# IBM Spectrum Scale Best Practices for Ultra Fast Data Acquisition

New Concepts in Ultra Fast Data Acquisition Workshop
PSI – April 10+11, 2018

Ulf Troppens, IBM Spectrum Scale Development

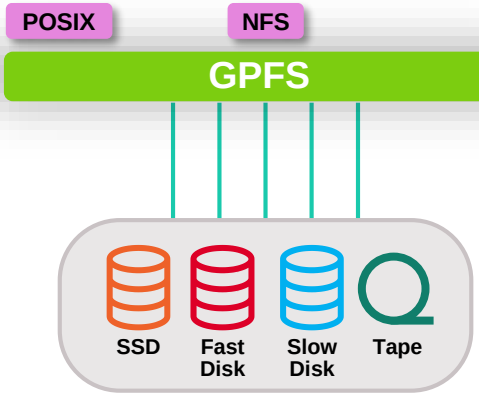
# Outline

1) ***IBM Spectrum Scale Overview***
2) Designing an IBM Spectrum Scale Implementation
3) Composable Infrastructure for Ultra Fast Data Acquisition, Data Analysis and Archiving
4) Best Practices & Discussion

# GPFS is changing …

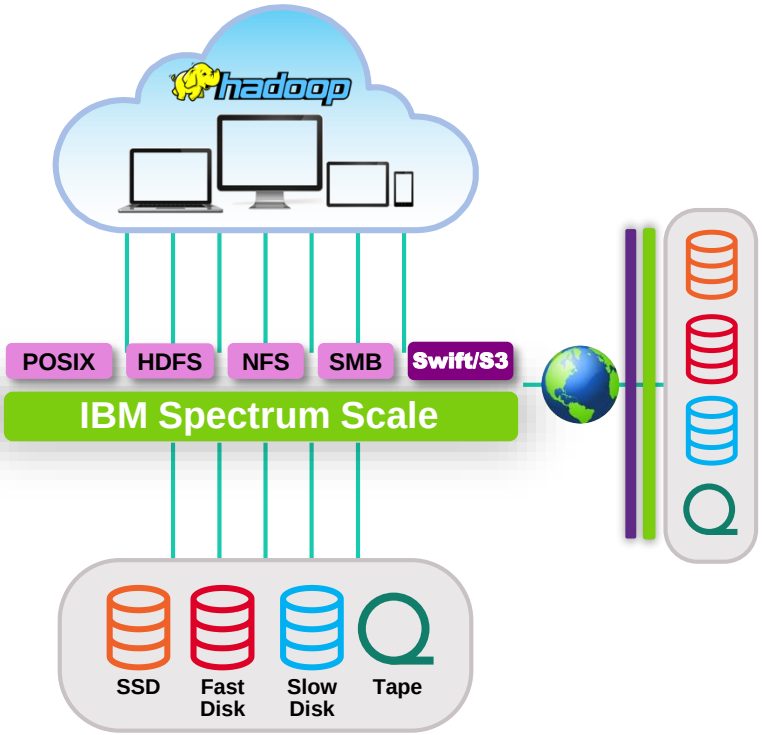
IBM Spectrum Scale

- 1993: Started as “Tiger Shark” research project at IBM Research Almaden as high performance filesystem for accessing and processing multimedia data
- Next 20 years: Grew up as General Parallel File System (GPFS) to power the world’s largest supercomputers
- Since 2014: Transforming to IBM Spectrum Scale to support new workloads which need to process huge amounts of unstructured data

