TAPE at EBI

Introduction

Previous document

The current tape solution used by FIRE needs to be replaced or substantially improved. Currently tape can only be used by FIRE, and if tape needs to be used by TSC then it makes sense to study all the business cases identified and received that could benefit of a new tape solution.
This document gathers the following:

- FIRE tape needs
- Benefit of expanding current VM backup into tape
- Allow other TSC teams to save data to tape
- Different Tape infrastructure improvement costs and features

Definitions

The following definitions are used outside of the EBI and should be understood as standard definitions for those terms.

Backup

Like a cp or rsync. You end up with two copies of a file, useful for daily used files that you do not want to lose.

Archival

When a file is no longer used on a daily basis, you move it to tape/cloud. Like a mv you only have one copy of the file.

Hierarchical Storage Management (HSM)

A data storage system that automatically moves data between high-cost and low-cost storage media.

Business Continuance (BC)

Depends on each company, but aims to prevent production disruption. In our case we have 2 DataCentre (DC) and if one fails the service would be provided from the other one.

Disaster Recovery (DR)

This one has a defined policy which would be triggered after a devastating event, including data loss and service disruption. It has two important measurements, RTO and RPO.

Virtual Tape Library (VTL)

A virtual tape library (VTL) is a data storage virtualization technology used typically for backup and recovery purposes. A VTL presents a storage component (usually hard disk storage) as tape libraries or tape drives for use with existing backup software.

Tape LTO

Reliability

When specifically talking about LTO , LTO tape provides a reliable, offline protection against on-line data corruption. Compared to the shelf-life of disk (3-5 years), LTO tape can withstand the test of time, providing users with a shelf-life of up to 30 years. The protection is also there when you need it, offering a strong error protection and correction algorithm, with an error rate of 1 bit in 1x10E19 bits (12500 Petabytes). That's four orders of magnitude better than disk, whose error rate is 1 bit per 1x8E17 (100 Petabytes). Once your data is written to the tape there's no need to worry about whether it's there, it's immediately verified by the LTO "Read after Write" technology.

No vendor locking

LTO technology is an "open" format, meaning that many different competing companies are producing LTO drives and LTO cartridges that are compatible with each other. Each generation is also compatible with its previous generation, so you know that you'll be able to make the most of your investment as technology needs change.
LTO specifications are an open format designed for interoperability.
Investment protection is assured through backwards compatibility.
LTFS provides easy data access and management allowing for easy file share.

OUR TAPE Libraries

TAPE_TYPE	Tape size	Full TB	Available TB
3592-JD	9	10134	756
LTO-6	2.2	7697.8	0
LTO-7	5.4	6798.6	124.2
LTO-7-M8	9	15129	3582
		39759.4	4462.2

COUNTA of TAPE_STATE	*TAPE_STATE*				
TAPE_TYPE	Corrupt	Empty	Full	Offline	Grand Total
3592-JD	28	84	1126	240	1478
LTO-6			3499		3499
LTO-7	18	23	1259		1300
LTO-7-M8	21	398	1681		2100
Grand Total	**67**	**505**	**7565**	**240**	**8377**

1500 LTO6 have been exported and sent to a safe place.

IBM TS4500

From wikipedia :

TS4500 Tape Library[edit]

High density tape library supporting Linear Tape-Open (LTO) 5 and 6 or TS1140 and TS1150 drives. Can scale up to 35.5 PB of native capacity with 3592 cartridges and up to 11.7 PB with LTO 6 cartridges. Supports up to 5.5 PB in 10 sq ft.[27]

From IBM page Store up to 351PB (1,053PB compressed) per library with IBM 3592 cartridges

Spectra T950

FIRE tape requirements

Currently FIRE uses an in-house solution to interact with TAPE called Object Tape Archive OTA. To fulfill its requirements FIRE requires the following capabilities, the ones available with OTA have been underlined .
Granularity on recovery, we should be able to recover one single file from tape without recovering the full tape. We currently use TAR and with bigger tapes we would benefit of LTFS to retrieve single files if needed.
Compression, would help reduce the number of tapes consumed.
Migration between technologies, so we can move data from LTO-N to LTO-N+1 without using the external buffer.
S3 client or alike to access the data, to reduce bespoke development if no benefit is added.
Interface to allow manual interventions like add/remove tapes without service disruption.
Clean drive control, to allow drive clean up without service disruption.

