The Role of Active Archive in Long-Term Data Preservation

September 19, 2016

Access to all your data, all the time.

Active Archive

- Access to all your data, all the time
- Open systems offering effortless means to store and manage all their data
- Address the key underlying requirements of an Active Archive
  - Ease of Use
  - Scalability
  - Cost
  - Compliance

Long Term Preservation

- Typically longer than 90 days or much longer
  - Justifying an approach other than leveraging active workflow layers
- Sometimes for compliance
- Sometimes for content value
- Sometimes for both content value and compliance

When Archive is Justified

- When an archive solution offers material benefits, and meets all requirements
  - Economic benefits can be substantial
  - Can enable user access to more data to yield greater productivity
- When an archive solution fixes an existing problem such as a broken backup window or hard to access retained content
- Key costs and functions must be assessed
  - Primary Storage
  - Protection Storage
  - DR Storage
  - Protection Software
  - Archive Software
  - Archive Storage
  - Backup window
  - Retained data access process

Active Archives are Needed Everywhere

Government and Defense

- Surveillance, Forensics
- Legislative records
- Infrastructure analysis and development
- Enforcement records

Media and Entertainment

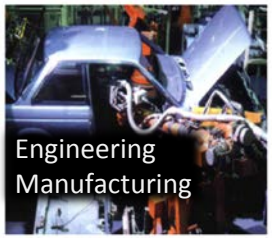
- Production Assets
- Transcoding
- Distribution Assets
- Raw Footage

Education, Research, Medicine

- Campus central archive
- Genomics analysis
- Particle physics
- Medical records

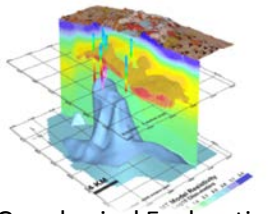
Finance, Insurance, Legal

- Transcactions logs
- Electronic trading logs and analysis
- Private records
- Case history



Engineering Manufacturing

- Sensor generated data
- Rendering and modeling output
- File and print
- Manufacturing quality and log analysis



Geophysical Exploration

- Seismic Analysis
- Climate logging and analysis
- Planetary-solar relations

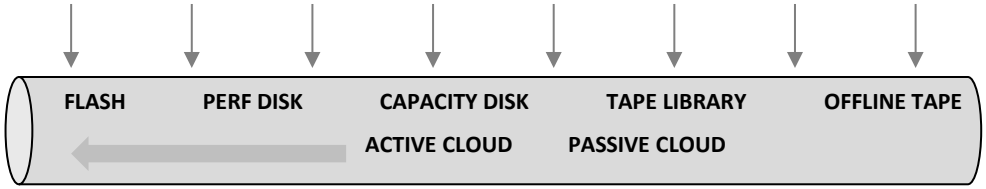
ACTIVE Archive

Access to all your data, all the time.

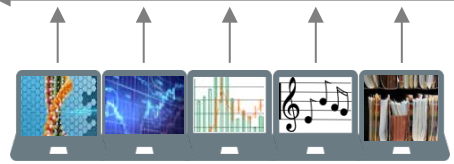
Storage and Workflow

Sometimes ***external*** ***processes capture or create data*** from sensors, cameras, machine generated data, transactions, etc.

Data is ***ingested*** into, or ***created*** in, a storage environment



Data is migrated
To meet process performance, access and budgetary requirements



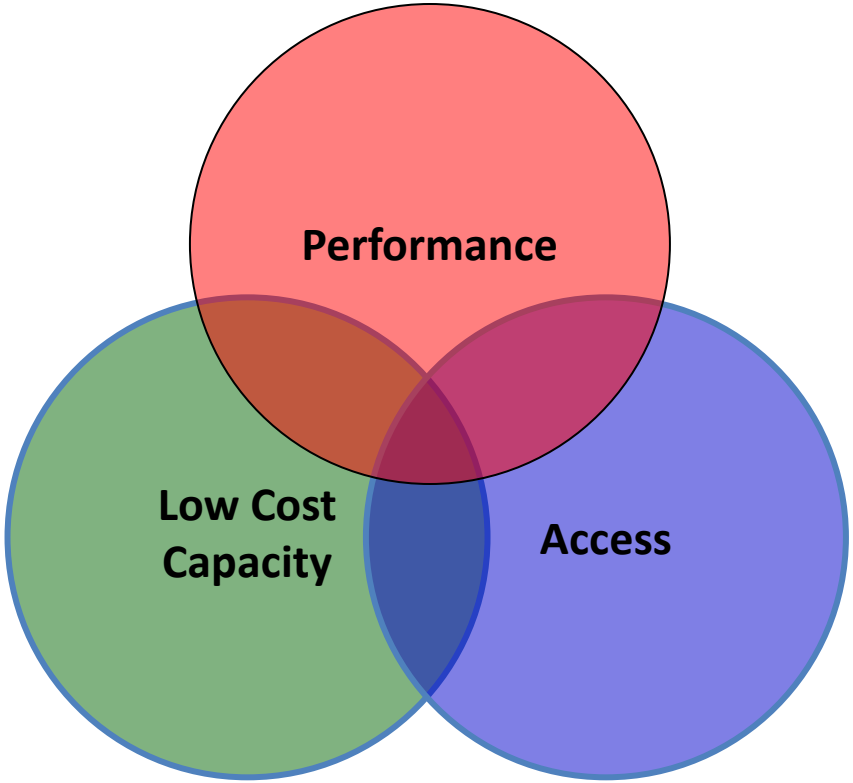
Applications/People/Processes operate on data
leveraging CPU and Storage resources appropriate for each process

Workflow ***Archive***

ACTIVE Archive

Access to all your data, all the time.

