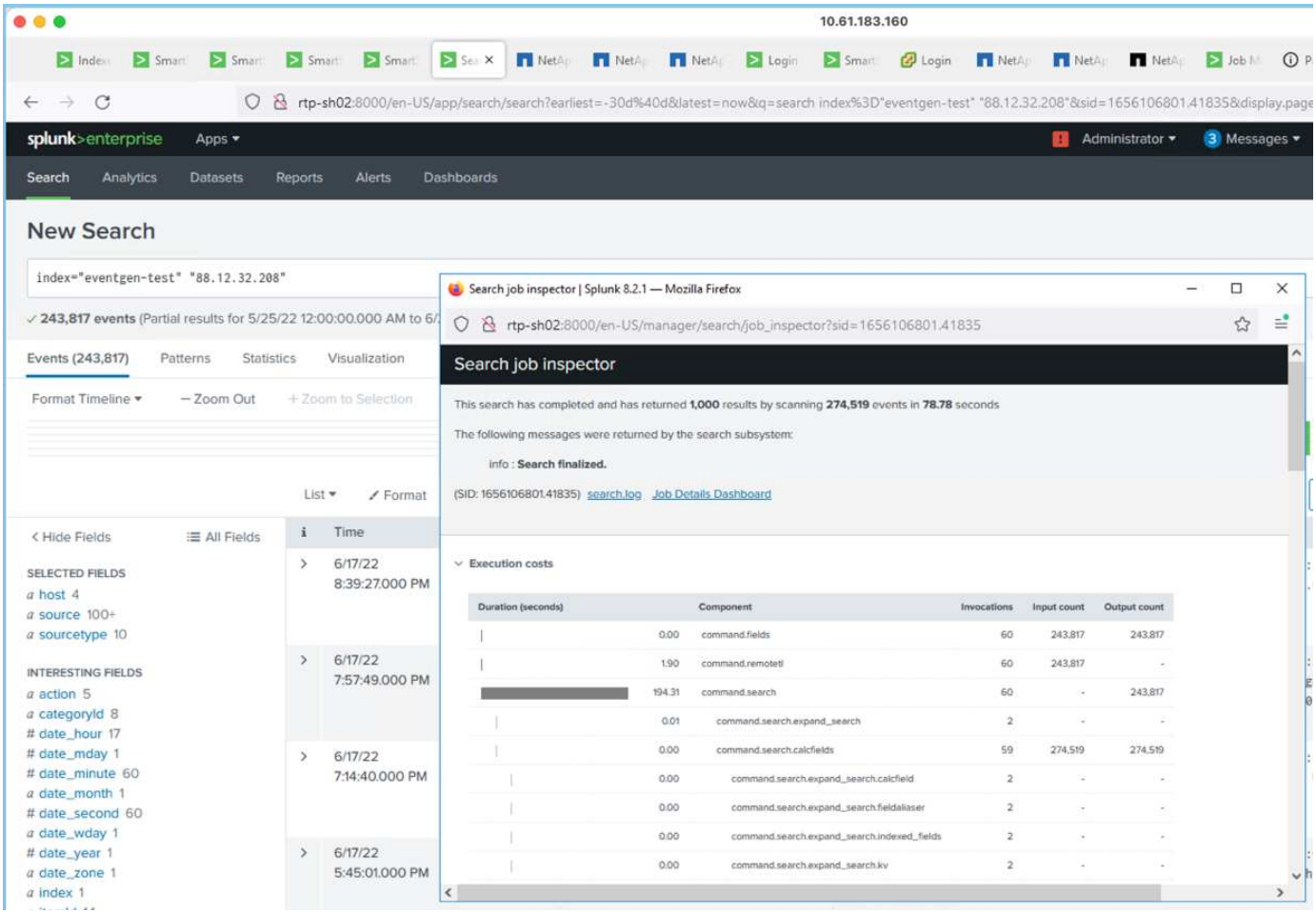
```
for i in rtp-idx0001 rtp-idx0002 rtp-idx0003 rtp-idx0004 rtp-idx0005 rtp-
idx0006 rtp-idx0007 rtp-idx0008 rtp-idx0009 rtp-idx0010 rtp-idx0011 rtp-
idx0012 rtp-idx0013 rtp-idx0014 rtp-idx0015 rtp-idx0016 rtp-idx0017 rtp-
idx0018 ; do  ssh $i " hostname;  date; /opt/splunk/bin/splunk _internal
call /services/admin/cacheman/_evict -post:mb 1000000000 -post:path
/mnt/EF600 -method POST  -auth admin:12345678;   "; done
```