In-house Object Tape Archive - OTA

This in-house tape solution has greatly helped FIRE growth, however it also has several weaknesses preventing its evolution.

- Does not allow any other tool or team to use the drives associated to it.
- Can not stop writing a tape without corrupting it.
- Written in python2.6 unportable to python3 without a full rewrite,
- Using a shared Oracle database with other already retired software from the same developer.
- Dependent on proprietary oracle SQL triggers not portable to postgresql.
- Forcing virtual tape association for each file before receiving it.
- Forcing full tape recovery to recover or check a single file.
- Internal tape to tape migration is not implemented.
- Its endpoints are not compatible with any specific tool or client.

This solution does not bring any specific benefit and stops other teams/applications of using tape to archive or backup their files.

Extend current backup retentions

Allow tape to be used by teams other than FIRE

Current infrastructure limitations

New tape infrastructure options

Spectra Logic Black Pearl

https://spectralogic.com/products/blackpearl/
davidt@spectralogic.com <davidt@spectralogic.com>

Status

- Waiting for spectra quotes

Pros

- Object store
- We have spectra already , can grow by adding TS4500 drives mixed enclosure
- Default setup is OTA replacement with 160TB of disk cache
- Can migrate data between tapes
- Priced by appliance, not related to size
- Compatibility matrix is very good

Cons

- BULK get and BULK PUT
- Does not support IBM TS4500 JD-3592 drives (13PB)
- Requires RIO to become more useful, is free but shows how bad their roadmap was

Eject Media
Look at Spectra RIO

Presentation from santa clara 2015 : link

Next meeting - done

- Black pearl S3 , not good for us
- Riobroker, windows only, API is not standard
- IBM TS4500 , JD 3592 , 13PB of tape ~ 150K + frames + library

Tuesday, 16 June 2020

They will try to find a company that can get us the S3 layer implementation that we want.

Arcitecta - Black Pearl (Mediaflux)

Link
Company is a data manager.