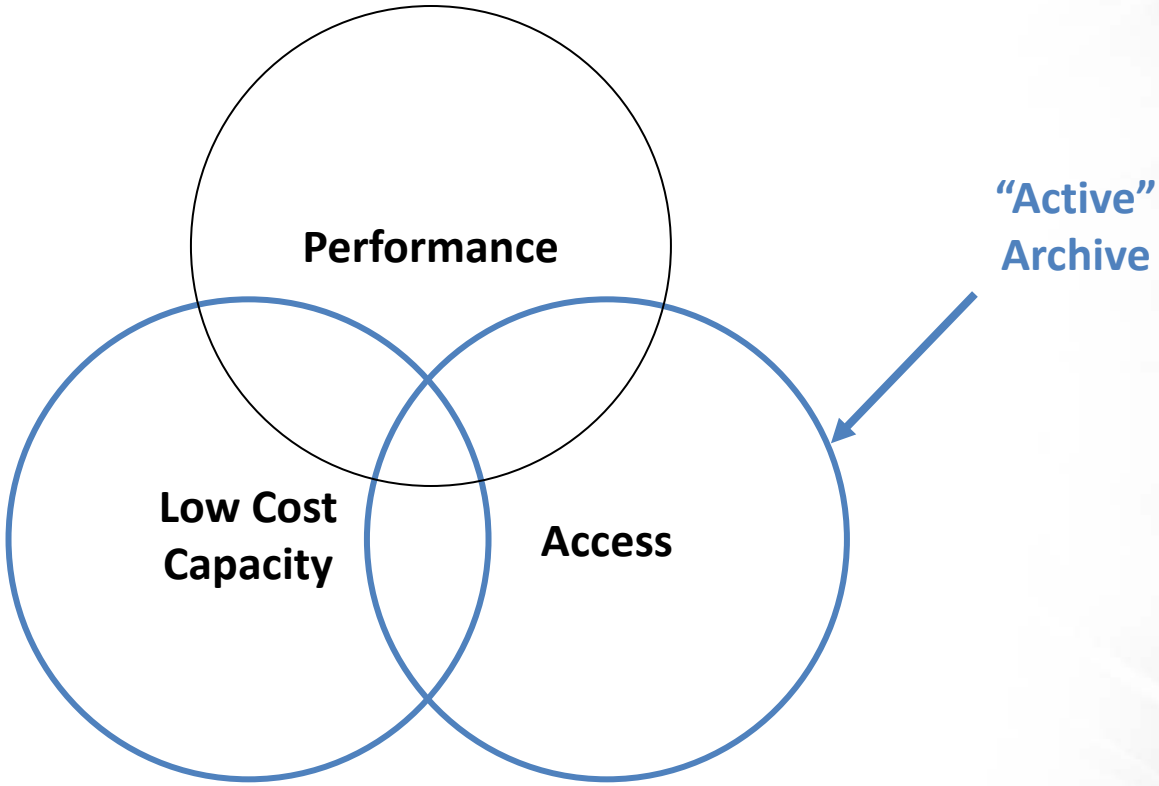
Retention Strategies Must Strike a Balance



ACTIVE Archive

Access to all your data, all the time.

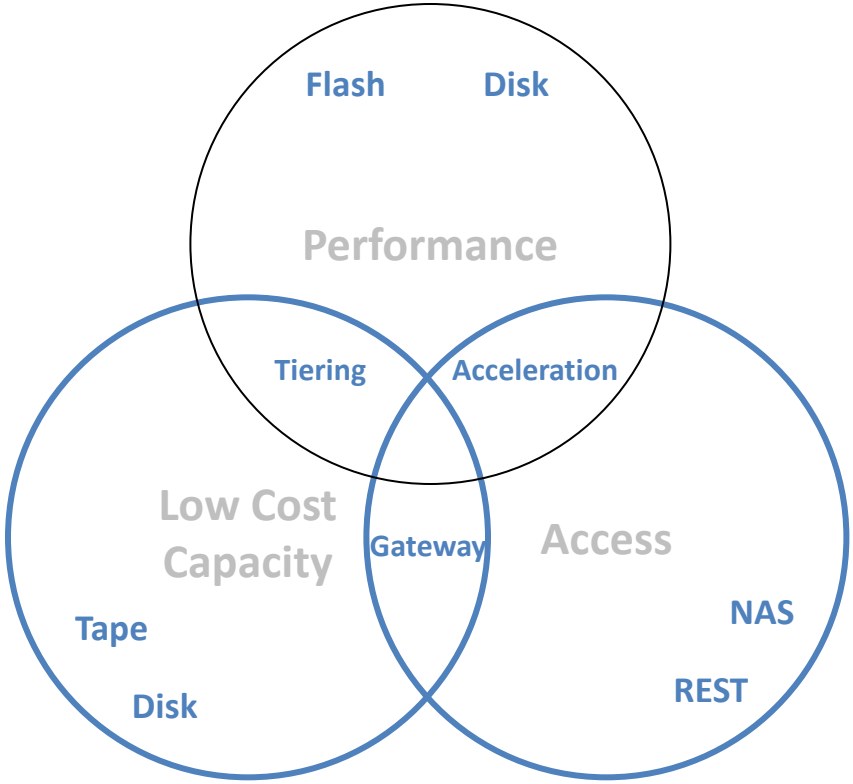
Active Archives Must Provide Low Cost and Active Access



ACTIVE Archive

Access to all your data, all the time.

Technology Choices are Critical



ACTIVE Archive

Access to all your data, all the time.