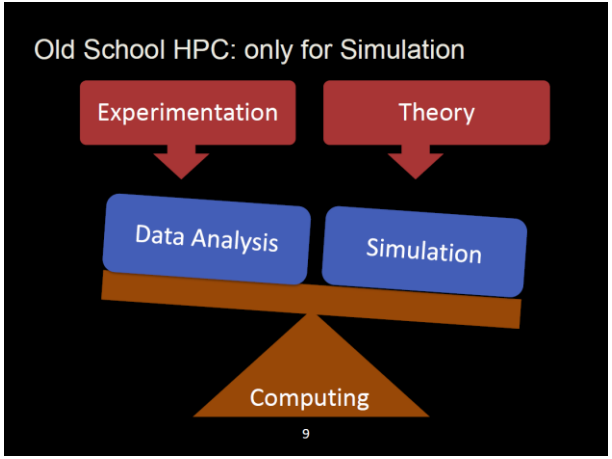


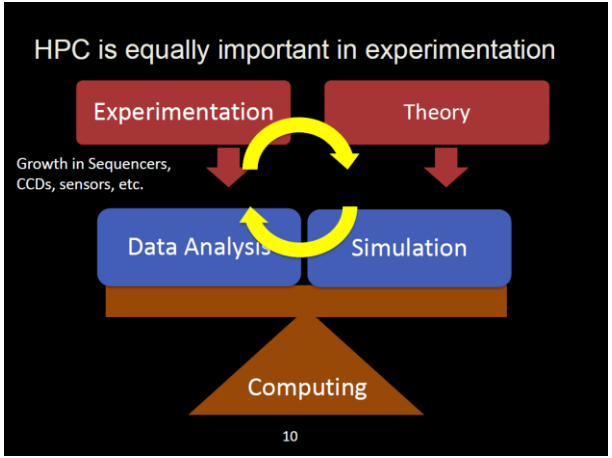
# IBM Spectrum Scale

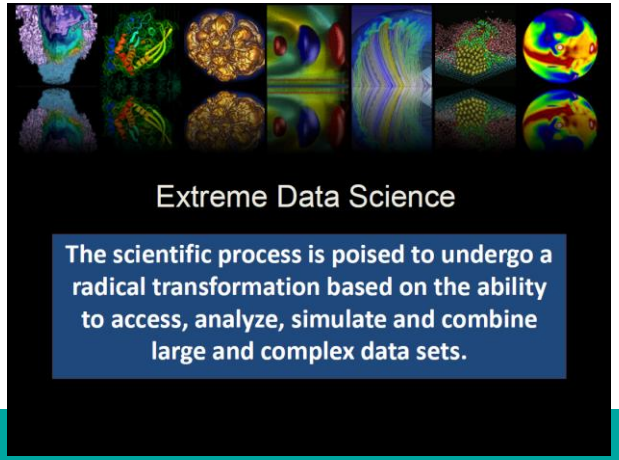
- Based on GPFS, a robust, fast and mature parallel file system
- BUT: If you still just think GPFS, you miss:
  - Support for workflows which for example inject data via object, analyze results via Hadoop/Spark and view results via POSIX
  - Storing and accessing large and small objects (S3 and Swift) with low latency
  - Automatic destaging of cold data to on premise or off premise object storage
  - Exchange of data between Spectrum Scale clusters via object storage in the cloud
  - Storing and starting OpenStack VMs without copying them from object storage to local file system
  - GUI, Grafana Bridge, REST API
  - iSCSI boot
  - And many, many more



# Changes in Science



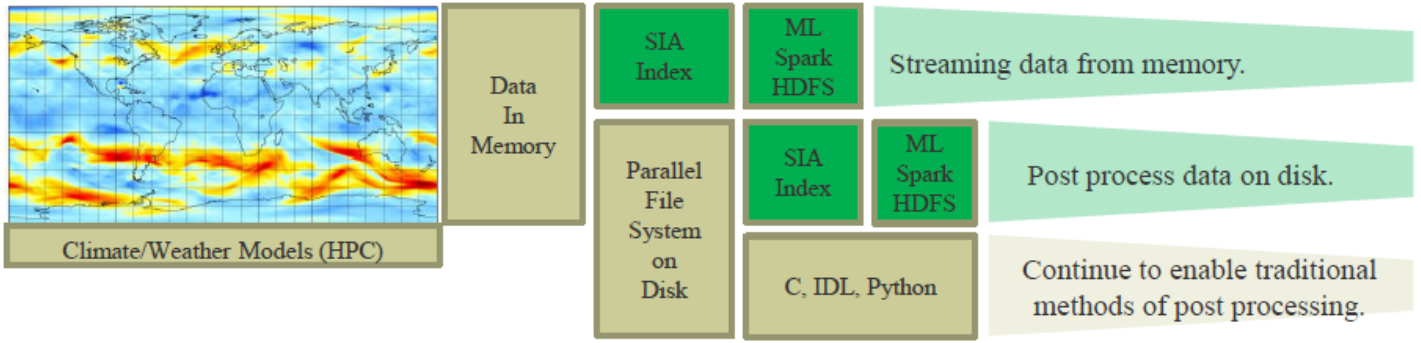




- Data driven science is getting the norm
- HPC gets integrated into the experiments to analyze huge amounts of measured data
- This shift is seen in scientific research and in industry (e.g. life-science, automotive)