- Customers at PB level, 2-3BP/months

Guy Grifftihs - Experience in isilon in HPC.
Emily King -

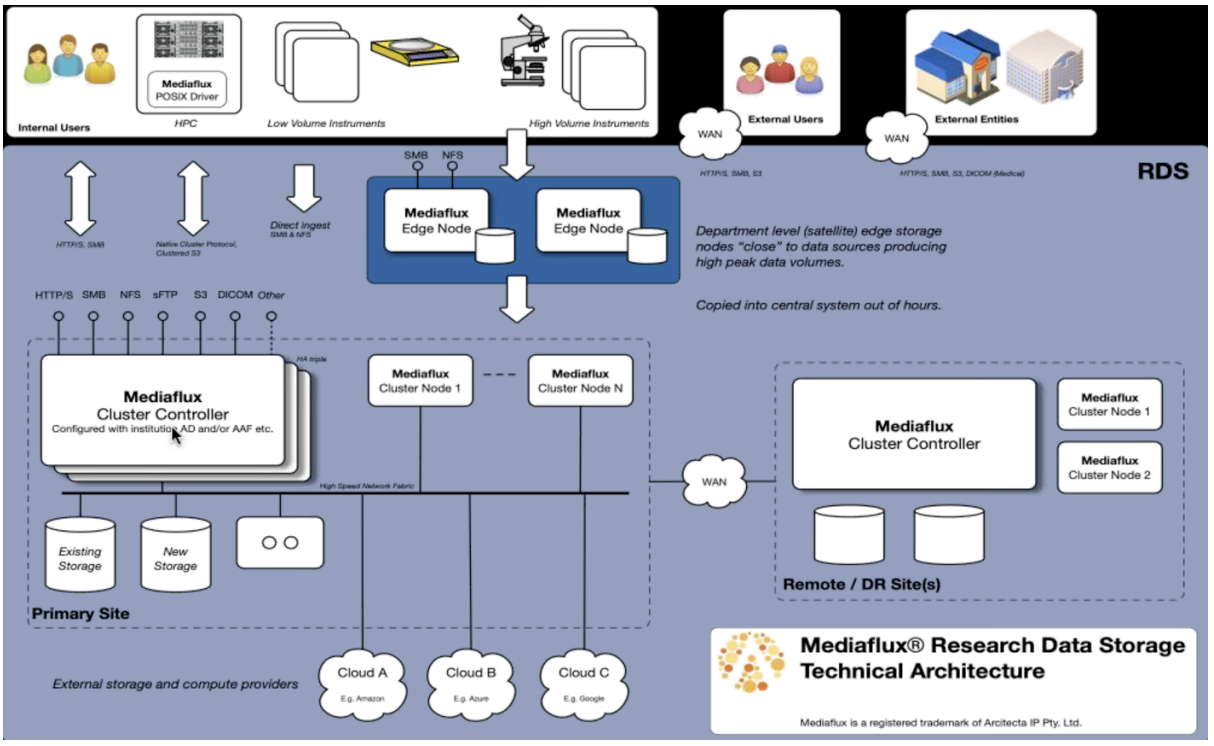
Key points

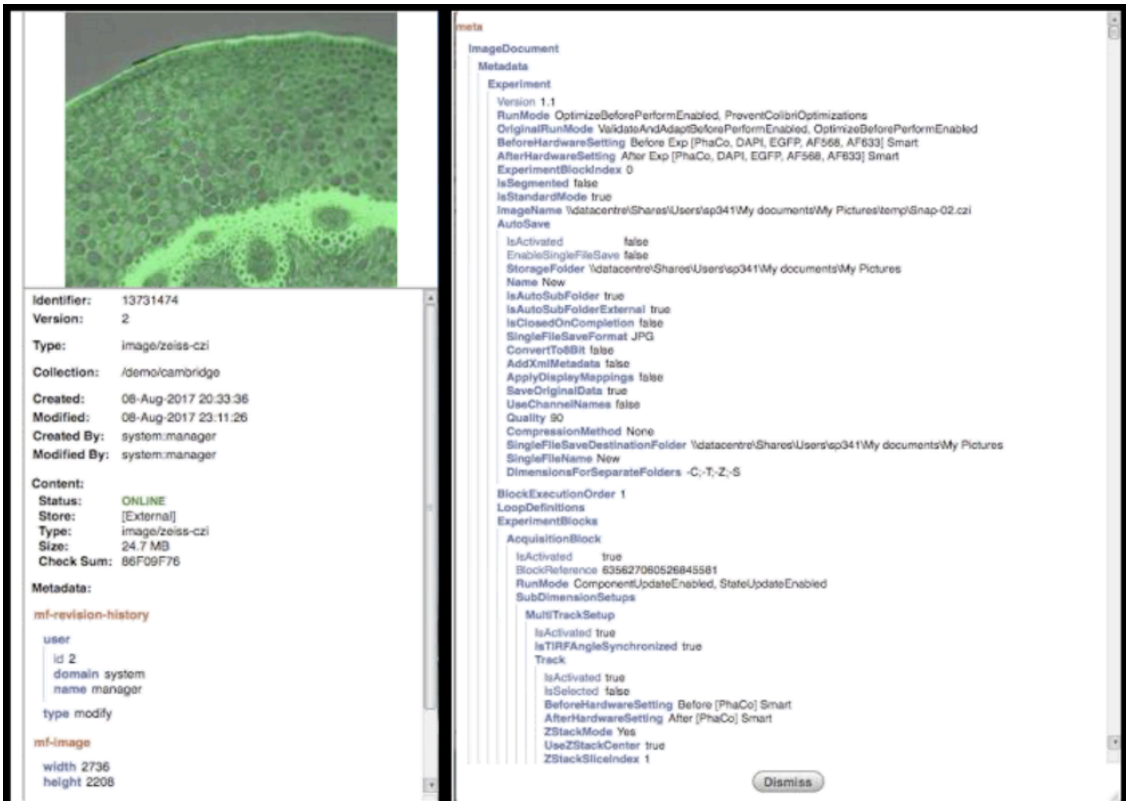
- Metadata database, with version management
- Mediaflux has a desktop application to browse the files metadata.
- Gets information from fasq files
- FAIR principles in metadata.
- Hardware migrations. The hardware migrations are done transparently because the user sees a Virtual Filesystem. Single global namespace.
- Example of a company 2014 - 2017 , move to HSM and cloud . In the presentation.
    - Up to 5 billion files
- Powerful data management and access system. Including metadata ACL .