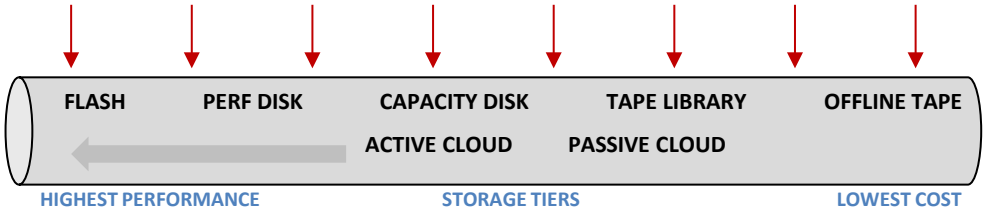
Common Attributes of Archive Storage Targets

- Tape
  - Lowest cost per TB
  - Latencies can include cartridge load time (30+ seconds)
- Public Cloud
  - Lowest entry cost
  - Archive services may carry significant latency and retrieve cost penalties
  - Monthly payments often amount to higher investment over time
- Object Storage
  - Usually include forms of multi-site protection such as replication and erasure code
  - Erasure code protection can be more cost effective than traditional RAID replication
- Gateways
  - Sometimes gateways offer substantial performance cache as a front end to high latency targets
  - Can change the world by enabling easy deployment of harder to connect targets (tape, cloud, object)

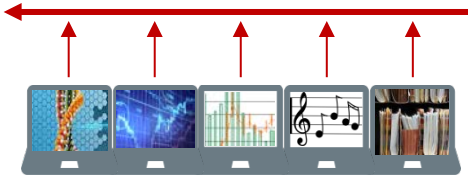
Users need data to move throughout its life

Sometimes ***external*** ***processes capture or create data*** from sensors, cameras, machine generated data, transactions, etc.

Data is ***ingested*** into, or ***created*** in, a storage environment



Data is migrated
To meet process performance, access and budgetary requirements



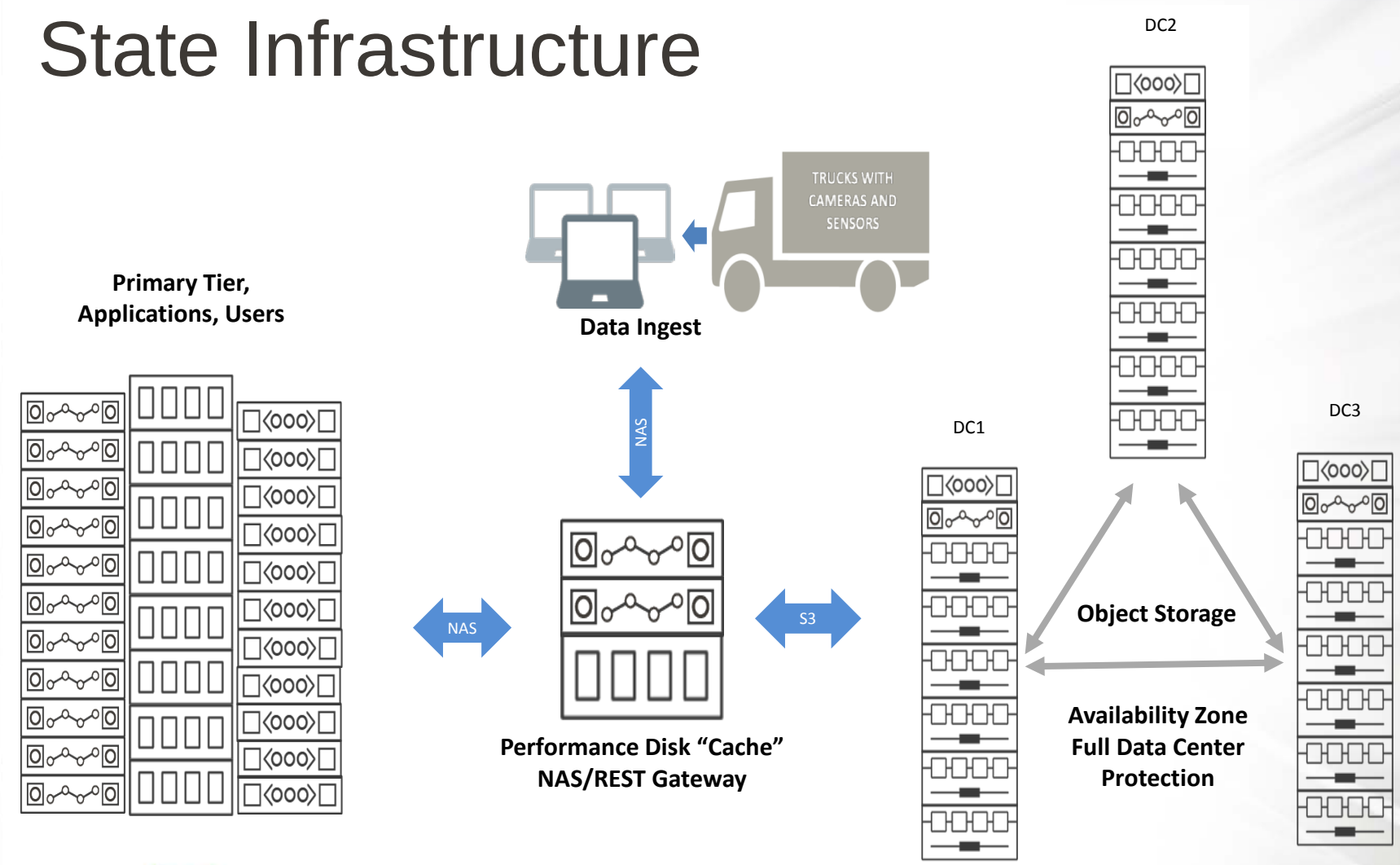
Applications/People/Processes operate on data
leveraging CPU and Storage resources appropriate for each process

Workflow ***Archive***

ACTIVE Archive

Access to all your data, all the time.