https://science.energy.gov/~/media/ascr/ascac/pdf/meetings/201609/Yelick_Superfacility-ASCAC_2016.pdf

# Future of Data Analytics

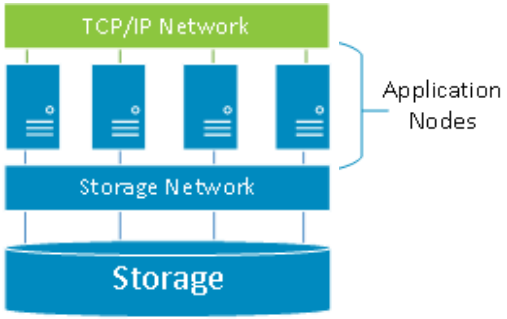



- Future HPC systems must be able to efficiently transform information into knowledge using both traditional analytics and emerging *machine learning* techniques.
- Requires the ability to be able to index data in memory and/or on disk and enable analytics to be performed on the data where it resides – even in memory
- All without having to modify the data

http://files.gpfsug.org/presentations/2016/SC16/06_-_Carrie_Spear_-_Spectrum_Sclale_and_HDFS.pdf

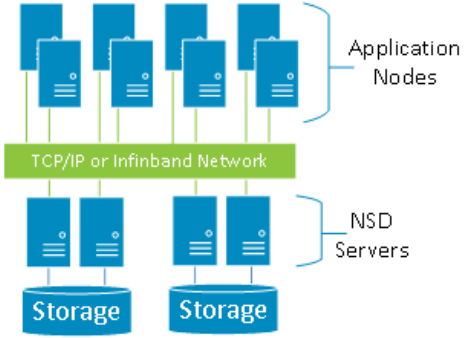