Questions

- Does it find duplicates of files and link them together?
    - Scan : scan all and import
- The example showed CZI , is that Zukerberg?
    - MRC cambridge university.
- What are the biggest deployments you have? And what issues you had there? How did you solve them?
    - Limited only by the number of nodes. 2PB/month. 65TB/hour .
- Can it have actions embedded as API calls with the data?
- Can you show me a report of data allocation from an existing system?
- What is the limit on metadata management?
    - No limite yet, just add nodes .
- HPC IO operations management?
- What is the hardware example of one of the nodes? Virtual?
    - Controllers Dell R640 / R740 (Failover active passive)
        - 64 processes high cpu speed
        - Ram 256G
    - Edge nodes Dell R240 (edge nodes as need to move)
        - Low memory . 128G RAM

        - More threads and lower cpu speed.
        - Network interfaces.
- What about two data centres?
    - One node can be located in X , and they know where the data priority should be.
    - Can know where the physical data is, including shipping libraries and media.
- Existing systems
    - Need a migration
- What about LTO migrations?
    - Spectra is part of the problem and solution. General problem with LTO9.
    - Black Pearl would be the one migrating the LTO lifecycle.
- Licensing?
    - Concurrent users? Accessing at the same time!
        - Anyone that authenticates with the server ,UID, password,





insurgo.co.uk

DDN - DataFlow [link]

Waiting for technical presentation.
Ivaylo Dikanov - TSM & Data Migration

Dataflow

- Data mover
- Archive
- Backup
- Manages tapes , many vendors
- Scale more movers when needed.
- With an extra DATA MANAGER it can put data in tapes in extended TAR,

IBM

IBM storage team
Alan nash - product manager for ireland storage. Including tape.
Jonathan willcock. - technical presales, ibm spectrum. Object storage and scale filesystem.
Adam willis - overall storage.
Clide - storage representative . Life Science clients.

Insurgo related. They do the move from lto6 to IBM scale filesystem enterprise format.

Spectrum scale is the overall controller. With stub files can have some size.

LTO9 can only read LTO8 .
Enterprise format?

Waiting for their

PoINT

EUROSON, tape archive

In Amadeus , they use StrongLink , and Scality Zenko is looking at them.

German company based in US.
https://www.point.de/fileadmin/user_upload/datenblaetter/2019_POI_Broschuere_Archival_Gateway_20200123_web.pdf

Things to take into account

- Windows based.
- Media support ? Look out for media migration for the long run.

Point Archival Gateway

Largest - Daimler
230GB/s
100 Billion objects per S3
Scales horizontally

Erasure , ~ raid, but 0+2 in 2 data centres, or 3+1 4 tapes with checksum.
    Performance increases, since we read and write to several tapes at once.