State Infrastructure

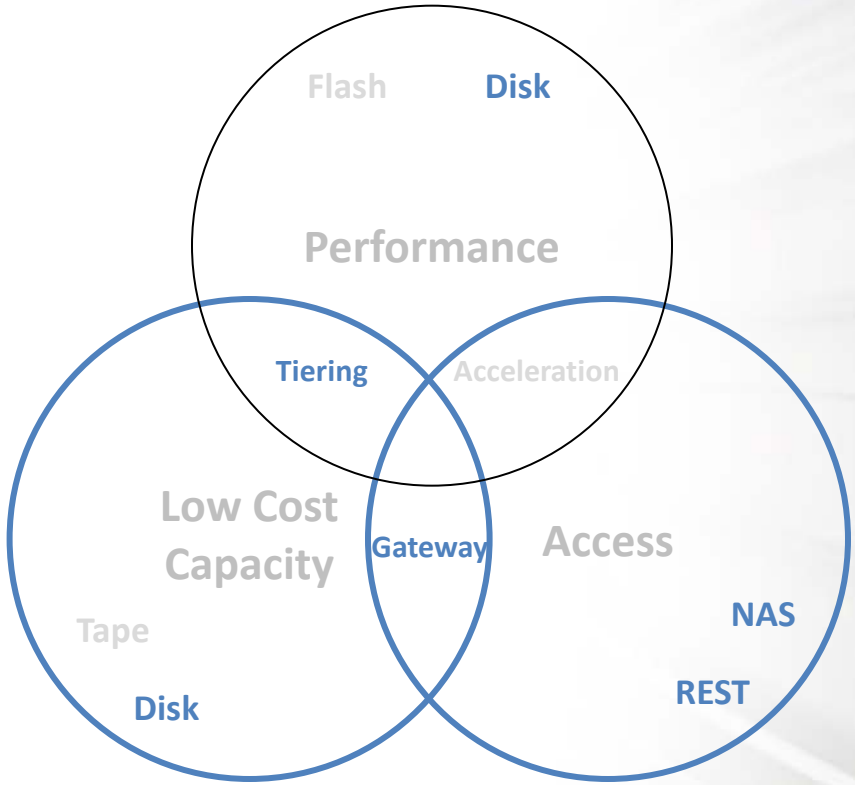


ACTIVE Archive

Access to all your data, all the time.

State Infrastructure

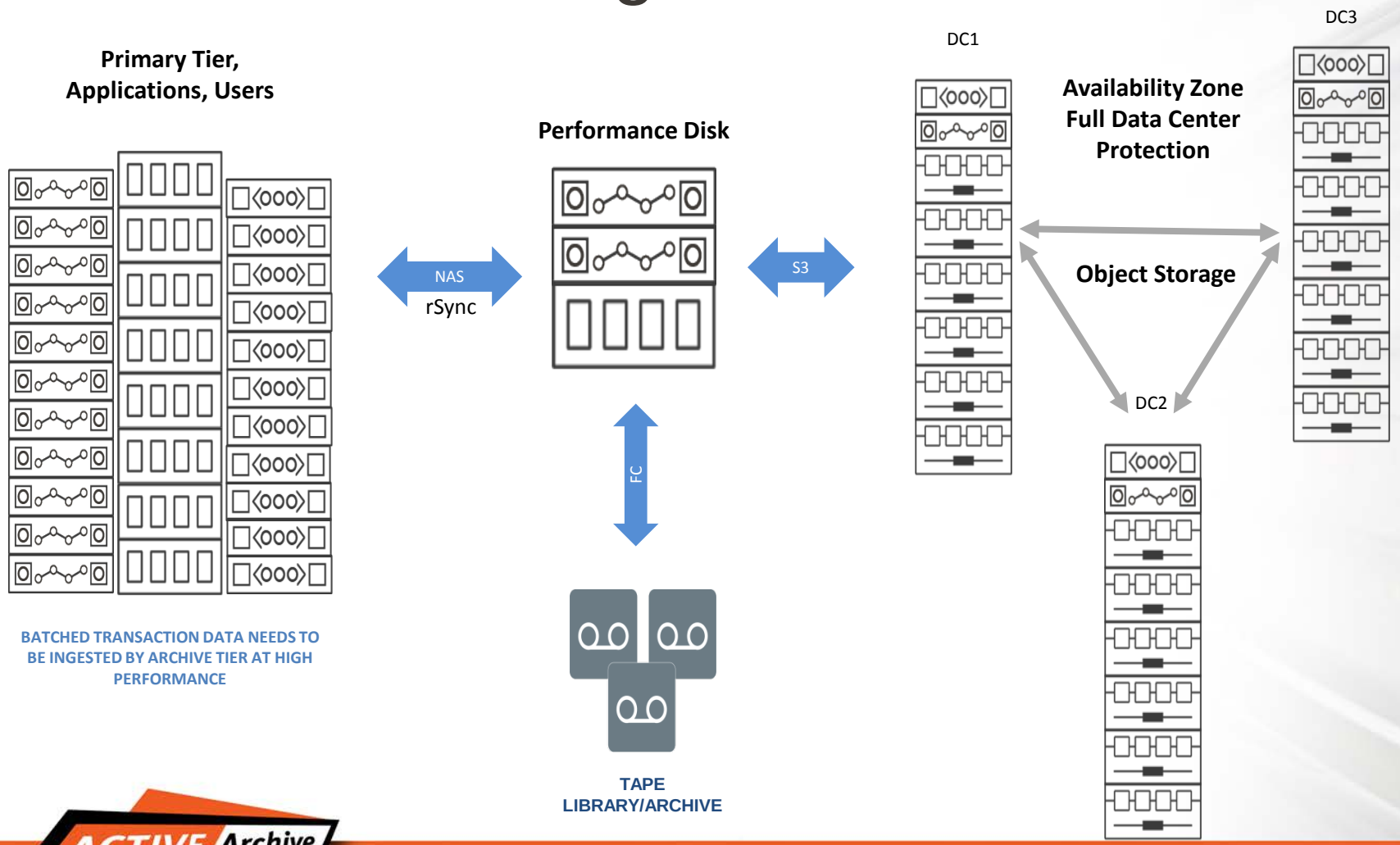
- Ingest captured data from ingest station over NAS to disk cache
- Migrate immediately to capacity archive object storage
- Retrieve when needed with intelligent NAS presentation of all archived data



ACTIVE Archive

Access to all your data, all the time.