The indexer configurations were pushed from SmartStore cluster master. The cluster master had the following configuration for the indexer.

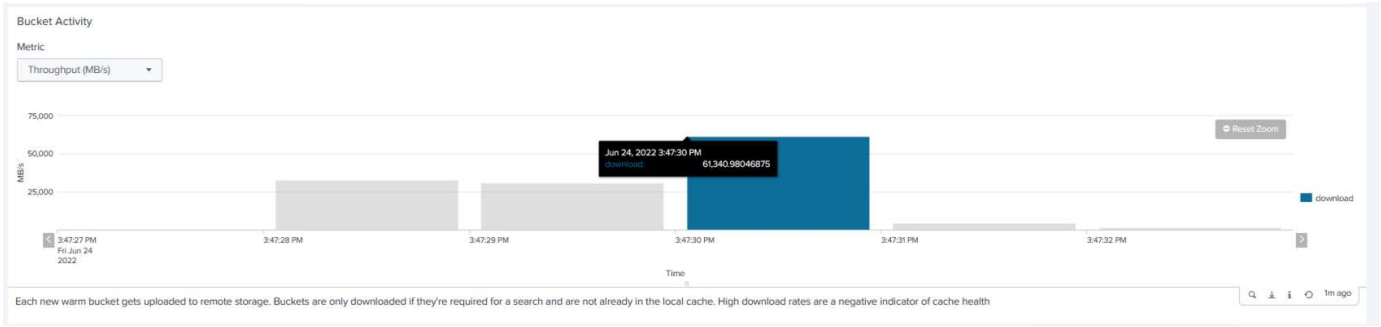
```
Rtp-cm01:~ # cat /opt/splunk/etc/master-apps/_cluster/local/indexes.conf
[default]
maxDataSize = auto
#defaultDatabase = eventgen-basic
defaultDatabase = eventgen-test
hotlist_recency_secs = 864000
repFactor = auto
[volume:remote_store]
storageType = remote
path = s3://smartstore2
remote.s3.access_key = U64TUHONBNC98GQGL60R
remote.s3.secret_key = UBoXNE0jmECie05Z7iCYVzbSB6WJFckiYLcdm2yg
remote.s3.endpoint = 3.sddc.netapp.com:10443
remote.s3.signature_version = v2
remote.s3.clientCert =
[eventgen-basic]
homePath = $SPLUNK_DB/eventgen-basic/db
```

```
coldPath = $SPLUNK_DB/eventgen-basic/colddb
thawedPath = $SPLUNK_DB/eventgen-basic/thawed
[eventgen-migration]
homePath = $SPLUNK_DB/eventgen-scale/db
coldPath = $SPLUNK_DB/eventgen-scale/colddb
thawedPath = $SPLUNK_DB/eventgen-scale/thaweddb
[main]
homePath = $SPLUNK_DB/$_index_name/db
coldPath = $SPLUNK_DB/$_index_name/colddb
thawedPath = $SPLUNK_DB/$_index_name/thaweddb
[history]
homePath = $SPLUNK_DB/$_index_name/db
coldPath = $SPLUNK_DB/$_index_name/colddb
thawedPath = $SPLUNK_DB/$_index_name/thaweddb
[summary]
homePath = $SPLUNK_DB/$_index_name/db
coldPath = $SPLUNK_DB/$_index_name/colddb
thawedPath = $SPLUNK_DB/$_index_name/thaweddb
[remote-test]
homePath = $SPLUNK_DB/$_index_name/db
coldPath = $SPLUNK_DB/$_index_name/colddb
#for storagegrid config
remotePath = volume:remote_store/$_index_name
thawedPath = $SPLUNK_DB/$_index_name/thaweddb
[eventgen-test]
homePath = $SPLUNK_DB/$_index_name/db
maxDataSize=auto
maxHotBuckets=1
maxWarmDBCount=2
coldPath = $SPLUNK_DB/$_index_name/colddb
#for storagegrid config
remotePath = volume:remote_store/$_index_name
thawedPath = $SPLUNK_DB/$_index_name/thaweddb
[eventgen-evict-test]
homePath = $SPLUNK_DB/$_index_name/db
coldPath = $SPLUNK_DB/$_index_name/colddb
#for storagegrid config
remotePath = volume:remote_store/$_index_name
thawedPath = $SPLUNK_DB/$_index_name/thaweddb
maxDataSize = auto_high_volume
maxWarmDBCount = 5000
rtp-cm01:~ #
```