OTACACHE equivalent buffer is optional, not mandatory.
Use internal tiering features of other tools, like storageGrid, HDFS , AWS , HCP , ...
They can fetch data from S3 buckets.

Everything is managed from the UI.
LDAP support.

Questions

?? Erasure Coding for tape storage ??
Daimler size ?
    They use HSM , maybe 2PB . They have 16 instances of Point around the world.
    20PB in a car manufacturer
    1PB a day, this has been tested. The bottleneck is the number of drives.
Architecture will require some SAN changes
Erasure, Is the storage FS proprietary ?
    In principle is their own format.
    You could read the data without our tool, they can provide the format specifications.

Overhead of 3+1 , 1Gb => 1.25
Can we start with erasure to 1 ,and then add drives and increase erasure?
    Yes. no problem.
Why the old media first? LTO6 ?
Can you deploy on our own hardware? Our own hardware, with requirements for throughput.
Do you need more drives ?
    If you want more performance, yes.
Can I add metadata to a file on upload? Yes .aws standard, 2Kb for each object.

Backroll migration process from LTON to LTON+2 , is in the roadmap but not yet implemented. They can move from one bucket to another, and if the buckets are different generations then that is all. The bucket name will be changed, but in the future will be a migration inside the same bucket(roadmap).

Fujifilm is working in a object storage tape format.

Point, they also have HSM software.

StrongLink (strong box)

30 years company.

Status

Simon Marrion
Miguel Castro

Properties & notes

- AWS , vmware, hardware
- HA setup available
- THey had tape appliance, they are going LTFS directly
- Supports microscopy metadata.
- Ingest by reference , looks interesting. No links!
- Supports LTO6 to LTOX migration.
- Like LVM but at a higher level.

At user leaving, add metadata to his files.
Query that checks users that have left for 2 months, notify about his files to GTL.
Move all users files to tape.

Also versioning.

Pros

- Supports Black Pearl , and extends it. Ingest everything on it.
    - What you wish Iron Browser was.
- When writing to tape , it can join files in containers.
- Active directory. & unix users mapping.
- Does not use links when moving data , even to tape .
- API access to export all content to other platforms.

Cons

- S3 frontend is still new. 2020

Questions

Where is it in production? Since 2015 .

- University harvard.
    - Charge departments , project management.
    - Similar research workflow.
    - 20PB managed

        - Single instance is enough , it depends on how many clients.
        -
- Self driving cars
    - Also tape
- Mexico : gas institution. Encrypted tapes.
    - 8PB on single data centre. All users using stronglink Virtual FS

Strongbox since 2010 and now retiring. StrongLink is taking over the customers. 80%
Command line API ?

Yes .

Support matrix :

Lustre file systems?
Panfs?
GPFS?
mooseFS?

Deduplication ? No, but can detect duplicate files. (sha512)
Tiering? Yes, through policies and jobs.
Number of files you can support ? proprietary database, it shardes over nodes. And then database cache is what is important.

Maximum number of files processed during hour or day ?

NExt steps

=> PoC . talk with Steven.
Prepare a requirements list for Miguel to look at.

Friday, 19 June 2020

Spectra will be supported
Ibm ts4500 is supported.

They can do the HSM + tape S3 archive.

Container of small files, better performance but each file can be fetch by itself

They need full visibility of the data. ?? what does that mean.
VFS , virtual filesystem needs to scan each mountpoint.

Smallest install: 1 node for migration
3 start cluster , HA . DR support ⇐ constellation

Training includes.
Stronglink has smartpool, aggregate storages, so it can migrate data around transparently.

Do I own the data?
Can we read the data without your tool?

Licensing by core of the servers. Yearly license.

Providers

Mick.Powell@trustmarque.com

- Sold Spectra T950 to us.
- Spectra & Quantum

Xendata

Phil .
https://xendata.com/

Limit x volume : 2 billion files

- **Present a disk interface**, with a managed disk size, up to 20PB.
- Disk is saved to disk, then it goes into a queue and is saved to tape.
- LTFS or tar. Tar as file marker.
- Runs on windows server .