Securities Trading

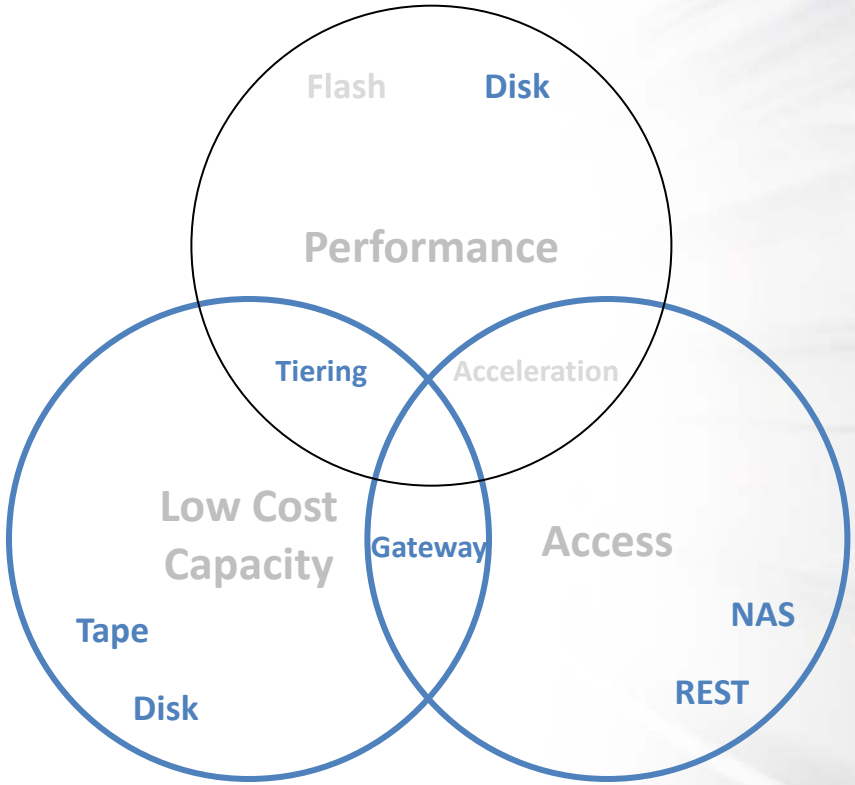


ACTIVE Archive

Access to all your data, all the time.

Securities Trading

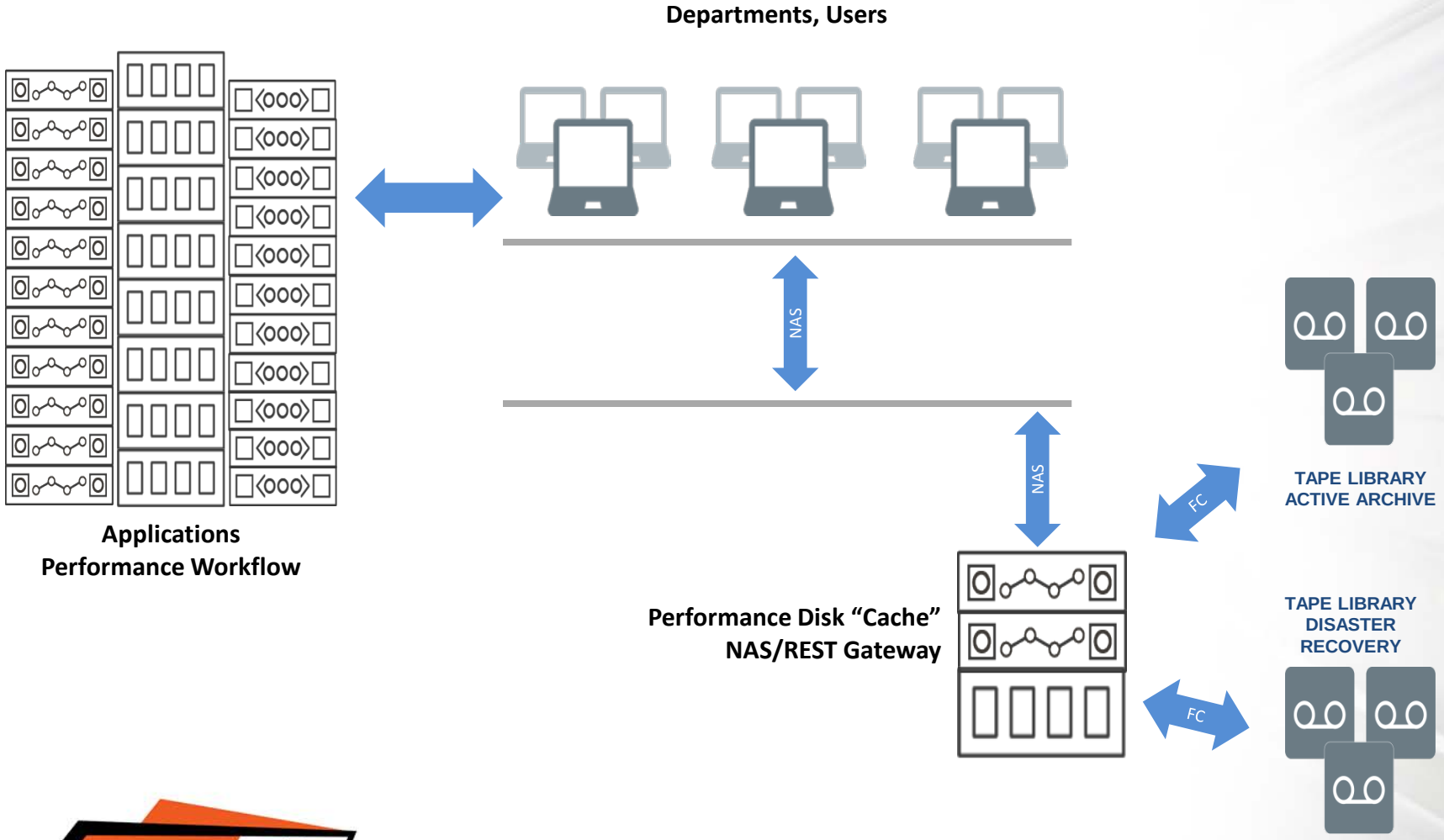
- High performance daily ingest via rSync to NAS disk share
- Long term retention for active retrieval and analysis on object storage
- Offline and compliance retention on remote tape



ACTIVE Archive

Access to all your data, all the time.