We ran the following search query on the search head to collect the performance matrix.



We collected the performance information from the cluster master. The peak performance was 61.34GBps.

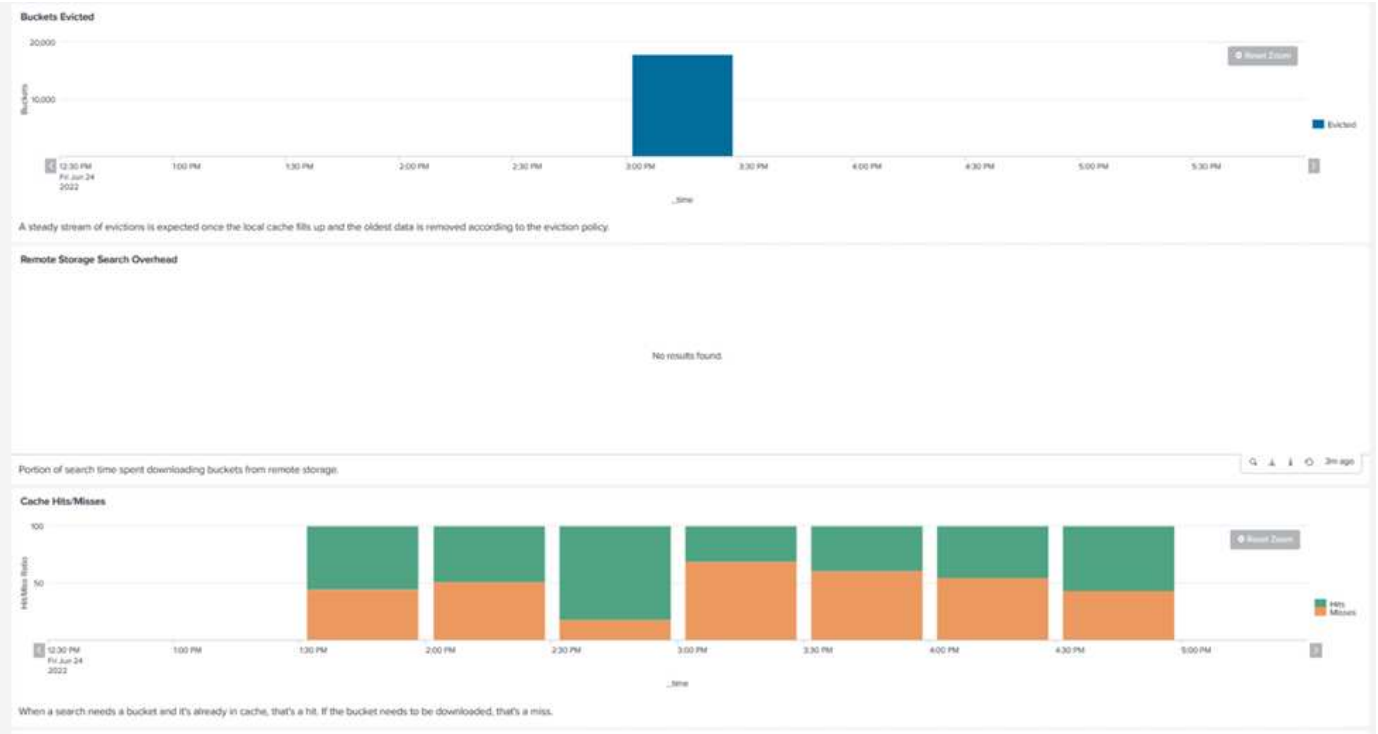


The average performance was approximately 29GBps.

StorageGRID performance

SmartStore performance is based on searching for specific patterns and strings from large amounts of data. In this validation, the events are generated using Eventgen on a specific Splunk index (eventgen-test) through the search head, and the request goes to StorageGRID for most of the queries. The following image shows the hits and misses of the query data. The hits data is from the local disk and the misses data is from the StorageGRID controller.

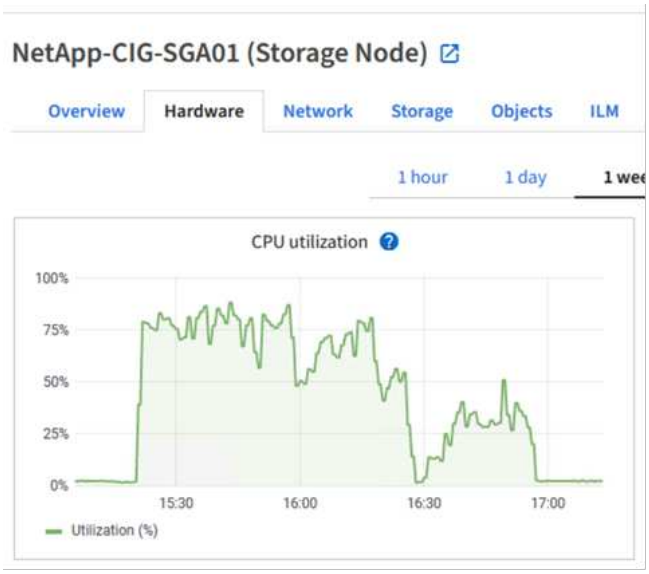
The green color shows the hits data and the orange color shows the misses data.