# Spectrum Scale deployment models

**Enterprise Integrated Model**



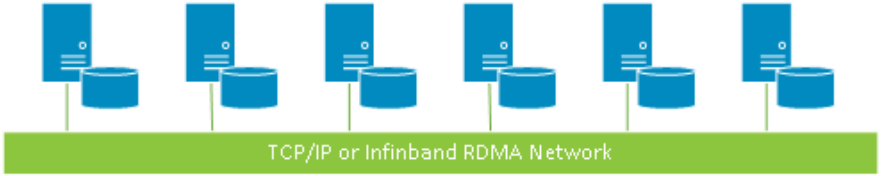

Unify and parallelize storage silos

**Network Shared Disk (NSD) Model**




Modular High-Performance Scaling

**Shared Nothing Cluster (SNC) Model**




Span storage rich servers for converged architecture or HDFS deployment

# Spectrum Scale Parallel Architecture

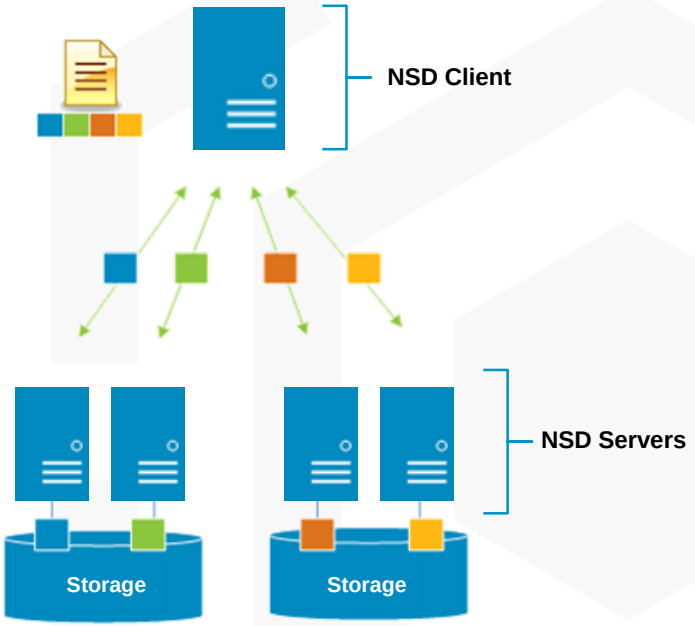
**No Hot Spots**

All NSD servers export to all clients in active-active mode

Spectrum Scale stripes files across NSD servers and NSDs in units of file-system block-size

File-system load spread evenly

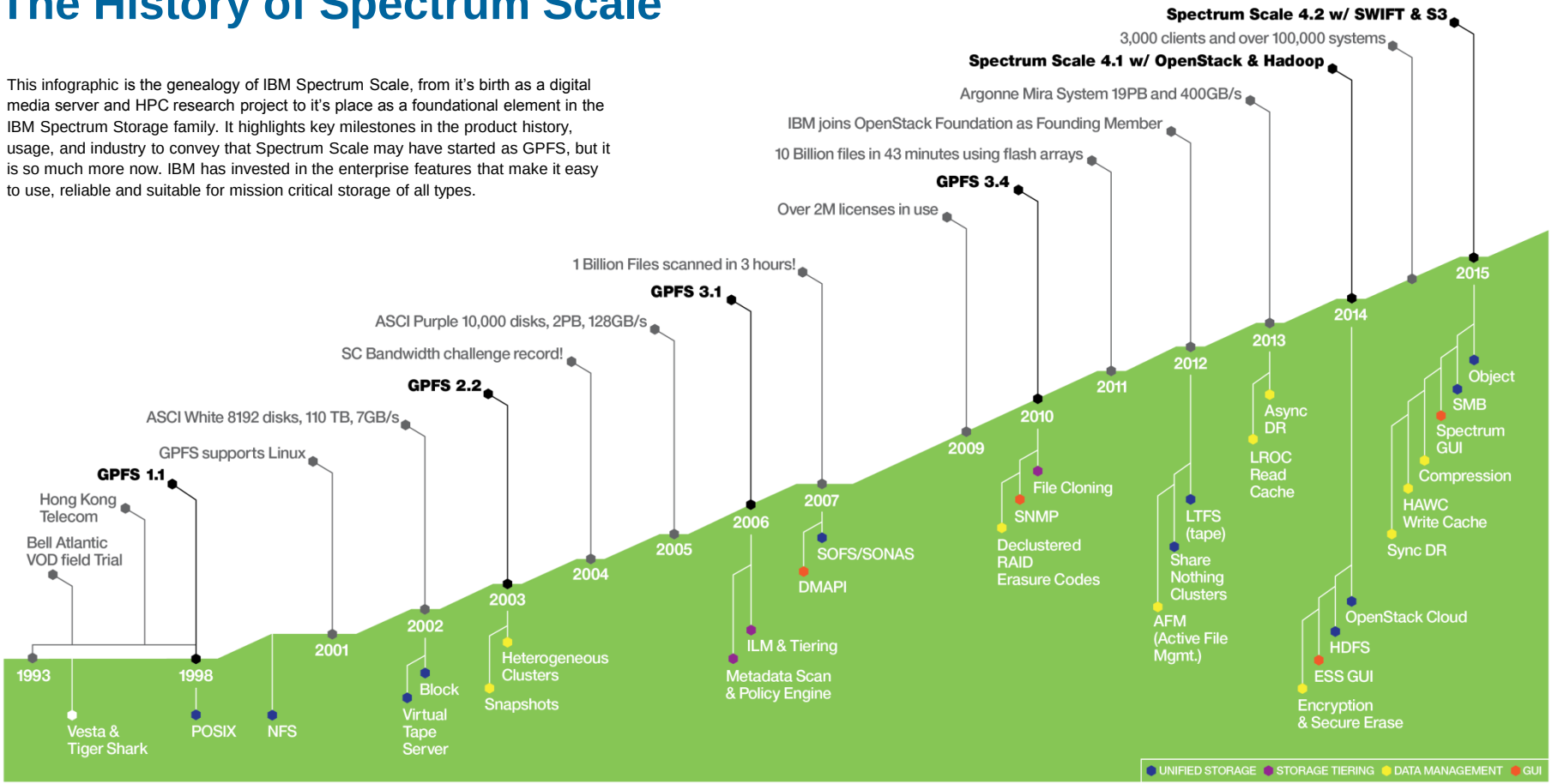
Easy to scale file-system capacity and performance while keeping the architecture balanced