University

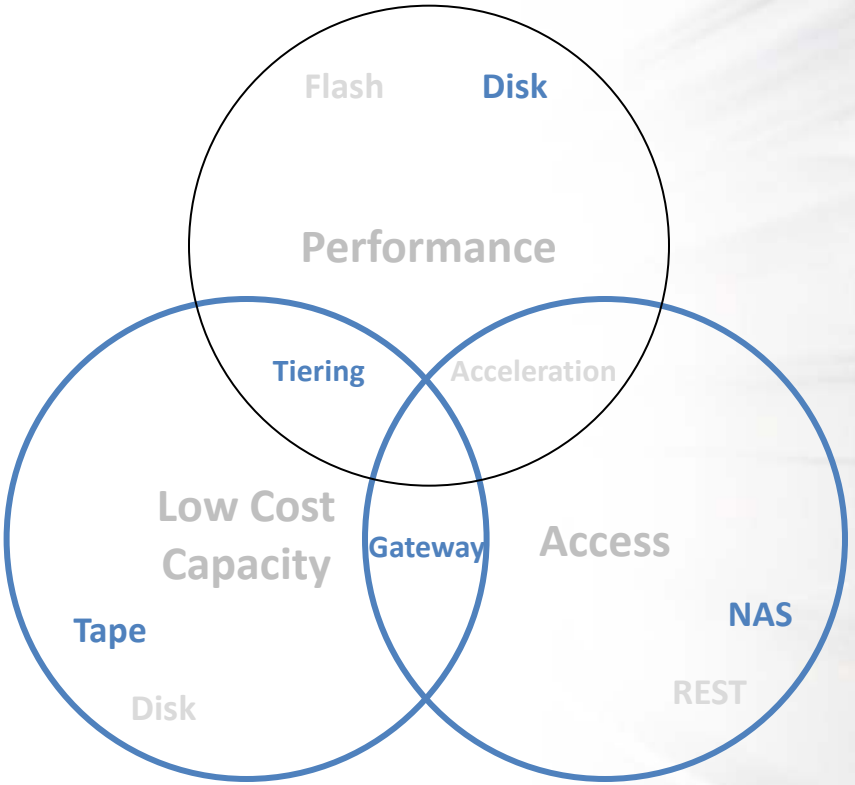


ACTIVE Archive

Access to all your data, all the time.

University

- At will movement to and from archive NAS disk shares
- Aging files tier to tape. Users see files in original share location regardless of media location.
- DR and compliance retention via 2nd remote tape copy

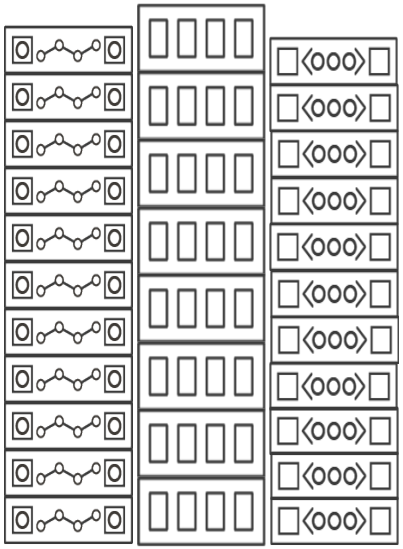


ACTIVE Archive

Access to all your data, all the time.

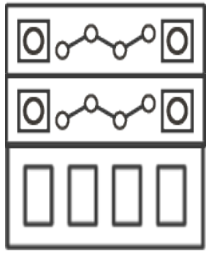
Media Production and Distribution

Production, Distribution, Asset Management



High performance retrieval of active content.
Built in, seamless, non-disruptive protection , DR, Scale

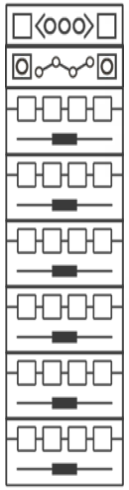
FC



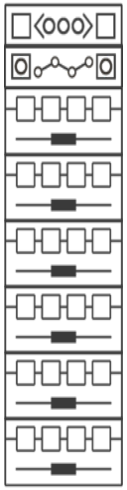
S3

Flash and Disk Workflow Protection/Archive copies to Object Storage

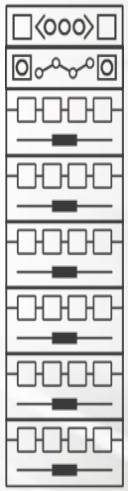
Denver



Los Angeles



New York



Object Storage

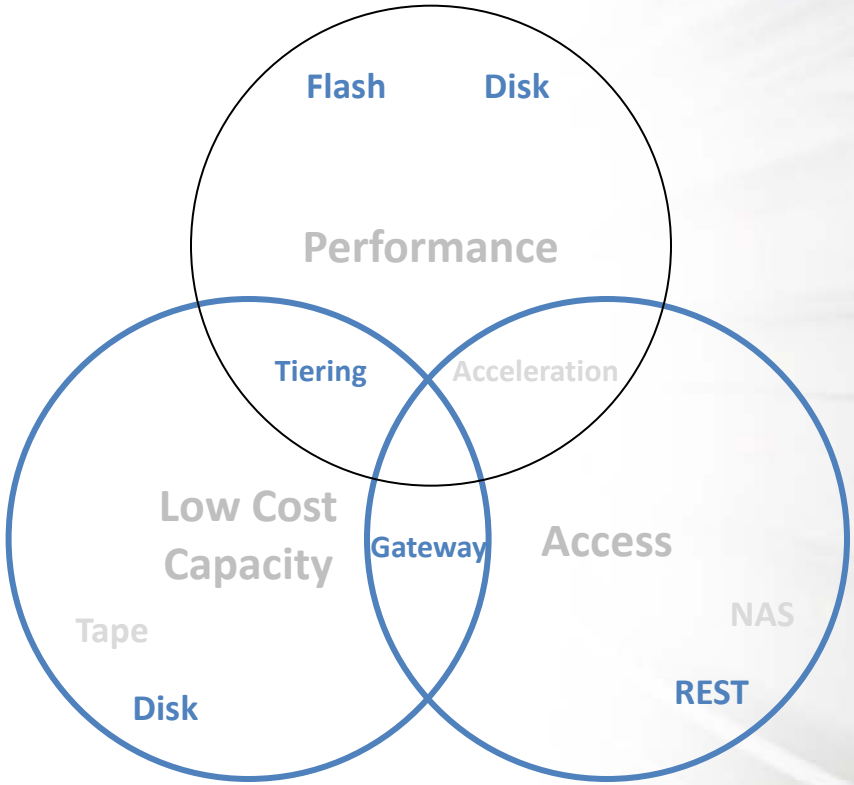
Availability Zone Full Data Center Protection

ACTIVE Archive

Access to all your data, all the time.

