ScoutAM Whitepaper

Table of Contents

Introduction

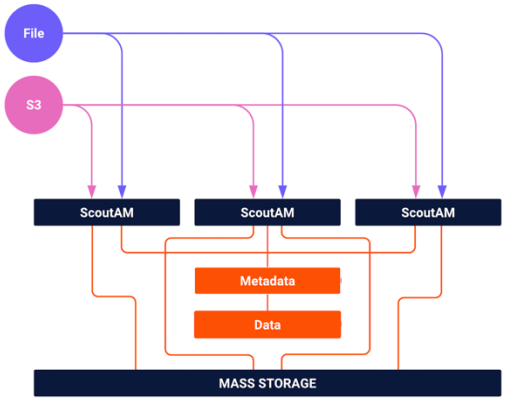
Scale Out Archive Manager (ScoutAM) is the first scalable, modular, economically efficient software defined mass storage solution to address the need for a simple but powerful tool to manage and preserve petabyte to exabyte scale collections of unstructured data.

ScoutAM ingests file and object data from primary data sources, establishes persistent searchable online metadata records and then automatically brokers data in parallel to a composable mix of mass storage devices and services such as public cloud, private cloud, disk, and tape storage systems.

The ScoutAM platform provides cloud scale services by scaling both horizontally and vertically and was developed to preserve and process large volumes of metadata while delivering exabyte-scale data throughput performance to meet the most demanding mass storage requirements where hundreds of millions of files and objects are rapidly moving in and out of mass storage systems.

ScoutAM includes a sophisticated and highly configurable policy engine with on-the-fly tuning capabilities to automate the process of aligning diverse incoming data streams with different protection levels, physical locations, and storage classes. For example, certain files or frequently used objects may be protected with one or two copies on tape, but also routed to a flash or disk storage class to improve read performance. Since workflows and data types are often changing, the ability to add new storage policies to a running system is crucial advantage for modern datacenter workflows.

While developed for exabyte-scale workloads, ScoutAM is designed for ease of deployment, configuration, and use. Our vision for the product was to create the most capable mass storage platform in the world while at the same time ensuring that it is easy to deploy and manage. ScoutAM is the first mass storage system with a quick start guide and an rpm-based installation that can be accomplished in 30 minutes.



This whitepaper describes the ScoutAM mass storge software platform including the features and benefits derived from the open source ScoutFS filesystem which is a Versity supported component of the ScoutAM subscription-based solution. ScoutAM is the runtime application and the name of the Versity platform. ScoutAM is always deployed with ScoutFS filesystem.

Unique Attributes

ScoutAM is the **only** mass storage platform in its class featuring:

- Open-source data formats on archival and mass storage media
- An open-source filesystem for metadata management (ScoutFS)
- Ultimate disaster recovery - data and metadata recorded on archival media
- Converged metadata architecture - no third-party database hand-offs required
- Support from an independent, sustainable, growing, and self-governed company

Key Benefits

- Cost efficiency

  ScoutAM allows large storage sites to operate their own infrastructure with cloud scale cost efficiency. Versity sites are able to obtain TCO levels up to 10x lower than Amazon Glacier while delivering unlimited amounts of data to end users over high-speed local networks without unpredictable egress or transaction fees.

- Vendor Neutrality

  The ScoutAM platform is hardware agnostic, enabling large storage sites to add, change, and mix hardware components or back end storage systems to take full advantage of cost and performance improvements.

- Hybrid Cloud

  ScoutAM provides enterprise class performance natively for both traditional on-premises tape systems and private or public cloud systems from a single application. This enables background data migrations to the cloud, or the hybrid use of both cloud and on-premises storage to accomplish different goals such as accessibility and disaster recovery.

- Transparent interface

  ScoutAM supports standard read and write workflows with a transparent interface. Users and applications can locate archival files and objects as if they were local copies by interacting with online metadata. When users drag and drop files, or applications read data, the retrieval is orchestrated by ScoutAM and the data is progressively read before it is completely staged back from the archival media.

Features & Capabilities

Scale Out Architecture >

In recent years, technologies emerged that enabled Versity to design a new system architecture. We created a scalable “nodes and services” architecture to deliver high availability, with modular building blocks that can be added to scale up performance of the platform. All of the nodes in a ScoutAM cluster deliver parallel data processing and scale-out metadata services. Please see the Architecture section of this whitepaper for additional details.