NSD Client does real-time parallel I/O
to all the NSD servers and storage volumes/NSDs

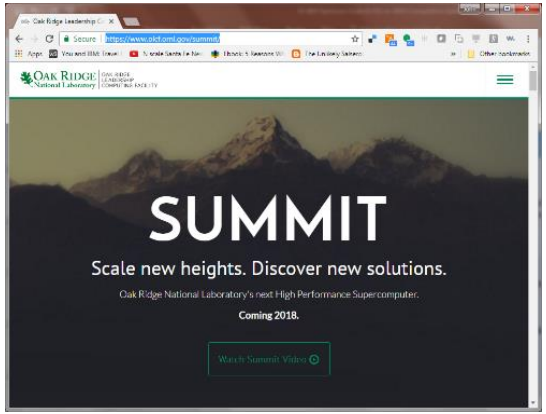
# The History of Spectrum Scale

This infographic is the genealogy of IBM Spectrum Scale, from it's birth as a digital media server and HPC research project to it's place as a foundational element in the IBM Spectrum Storage family. It highlights key milestones in the product history, usage, and industry to convey that Spectrum Scale may have started as GPFS, but it is so much more now. IBM has invested in the enterprise features that make it easy to use, reliable and suitable for mission critical storage of all types.



# Performance engineering matters

CORAL COLLABORATION
OAK RIDGE • ARGONNE • LIVERMORE




https://www.olcf.ornl.gov/summit/

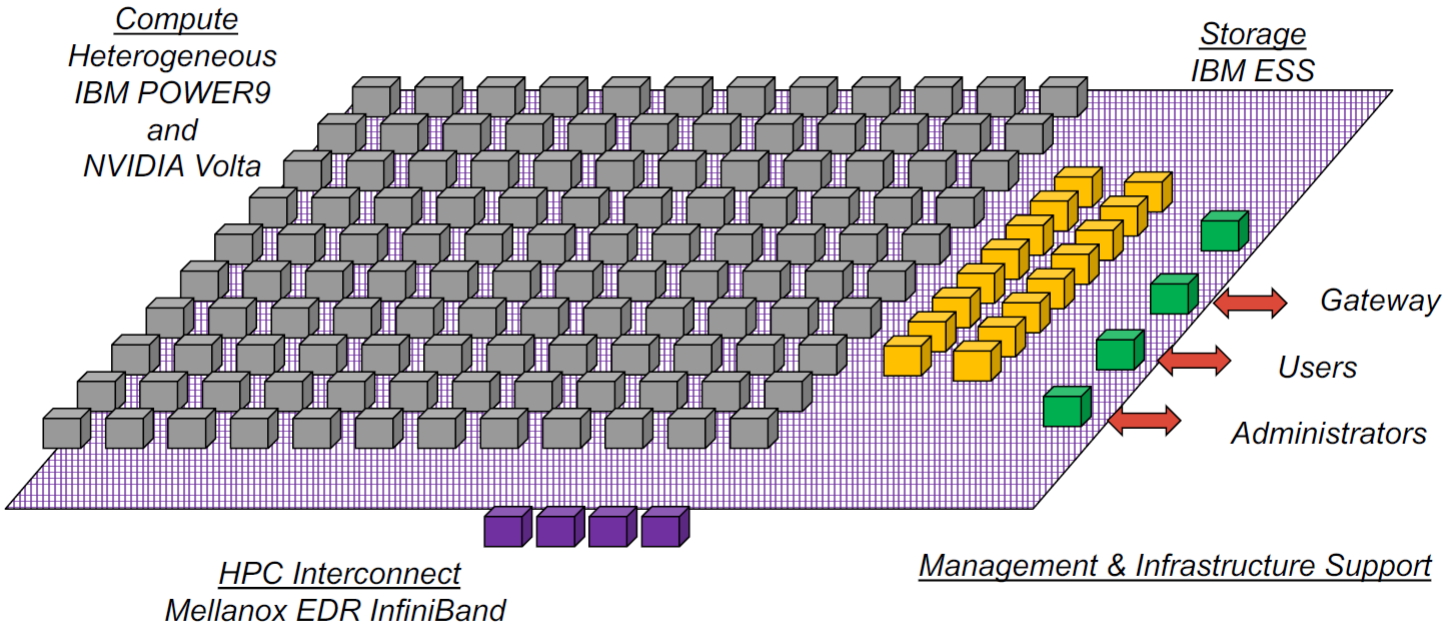
**Imagine you need to deliver the following goals:**