Media Production and Distribution

- Integrated workflow with automated protection
- Multi-geo object storage disk archive and DR



ACTIVE Archive

Access to all your data, all the time.

Other Key Considerations

- Cloud
- Data Movement
- Reporting
- Compliance
- Scale

Cloud

- Is just another RESTful target
- Is just someone else's datacenter
- Often
  - Lowest cost of entry (storage)
  - Higher storage costs in the long run – particularly for active data
  - Better if workflow is in the cloud

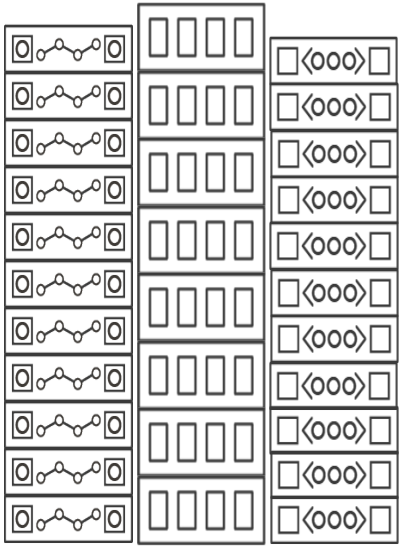
Data Movement

- There are two common areas of data movement
  - Move to archive infrastructure
  - Manage within archive infrastructure
  - Provide acceptable ongoing access models

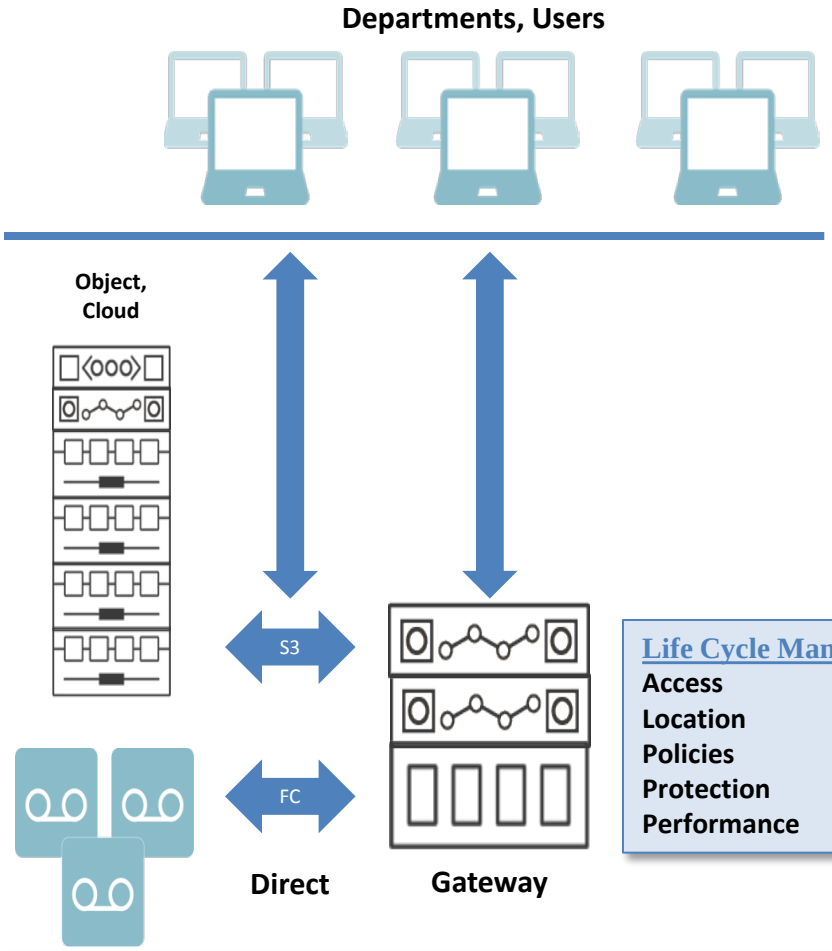
- Move to archive
  - High performance storage is no longer the best resource use for this content

- Manage within archive
  - Meet access requirement such as location and latency
  - Protect to durability and other compliance requirements
  - Meet cost requirements

Data Movement



Applications
Performance Workflow



Archive

File crawlers
- Policies
- Content
- Attributes

Location
- Project
- Geography

User selection

Life Cycle Management
Access
Location
Policies
Protection
Performance

ACTIVE Archive

TAPE LIBRARY
DISASTER
RECOVERY

Access to all your data, all the time.

Data Movement

- Today, move to archive and lifecycle movement are often two different operations
  - Move to archive can be as simple as drag-and-drop, or can have complex data aware policies
- Separate movement solutions may be typical if not necessary for heterogeneous environments
  - Optimizing cost and performance
- Homogeneous environments may come with comprehensive data movement solutions
  - Minimizing potential complexity

Compliance and Integrity

- It's not always about the storage target
  - Access control and event logging software layers may be what's needed, and storage can be just storage
- WORM
  - Some storage hardware is fully compliant with enterprise or government regulations
    - CD-R, DVD-R, LTO-WORM
  - Some software layers can add compliance WORM functionality where the storage system does not meet those requirements
- Ongoing data integrity checking
  - Upon write
  - Upon read
  - Periodically throughout data life

Scale

- A central tenet of an archive solution
- All content ends up here – the ability to scale is an imperative
- Tape libraries, Object Storage, and Cloud all have inherent scale models
- It is critical to understand the scale and limitations of data presentation layers
  - Object count
  - File count