Workflow API (data mover)
- Xml based API to a network socket on xendata machine.
- CIFS,ftp,nfs ,

Supported backends:
- Tape
- S3

Cons

- They do not support the IBM TS4500
- Not a native object store

Status

Not useful since it requires mounting the storage as NFS.

They are looking for ways of providing an S3 frontend.
Phil : <<<I believe that our architecture will provide better scaling than that of BlackPearl. A single XenData appliance will match the performance of the Spectra's S-model. And, we then scale by adding additional appliances.>>>

Friday, 19 June 2020

They added an S3 frontend for their solution.

- Ready by the end of summer.
- They have seen it helps position themselves in the market.
- With an SSD cache, they can feed 3 drives with one server.
- Can go to 30PB/year 1000MB/s with one server, can be HA Active-passive.
- Also second phase 60PB/year.

Quantum StoreNext

https://www.quantum.com/StorNext
https://qsupport.quantum.com/kb/flare/Content/stornext/SN6_DocSite/Default.htm

Status

- Waiting for first contact

Pros

Cons

Atempo

Miria - Backup & Migration for Large File Storages

Status

- Price given.

Pros

- Easy to use by other teams as an archival mechanism.

Cons

- It's not a native object store
- Does not have an API, it has a command line.
- Integration would be difficult.

PoC => link

- S3 as source is under development.
-

Price

30PB => 150K£
50PB => 250K£
This is not the final price. ! They can provide Site license, or company license. 3 or 5 years.

Hello Marc,

Indicated price was for 30PB not TB (sorry for this).
For a 50PB licence it will cost circa 250k£.
Over this volume we do site licence.
Kind Regards,

Christophe

Hardware requirements

Hello Marc,

We usually calculate 1 MIRIA Data Mover for 2 LTO8 Drives.
The LTO8 Drive (uncompressed) are working at 25TB/Day or 750TB/month
16 drives means 12PB/month , so we will need 16 LTO8 drives and 8 data movers.
Data movers specs are as followed:

Miria Server (1 or 2 of redundancy)

- Windows 2016 / RedHat (CentOS)
- Dual-CPU / 16 cores

    - 128 GB RAM
    - Two HDD 300 GB 10K SAS drives (Operating System)
    - Catalog (MaxDB) on separate disk: 8 TB minimum
    - Dual 10 GbE cards (for Data-Mover communication)

Miria Data Mover (8) (1 data mover for each 2~3 drives)

    - Dual-proc 16 / 18 cores (minimum)
    - 32 GB RAM
    - 10 GbE / 40 Gbe / 100 Gbe / Infiniband / Fiber

Might you want to, we can set another call with Thomas tomorrow or Friday,
Kindest Regards,

No worries, maybe it was not detailed enough.

Our offer is software only, the customer needs to provision the hardware.
Indicated prices are for perpetual license based on volume with all options.
SKD and integration support with third party applications is included as well.
Installation is not included but typically should be 3 days of Professional Services (5k£).
Training is not included (to be discussed).
Support and Maintenance is not included. We have 2 options:

- Regular working hours and days: 15% of the licence cost per year,
- 24/24, 7/7 : 25% of the licence cost per year.

A site licence can be envisioned with a duration period (5 years for instance) including or not Support and Maintenance (S&M can also remain annual).
Hope this help, but please don't hesitate if you need more information at this stage.

Kindest Regards,

Contact

From: Marc Riera <mriera@ebi.ac.uk>
Sent: Wednesday, January 29, 2020 11:55:06 AM
To: Christophe Darras <christophe.darras@atempo.com>
Subject: Re: Atempo contact following your request on our website

Friday, 19 June 2020 Command line interface , web API

They will look at minio , and how to integrate it with their software.
They do not provide hardware. :/

Object matrix

https://object-matrix.com/
Uk based, Cardiff
Media industry, and investment banks.
Used to replace legacy archives, and lto libraries.
Partner with Aspera, spectra, and any S3 AWS .