Parallel Transfer >

Extremely large and active data collections require very high reading and writing throughput rates. ScoutAM has employed an innovative, highly scalable packet-based architecture to move many data streams in parallel from each node in a ScoutAM cluster.

Exabyte-Scale Capacity >

The ScoutFS global namespace is designed to accommodate hundreds of billions of files and objects while delivering responsive search and query results. Data storage capacity scales to hundreds of zettabytes. Versity sites can store and retrieve petabytes of data per day under real world conditions.

GUI >

ScoutAM includes a modern API-driven graphical user interface for administration of the system, monitoring, alerting, and configuration. The GUI is an offline web app built with React and is powered by the ScoutAM REST API. The graphical interface has been designed to reflect the same functionality as the CLI.

Modular >

Since storage requirements change frequently, it is important to adopt a platform that is comprised of discreet capacity and performance-based building blocks. Each node on a ScoutAM cluster is capable of individually moving ~ 10 GB/s, and aggregate throughput increases modularly by adding nodes. Each element of the ScoutAM platform is modular and may be scaled up including the metadata devices, the data cache devices, the ScoutAM nodes, and the mass storage devices and services.

Extensive API Coverage >

ScoutAM is a mass storage platform designed to interoperate with a wide array of scientific and enterprise applications. Broad API coverage is available to expose functionality to external systems that are part of data acquisition and storage workflows. All significant features and capabilities within the ScoutAM platform are supported by published RESTful API's.

High Availability >

ScoutAM remains available despite the loss of servers in a cluster. Depending on the size of the cluster and the quorum definitions, ScoutAM can tolerate the loss of one or more servers with no impact to availability or continuity of services. High availability is built into the ScoutAM platform and does not require complex external failover or HA tools.

Replication & Federation >

ScoutAM supports asynchronous read-only remote namespace replication. Metadata, cache data, and mass storage or archival data may be replicated to one or more disaster recovery or secondary locations to ensure continuity of operations. If the primary site is lost, a remote location may be designated as the primary location and switched from read-only to full read/write functionality. In addition to dual site replication, the same framework may be used to federate the namespace and data at multiple read-only sites including cloud native sites.

Extended Cache >

ScoutAM supports the ability to increase cache capacity by adding low cost S3 object storage devices to augment the capacity of the primary cache. The space available on the extended cache will be fully utilized and the least used data will be automatically evicted as newer data arrives. Pairing extended cache storage with an all-flash primary cache enables an optimal mix of performance and capacity for sites that require high throughput performance and also wish to have a higher percentage of their recent data cached to improve recall times.

Object to Tape >

The ScoutAM platform supports scalable ingesting of S3 object data and delivery of the object data to tape. This capability is crucial for data analytics platforms and data lakes both for protection and capacity extension. Versity's object to tape solution is deployed at scale and requires no special tooling, or implementation of custom variations to the S3 protocol.

Ultimate Disaster Recovery >

ScoutAM is the only platform in its class that records all of the data needed to restore the collection on the mass storage media including file & object metadata (file name, path, etc..), data, and checksum information. With ScoutAM it is possible to recover an entire data collection directly from the storage media.

Open Source >

ScoutFS is an open-source GPLv2, in-kernel Linux filesystem. Open-source metadata combined with open-source data formats ensure that customers can maintain complete control over their data collections and maintain alignment with their long-term data preservation and data autonomy goals.

POSIX that performs >

ScoutFS is fully POSIX compliant. Minimum synchronization points ensure maximum performance, while interfacing seamlessly with other POSIX applications.

Advanced search & indexing >

Metadata and data sequence numbers are indexed so both inode attributes and file content changes are quickly discovered. Specific attributes are indexed and can be queried with responsive performance, even among very large file and object populations. Costly filesystem scans are no longer required.

Custom metadata >

ScoutAM supports extensive user-defined metadata and automated rich metadata harvesting and management. User defined metadata may be indexed for fast searching.

Small file friendly >

Small files are commonly the downfall of mass storage systems. ScoutAM handles both large and small files with maximum efficiency.

Automatic & Smart >

ScoutAM employs a rich policy engine to automatically move data to the desired media, and to automatically manage cache space. Smart policy rules enable efficient grouping of data to optimize locality of data and overall system utilization and throughput. The application of policies, packaging of work, and execution of archive jobs is packetized, and executed in parallel utilizing all available node resources.