Reporting

- Archives often span across functional organizations
  - The best economy of scale may achieved when archive consolidation is leveraged
- Functional organizations manage individual budgets

- Utilization reporting is often a key requirement for IT to enable charge-back
  - Capacity per tier
  - Department
  - User
  - Throughput

Format Migration

- Archives over 10 years in duration may need to consider format migrations
- Software, physical formats, file system updates
- The good news, solutions and services are emerging to address these very issues
  - Many software products can migrate physical or logical formats
    - Tape cartridge generations
    - Proprietary software generations

Takeaways

- Active archive is a common requirement for long term retention infrastructures
  - In all industries
  - Archive solutions offer substantial economic benefits
    - May also address existing functional issues
    - Can enable access to more data
- Active archive solutions must deliver the right balance of cost, access and performance
- Many functional considerations beyond cost, access and performance often need to be addressed

Active Archive Alliance

- Promote open industry solutions
- A forum for discussing relevant topics, pain points and challenges of managing data at scale
- Develop customer-centric value messaging to evangelize and disrupt traditional methods of managing and monetizing useful data at scale
- Providing thought leadership to consumers of storage systems and solutions stacks to reduce alternative storage architecture decision risk

Leave your card; get our report.

ACTIVE Archive — Access to all your data, ALL the time.

Active Archive and the State of the Industry

Innovative Solutions Drive New Markets and Use Cases

Abstract

The Active Archive Alliance launched on April 27, 2010 as a collaborative industry association formed to educate end user organizations on the evolving new technologies that enable reliable, online and efficient access to their archived data. The Active Archive Alliance goal is to align the education and technologies needed to meet the rapidly increasing requirements for archival data. Alliance members strive to extend solutions to the greater general IT audience that needs advanced online data archive options. This report describes the current state of the active archive market.

The State of the Digital Archiving Market

According to the latest Digital Universe report approximately 90% of the data in the world today was created within the past two years and the vast majority of it is not changed once created reaching archival status in a relatively short period of time. Overall, IT budgets are relatively flat, yet newly created digital data is growing at over 40% annually, and is now being generated by billions of people, not just by large data centers as in the past, mandating the emergence of an ever smarter and a more cost-effective and secure long-term storage infrastructure. Top external factors driving long-term retention requirements include compliance regulations, business risk, security exposure and the rapid emergence of Big Data analytics. For most organizations, facing hundreds of terabytes or several petabytes of archive data for the first time can trigger a need to redesign their entire archival storage infrastructure.

Exabyte ($1x10^{18}$) sized archives are now on the horizon. In the era of Big Data analytics, stored long-term data needs to be accessible for large-scale data queries long into the future. Redefining the term "archive" to become an online, accessible, affordable data management platform is necessary to solve data growth, improve performance, and to meet retention challenges going forward. Leveraging innovative and integrated solutions for disk, tape, and the cloud will be required to fully enable the active archive experience. An active archive can meet the challenges that lie ahead for data archiving.

Backup and Archive Are Different Processes

It is important to understand the differences between backup and archive as they are entirely different processes and have different objectives. Surprisingly, the differences between these processes are not well understood. Backup is best utilized during an emergency, when data

opies of data are kept to mitigate short-term risks. Archives ent, compliance and information re-use purposes.

tection) The backup process creates a copy(s) of data for used to restore the original copy after a data loss or data ed and updated frequently to account for and protect the sets. Backup data is typically overwritten and is difficult to historic relevance.

ger-term retention) The appropriate process of archiving new location(s) and refers to data specifically selected for an one copy of archive data should exist since having a ta presents an exposure should the only copy become ally secure and is not overwritten. Archiving is an ideal data with historic value and it is easier to search.

n system files, catalogs, indices, OLTP and directories, the , of most data types declines as the data ages and typically or less. Archival data is accumulating faster than ever, as kept indefinitely. A customer survey from the 2013 SNIA cated a retention period of 20 years or more is required by r more retention is required by 57% of respondents making

ed HDDs coupled with tape's highly favorable economics e archives. An active archive provides a persistent online one or more storage technologies (flash, disk, tape *and* that gives users a seamless means to store and manage alized storage pool. Data archives are indexed and have rts of files can be easily located and retrieved. An active a cache for higher performance, with tape providing online ounts of archival data. The increasing value of archival data storage systems with improved management and security

use your existing storage equipment to build an integrated d can incorporate enhanced file systems such as LTFS se who do not want to put together their own repository, active archive appliances or "NAS heads" with various file ape library back-end. Active archive solutions support file, aking them extremely versatile for any data type.

transformation, driven by much larger file sizes and new access requirements. Object storage enables IT managers to organize archive content sooner.

2

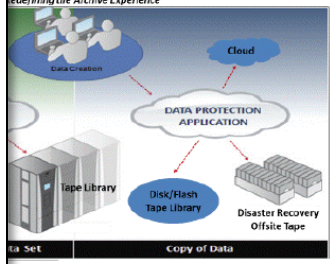
es can transform an archive into an 'active' archive that is e storage and low-cost disk or tape.
e simplicity *and performance* of disk with the economics of pe in a highly scalable solution.
a billion or more files and holds more frequently accessed

nes flash, disk, tape, in various configurations with data ill deliver enhanced cloud services.

you can easily implement an active archive.

The Active Archive

Redefining the Archive Experience



Source: Active Archive Alliance, Horison, Inc.

include:

ers with file-level access to all of your data, all the time – ether your organization stores active archive data in a rage cloud, on tape or in an off-site public cloud, an active ent the flexibility to select the correct storage media while tion access to the data.
reducing the total amount of data to manage, or partitioning organizations may see substantial improvement in system further improved by using disk drives to serve as a cache

the advent of Big Data, organizations are learning the unts of archival data to gain a competitive edge in the cs demand active archive solutions.

3

ACTIVE Archive

Access to all your data, all the time.

ACTIVE Archive

Thank you.

Access to all your data, all the time.