Status

- First meeting . Good.

Pros

- Support in Cardiff
- Software company, sitting on standard hardware. Non proprietary.
- Create vaults per projects, and policies for each vault.
- Integrates with public cloud, good to scale.
- No database, metadata is stored with the object.
- Usage metrics.
- S3 , smb, nfs, ftp, ….
- Move 2 S3 => HSM mechanism.
- Worm mechanism and rules.
- DR/BC
- Not proprietary, no lock-in
- They can Scan third parties and then HSM
    - Orchestration layer can move to ‘US’ , and then lifecycle to backup.
    - Can scan HGST S3

Cons

- Internal connectivity between nodes.
- To move to tape works
    - Xendata or blackpearl is needed.

Deploy time

- Usually , lead time, 2~4 weeks
- ISO , they need hardware and ILO access.

Pricing

- Capacity pricing
    - Appliance, standard hardware
    - Software only.

- 21PB ~ with everything ~> 1.35M (24PB raw , 21PB usable)
    - 3 racks space (97U excluding hubs)
    - Not geo-disperse

Status

- They do the exercise of costing if they can compete with google storage.

Information

Hi Marc

Ok great
There are lots of blog posts and datasheets on the website, link below
https://object-matrix.com/blog/

We have traditionally been deployed as a nearline/tier 2 platform on top of an LTO library, managing the automated data movement with our hsm move2, details below
https://object-matrix.com/products/access-tools/hierarchical-storage-management/

I have also added a few interesting links to have a browse through

2020 Storage Predictions
If You Could Change Just One Thing
Customer Pain Points
$2 an Hour for 1PB of Hybrid Cloud Storage
What is the point of LTO
5 Things Driving the Migration from LTO to Object Storage

Cheers
Pete

Peter Watling
Object Matrix - EMEA
Mobile: +44 7713 141525
Skype: peterwatling1980
Web: www.object-matrix.com

Technical guy: Mark Habberfield .

They support the IBM TS4500

With partner :

a) Integrate IBM directly.
b) Partition the spectra and adding the JD drives. NOPE !
c) Black Pearl can talk to IBM , they will do it. Not advertised but they can do it.

- They will look at integrating some tape management directly.

GPFS/Spectrum Scale

https://www.ibm.com/support/knowledgecenter/STXKQY_5.0.4/com.ibm.spectrum.scale.v5r04.doc/bl1ins_objectoverview.htm

Status

- Waiting for first contact, mail sent.

Pros

Cons

Commvault

Status

- Mail sent to global site, will see.

Versity

Status

-

Pros

- Both libraries integrated
- High scale at its core
- https://news.ycombinator.com/item?id=18022806
- Integrate with Starfish (recommended by them)
- Small file friendly, on the disk based

Nice sales presentation and list of features. Link
Presentation from Santa clara conference : link

Cons

- Read only if we stop paying license.
- Pricing, not expensive but at this point is obscure to me

OCF company

They are looking for solutions to our problem.
David Yip
He is talking with Atempo, aiming to provide the hardware for them.

Scality - Data orchestration software

They support Point
They support strongLink <= work together
Support Fujifilm
Software based, RedHat based, rpm release.
CDMI ~ first API implementation with S3 .

- RING => Pete tested, EBI purchased 2017
    - To replace isilon
    - Single namespace , NFS and S3(AWS compatible)
        - Microsoft paid them to support BloB API (azure)
    - Scale out FS
        - Replication, Erasure and multi-site
        - Stateless connectors, servers
    - Bloomberg 6PB ring, 20 connectors , 20GB/s
    - Other solutions , multi-site , 6GB/s ~> 14PB/month
    - It is also a search application
- Zenko => HGST competitor, they bid with HP hardware
    - Open Source
    - Is a data workflow function, a data controller
    - Can use as backend, Ring, S3, BloB,
    - Does not talk to tape directly.
        - Fujifilm, S3 gateway to tape, Using Zenko open source
    - They store in native format, so AWS in S3, NFS is file. Open availability.
    - It can scan NFS , and apply policies.