Software Architecture

Horizontal & Vertical Scaling

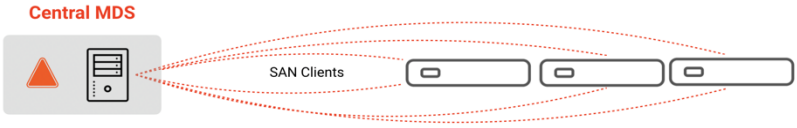
ScoutAM R&D efforts have focused more on scalability and simplicity than any other attributes. The growth of unstructured data is unrelenting, and the continued rapid advancement of science and industry is dependent in many ways upon dramatically improving the scalability of storage systems. ScoutAM is built upon a nodes & service architecture wherein additional performance may be added in modular chunks by scaling horizontally - adding more nodes to a cluster.

ScoutAM scales vertically by taking advantage of more cores, memory, and networking interfaces on a given node. The Golang development platform used in ScoutAM plays a significant enabling role for both the horizontal and vertical scalability of ScoutAM.

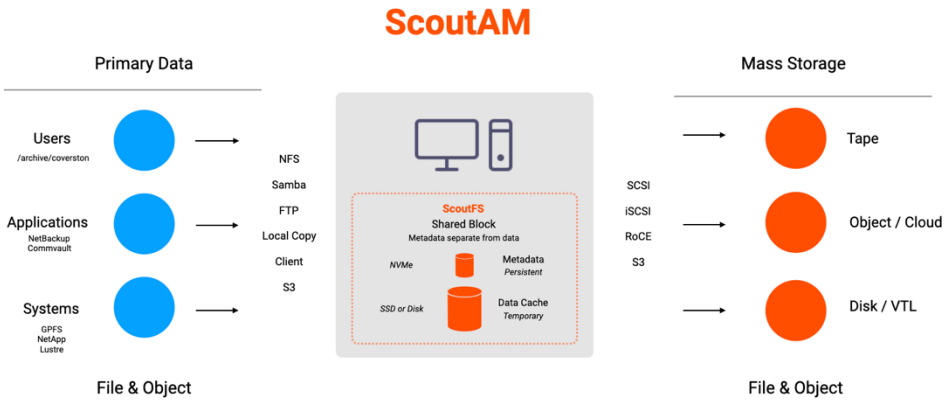
ScoutAM was designed to radically improve the scaling limitations inherent in legacy HSM, large archive, and mass storage products. The primary scaling bottlenecks in traditional product architectures are metadata management and sustained data throughput.

Metadata management

Metadata scaling limitations arise from the use of a centralized metadata server, also referred to as a metadata controller (MDC) or from other architectures that are unable to take advantage of additional compute and storage resources. Poor metadata scalability can result in degraded performance, even for routine tasks. Single server architectures limit the computational resources available for generating the massive number of IOPS needed to read, write, and search large volumes of metadata. With file and object populations in single data collections sometimes exceeding one billion items, even a single search that requires walking the full namespace might take many hours or days. Similarly, administrative functions like metadata dumps and restores can require extended periods of time.



The ScoutAM platform utilizes the open-source Scale Out Filesystem (ScoutFS) to maintain and manage metadata. ScoutFS was designed to manage tens to hundreds of billions of files and objects in a single namespace. ScoutFS is a multi-node, clustered metadata management system. In the ScoutFS architecture, there is no central metadata server. All of the nodes in a ScoutAM cluster (or a subset of the nodes) may be utilized by the ScoutFS filesystem allowing the combined resources of the cluster to be leveraged for scalable metadata services.



ScoutAM is the only mass storage platform built on a converged metadata architecture, meaning that metadata is stored within the application stack. Other products utilize a diverged model wherein metadata is handed off to a separate third-party database tool like Cassandra, or DB2 for storage and processing. Many of the legacy mass storage systems refer to their metadata collection as “the database”, reflecting the fact that the metadata is managed by a separate tool.