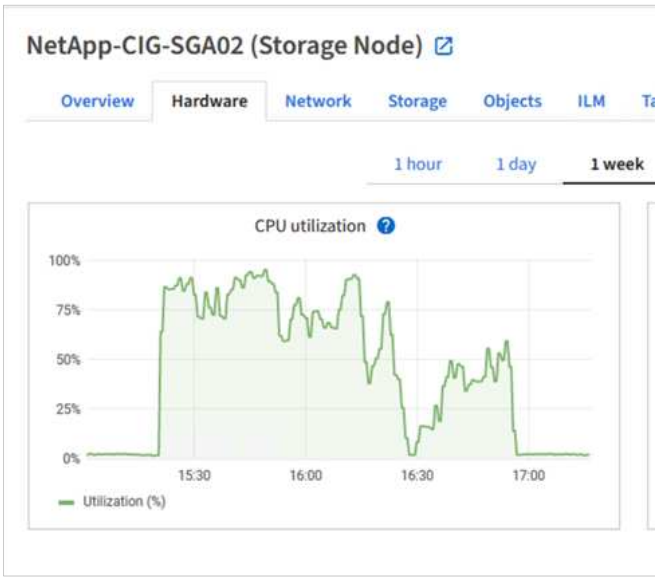


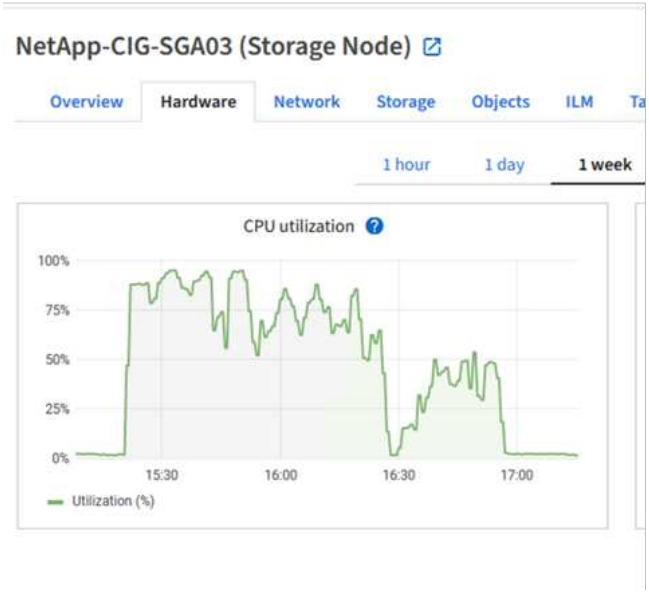
When the query runs for the search on StorageGRID, the time for the S3 retrieve rate from StorageGRID is shown in the following image.

StorageGRID hardware usage

The StorageGRID instance has one load balancer and three StorageGRID controllers. CPU utilization for all three controllers is from 75% to 100%.







SmartStore with NetApp storage controller - benefits for the customer

- **Decoupling compute and storage.** The Splunk SmartStore decouples compute and storage, which helps you to scale them independently.
- **Data on-demand.** SmartStore brings data close to compute on-demand and provides compute and storage elasticity and cost efficiency to achieve longer data retention at scale.
- **AWS S3 API compliant.** SmartStore uses the AWS S3 API to communicate with restore storage, which is an AWS S3 and S3 API-compliant object store such as StorageGRID.
- **Reduces storage requirement and cost.** SmartStore reduces the storage requirements for aged data (warm/cold). It only needs a single copy of data because NetApp storage provides data protection and takes care of failure and high availability.
- **Hardware failure.** Node failure in a SmartStore deployment does not make the data inaccessible and has a much faster indexer recovery from hardware failure or data imbalance.
- Application and data-aware cache.
- Add-remove indexers and setup-teardown cluster on-demand.

- Storage tier is no longer tied to hardware.