Price

Zenko as part of Ring is free.
Total data management cost, price per PB,

- Appliance model
- Subscription model
- Perpetual <= !suggested 50PB, List price is 10Million$, 5 years , 24/7

They are a software company, so they can be flexible. If we buy 5PB RING they can give 50PB of zenko.
They can talk with Fujifilm, they can work together.

Fujifilm

Read all this
https://blocksandfiles.com/2020/06/05/fujifilm-software-framework-object-storage-on-tape/
Uses Scality Zenko

Others that did not respond positively

StorageDNA -> They say the cloud is the future, they can provide cloud.

Other names to check ?

https://www.osnexus.com/

They have a commercially supported version of CEPH that comes with all the enterprise support you'd expect and allows you to build a very very cost effective S3 cloud on any JBOD you prefer...Qanta (QCT), WD, Supermicro, etc etc.

Other relevant business or software

Starfish

https://www-356.ibm.com/partnerworld/gsd/solutiondetails.do?solution=55823&lc=en&stateCd=P&tab=2

Business problem

Starfish solves the problem of massively large filesystems where it takes an enormous amount of time to maintain the storage performance, cost and growth as objects or files are added at a very fast pace. Having instant disk utilization, enabling tag and key-value pairing, quickly locating files and groups of files and being capable of keeping up with change rates on billions of files and petabytes of data is made more simple.

Business opportunity

Starfish creates and maintains a unified database of file metadata, file history, and customizable data such as ad hoc or metadata extracted from the files. The file metadata can be gathered using file system event integration and high performance scanning. The Starfish database can be queried and updated using our sophisticated CLI, HTML5 GUI, and API interfaces.
With this innovative approach to visualizing files across many different filesystem types and locations it provides the end user a unique way of reporting and knowing exactly where their growth is, in each file system. It also provides reporting on who is consuming the most amount of storage capacity, which files have and have not been accessed in customizable periods of time and finally a way for the owner of the files to either delete duplicates or files they no longer need. Move files to a lower cost storage or even archive them to S3 object stores locally or in the cloud. Using the IBM COS as a target provides the end user a way to use IBM an integrated lower cost target defined in Starfish and provide policies to archive and restore their data to the IBM COS.
Starfish Storage is a software solution that ultimately provides the end user with a total wholistic end to end life cycle management of their data even with billions of objects and petabyte scale file collections. By using Starfish the user can have clear visibility into file systems, directories, and files. Starfish orchestrates the data management lifecycle from detection to data migration among file systems and object stores.

Return on investment

Starfish provides a return on investment by being able to apply a cost per GB to a volume that Starfish is managing. This can then provide a view to the directories, filesystems and files and group them for a total cost. Having the metadata around these objects Starfish can then manage archive targets and lower cost volumes that the end user can introduce to the configuration. The user can use policy driven actions to move or migrate the data that is not frequently utilized to lower cost volumes or to archive them to the archive targets, which will immediately show an ROI savings and provide the ability to make the primary expensive storage last longer without the need for upgrades.

Competitive advantage

Starfish is based on a high performance scanner. Core to its speed is the use of multiple concurrent crawlers for each job, and the use of differential and incremental scans after the initial full scan. Concurrent scanning of multiple volumes using multiple agents can also be used to scale scan performance. The job manager allows Starfish to perform policy based workflows to provide enhanced functionality. By leveraging previously scanned filesystems, the job manager can run processes that cannot be realistically run using existing tools. Starfish includes high level, mid-level and detailed reporting. High level reporting leverages includes reports such as, Volume details: Number of files, size of volume, maximum file size, maximum directory size, and similar high level volume information, user and group reports and file size charts.