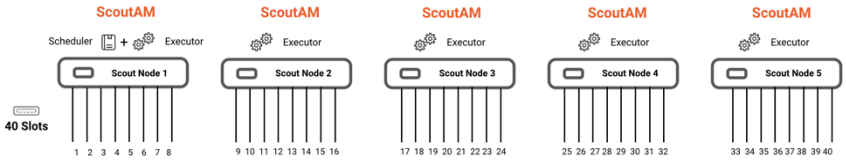
In addition to adding significant complexity to the application, external databases must be populated by transferring or handing off metadata from an underlying filesystem source. The hand-off introduces latency and is error prone even under optimal operating conditions. When errors occur, the ingested data stream must be interrupted and replayed. If the database server is unable to keep up with the pace of incoming metadata information then the database create rate establishes a fixed upper boundary on performance for the entire system. By contrast, ScoutAM utilizes ScoutFS to manage metadata internally with its own highly optimized data structures. This converged architecture is simpler, faster, more reliable, and more scalable.

Data Throughput

Data throughput scaling problems arise from slow metadata performance and a lack of parallelism. If metadata performance is slow, then the number of read/write requests will

decline or reach a GB/s performance speed limit strictly correlated with the metadata processing speed. Small file workloads are often slow due to this phenomenon. The underlying storage devices are capable of much faster performance, but an internal speed limit is imposed by the capabilities of the metadata management system. Assuming metadata processing scales well, data throughput can still be architecturally limited by a lack of parallelism or an inability to smoothly schedule and dispatch extremely large volumes of work across a cluster of server nodes.

ScoutAM is a scale out solution capable of utilizing all of the resources within a cluster for highly parallel data transfers and extremely high sustained aggregate data throughput. The workflow within ScoutAM is packetized and dispatched by a "scheduler" to each node where it is accepted and processed by an "executor". Each node in the cluster may have as many data "slots" or channels (connections to tape drives or interfaces to object storage or cloud resources) as the server nodes are equipped to provide. All of the slots on all of the nodes are available for parallel reading and writing of data. ScoutAM is capable of segmenting a single large file or object using a configurable segment size then sending the segments to all or a subset of available slots in the cluster thus enabling the parallel writing of a very large file across potentially dozens of tape drives or object interfaces simultaneously. Smaller quantities of data may be dispatched across the available slots in a cluster using a round robin algorithm. The scheduler and executors work together to ensure parallel uninterrupted high bandwidth streams of data are delivered across the slots for maximum aggregate streaming throughput. Additional throughput can be added by increasing the number of slots per server node (vertical scaling) or increasing the number of nodes (horizontal scaling).

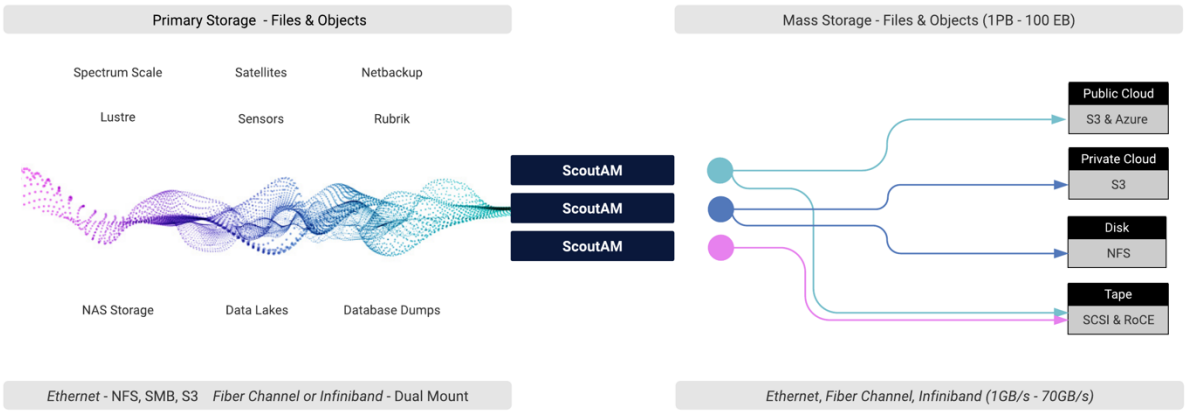


In addition to high streaming throughput performance, ScoutFS includes a powerful indexing capability that enables extremely fast searching of indexed attributes created by ScoutAM, as well as user-defined file and object attributes. The speed of ScoutFS metadata searches is only relative to the number of files matching the search criteria, and not the total population of files. Even searches that require walking the namespace are fast because the resources of the entire cluster are available to execute the search.