Next: Conclusion.

Conclusion

Previous: Single-site SmartStore performance.

Splunk Enterprise is the market-leading SIEM solution driving outcomes across Security, IT, and DevOps teams. The use of Splunk has increased considerably across our customer's organizations. Therefore, there is a need to add more data sources while also retaining the data for a longer period, thus stressing the Splunk infrastructure.

The combination of Splunk SmartStore and NetApp StorageGRID is designed to provide a scalable architecture for organizations to achieve improved ingest performance with SmartStore and StorageGRID object storage and increased scalability for a Splunk environment across multiple geographical regions.

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

- NetApp StorageGRID Documentation Resources

  https://www.netapp.com/data-storage/storagegrid/documentation/

- NetApp Product Documentation

  https://docs.netapp.com

- Splunk Enterprise Documentation

  https://docs.splunk.com/Documentation/Splunk

- Splunk Enterprise About SmartStore

  https://docs.splunk.com/Documentation/Splunk/8.0.6/Indexer/AboutSmartStore

- Splunk Enterprise Distributed Deployment Manual

  https://docs.splunk.com/Documentation/Splunk/8.0.6/Deploy/Distributedoverview

- Splunk Enterprise Managing Indexers and Clusters of Indexers

  https://docs.splunk.com/Documentation/Splunk/8.0.6/Indexer/Aboutindexesandindexers

Version history

Version	Date	Document version history
1.0	July 2022	Initial release.