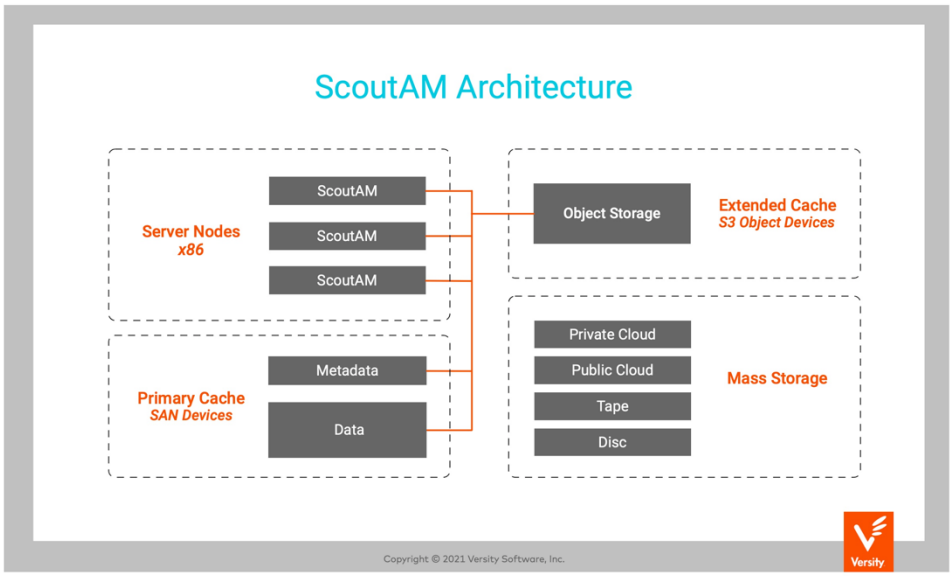
Interfaces

On the front end, ScoutAM interfaces with existing storage systems and enterprise business applications through a standard POSIX file interface as well as an S3 object interface. The POSIX interface allows the platform to be mounted on a network via NFS or Samba. Data may also be ingested using a "direct copy" mechanism by installing ScoutFS locally on servers hosting other filesystems. Versity has plans to release a ScoutFS Linux client capable of reading and writing to the ScoutAM cluster without participating in metadata management. For S3 object storage, ScoutAM includes a scale out S3 object gateway for ingestion and archiving of object data to tape.

On the back end, ScoutAM interfaces with mass storage devices though SCSI, iSCSI, RoCE, and native S3 & Azure object storage protocols.



Hardware Architecture



The ScoutAM platform is deployed on standard hardware components in five functional categories.

1. **Server nodes** - run the ScoutAM application and the ScoutFS kernel filesystem.
    - Linux x86 servers with FC/IB and Ethernet ports
    - A typical configuration would include 3-5 mid range server such as Dell R740
    - 2 Sockets
    - ~ 128 GB RAM
    - 1TB local storage for OS and system files
    - FC or IB ports for connectivity to shared block storage (components 2 & 3)
    - FC or IB ports for connectivity to Mass Storage (tape drives)
    - 10-100GbE ports for connectivity to primary data sources
    - 1GbE or greater ports for the cluster management network

2. **Metadata storage** –persistent high-performance storage for online filesystem metadata.
    - A dedicated SAN controller or a dedicated LUN within a SAN
    - RAID 10
    - Distributed RAID not recommended
    - Usable capacity = ~ 4GB per million files/objects
    - High random 4k IOPS performance is required

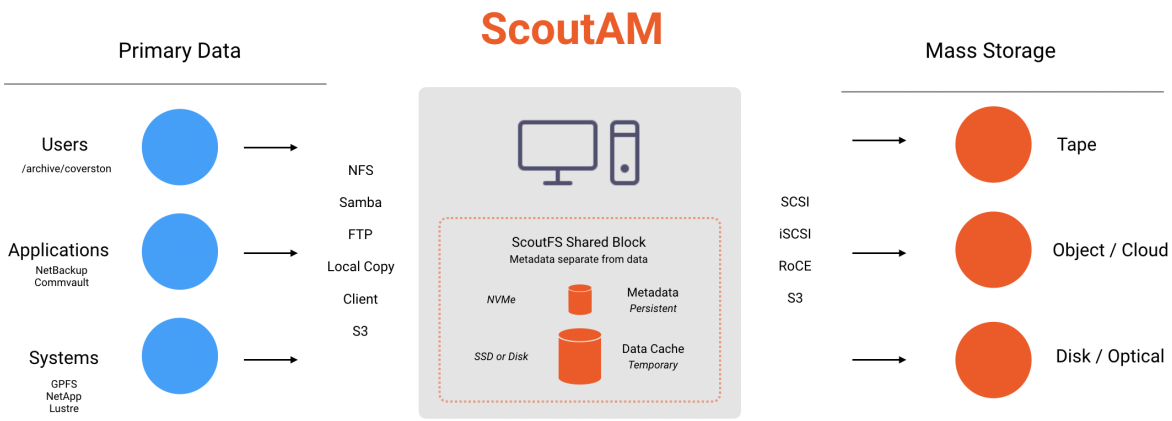
- SSD or NVMe devices required
- Compression & de-duplication features strongly discouraged

3. **Primary cache storage** – temporary high-performance storage for incoming data and data staged from mass storage resources.
    - A dedicated SAN controller or dedicated LUN's within a SAN
    - RAID5 or RAID6
    - Distributed RAID not recommended
    - Usable capacity should be adequate to accommodate at least 48 hours of peak data ingest, assuming no data is moving to mass storage destinations
    - High streaming 512k read/write performance is required
    - Flash or disk systems supported depending on performance requirements
    - Compression and de-duplication features are strongly discouraged

4. **Extended Cache Storage** – an optional low-cost capacity tier for maintaining a higher percentage of online cache data without specifying a larger SAN. Extended cache pairs well with an all-flash primary cache.
    - S3 object storage
    - Maximum throughput performance for parallel multi-object, multi-part uploads of 1GB per object and 64MB per multi-part chunk.
    - Capacity is optional and may be tuned to specific use cases

5. **Mass storage** – persistent low-cost storage for data protection and preservation copies.
    - Tape libraries and tape drives from all major vendors including IBM, SpectraLogic, and Oracle.
    - S3 compatible on premises object storage from all major vendors
    - S3 compatible optical libraries from all major vendors
    - Amazon public cloud
    - Azure public cloud
    - Google Enterprise public cloud
    - Wasabi public cloud

How it works



Writing Data

Data is ingested from primary sources to the ScoutAM Cache Filesystem (ScoutFS) over industry standard POSIX file and S3 object protocols. Primary data sources include end users, analytics platforms, applications, scratch filesystems, production databases, and utilities that generate large volumes of data such as enterprise backup products.

As files and objects are ingested, metadata is separated from data and written to a flash device where it remains online permanently for browsing, searching, and use by applications. Mass storage information such as the number of copies and the location of the copies is stored within the ScoutFS metadata.

Data resides in the cache for a configurable time interval or until the storage pool reaches a predetermined capacity threshold at which time policies are applied to efficiently move data to mass storage systems.

Cache space on shared block devices is managed between high and low thresholds that are specified by the site administrator. ScoutAM automatically manages space to balance new incoming files, archiving activities, and data retrieval. Archived files are optionally released (removed) from the cache only when the high threshold is reached so that files remain accessible on the fastest storage whenever possible. Storage administrators may specify that files with certain attributes should always remain in the cache for fast access.

Reading Data

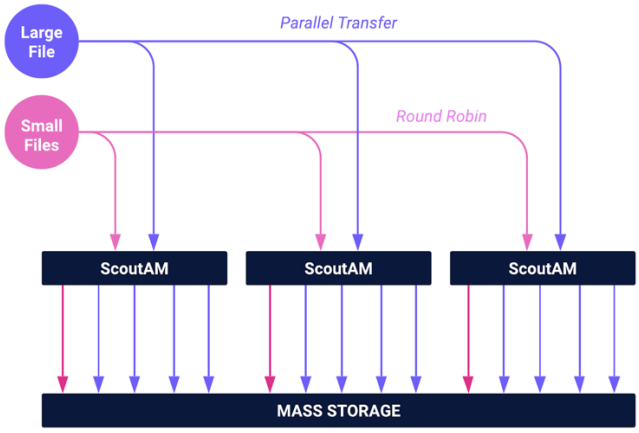
When data is requested by an application or user, it enters a process called staging. Staging activity is fully automated and transparent to applications, although there is visibility into the stage queue for the storage administrator if needed. Staging order is set by policy and may be configured to favor the copy that is most readily available. This is usually the copy on the fastest media, but this depends on site specifics like connection speeds and access fees. Like archive workloads, stage workloads are sorted and optimized to enable efficient media handling and maximum throughput. Stage resources may be managed to ensure optimal system availability. For example, the number of tape mounts or drives utilized by a specific user or by a specific set of files may be limited.

Policy Engine

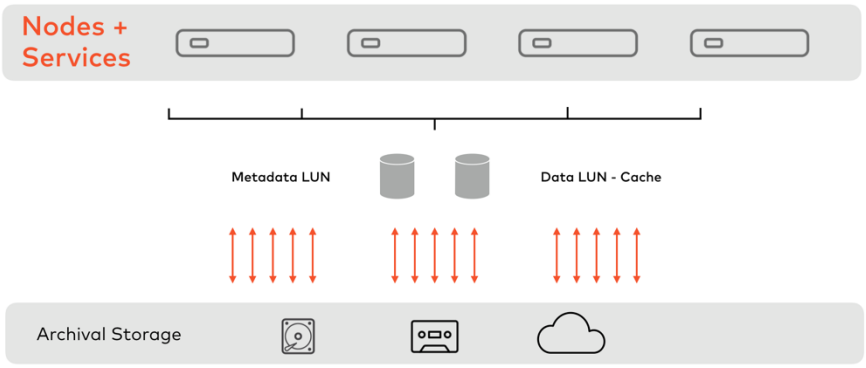
The ScoutAM Policy Engine controls the number of data copies, grouping of data, and placement of data on premises and in the cloud.

Different sets of automated policies are configured in advance and then applied to cache data to generate the desired number of copies and orchestrate writes to a composable mix of mass storage destinations. Policy definitions enable sophisticated control over the grouping of data by user, file type, application, size, group, creation/modification time, access time or any combination of these attributes. Policy definitions also set boundaries on the time that can pass prior to the creation of copies, as well as boundaries on the data set size. The ability to apply differentiated policies to coalesce random incoming data streams into efficient streaming data sets is one of the core functions of the platform.

Data sets are containerized into the open source GNUTar format, for read/write efficiency. This format has distinct benefits for small files, which are grouped and then written to mass storage, in package sizes that are optimal for obtaining maximum throughput of the target archive storage devices and media. When retrieving data, the system can recall files or objects individually without reading the entire GNUTar file. Large files may be divided into segments of configurable size and then efficiently steamed to tape drives in parallel. This capability allows ScoutAM to stripe large files across an arbitrary number of drives simultaneously. ScoutAM allows files that were written in parallel across many drives to be read back with a different number of drives so that administrators can control drive resources based upon workloads. Small files are written from all nodes using a round robin algorithm.



Archive resources can include any combination of tape, private cloud, public cloud, and disk storage systems. The archive resources are specified by configurable policies. For instance, a customer could specify a configuration for a specific user or file type that would write copy 1 to a tape library, copy 2 to a different tape library, copy 3 to AWS, and copy 4 to a private cloud system at a remote site. Storage resources may be reserved and prioritized, making things like placing certain types of data on separate devices within the same data collection possible. It is also possible to ensure that multiple file copies written to the same tape library are always written to different pieces of physical media.



How it is priced

ScoutAM is available through a monthly, quarterly, annual, or multi-year licensing and support subscription. The price level of the subscription is determined based upon a comprehensive evaluation of the site requirements and is then fixed for the term of the subscription and does not change based upon the number of servers, number of cores, number of users, or the aggregate quantity of data stored. The only variable pricing element is a pre-determined annual growth factor. The annual growth factor is an annual percentage increase in the base subscription fee determined in advance. The Versity fee structure is 100% deterministic, fully transparent, and independent from storage capacity.

Conclusion

ScoutAM is the first scalable, modular, and efficient platform for cost efficient management of exabyte-scale data collections. The company behind the product is independent and uniquely positioned to serve the needs of customers. For more information about ScoutAM or to arrange a briefing, please contact us at info@versity.com

About Versity

Versity is an independent, software-defined mass storage company focused on rapid innovation and long-term growth. We build scalable, modular and efficient exabyte-scale data storage solutions and aggressively invest in new technology. Our customer-friendly business model, exclusive focus on mass storage and highly rated customer support set us apart from the legacy storage vendors.