- 2.5 TB/sec single stream IOR as requested from ORNL
- 1 TB/sec 1MB sequential read/write as stated in CORAL RFP
- Single Node 16 GB/sec sequential read/write as requested from ORNL
- 50K creates/sec per shared directory as stated in CORAL RFP
- 2.6 Million 32K file creates/sec as requested from ORNL

**What innovations in storage would this require?**

# Performance engineering matters … scaling

## High Level Layout of the complete system



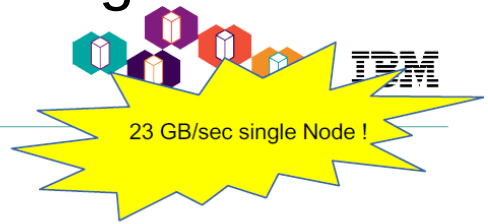





http://files.gpfsug.org/presentations/2017/SC17/SC17-UG-CORAL_V3.pdf

# Performance engineering matters … single ingest node

## Single client throughput enhancements




```
Began: Sat Nov 11 20:47:05 2017
Command line used: /perform/io-500-dev.ppc64le//bin/ior -w -C -Q 1 -g -G 27 -k -e -t 16m -b 128g -F -o /ibm/fs2-16m-10/shared/iorfile
Machine: Linux p8n19hyp

Test 0 started: Sat Nov 11 20:47:05 2017
Summary:
        api             = MPIIO (version=3, subversion=1)
        test filename   = /ibm/fs2-16m-10/shared/iorfile
        access          = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= constant task offsets = 1
        clients         = 8 (8 per node)
        repetitions     = 1
        xfersize        = 16 MiB
        blocksize       = 128 GiB
        aggregate filesize = 1024 GiB

access    bw(MiB/s)  block(KiB) xfer(KiB)  open(s)    wr/rd(s)   close(s)   total(s)   iter
------    ---------  ---------- ---------  --------   --------   --------   --------   ----
write     22261      134217728  16384      0.021629   47.08      0.001264   47.10      0

Max Write: 22261.01 MiB/sec (23342.36 MB/sec)

Summary of all tests:
Operation   Max(MiB)   Min(MiB)  Mean(MiB)     StdDev   Mean(s) Test# #Tasks tPN reps fPP reord reordoff reordrand seed segcnt blksiz xsize aggsize API RefNum
write       22261.01   22261.01  22261.01       0.00   47.10371 0 8 8 1 1 1 1 0 0 1 137438953472 16777216 1099511627776 MPIIO 0

Finished: Sat Nov 11 20:47:52 2017
```





http://files.gpfsug.org/presentations/2017/SC17/SC17-UG-CORAL_V3.pdf

# Performance engineering matters … single ESS

## IOR with GS4c – ESS with CORAL Enclosure (reduced output) – 16M


```
Began: Fri Oct 27 01:34:10 2017
Command line used: /tmp/ior-binary-dir/ior -F -i 3 -d 180 -w -r -e -t 16m -b 4064g -o /ibm/fs2-16m-10/ior-test-dir-1/iorfile -L
Machine: Linux fire01.sonasad.almaden.ibm.com

Test 0 started: Fri Oct 27 01:34:10 2017
Summary:
        api                = POSIX
        test filename      = /ibm/fs2-16m-10/ior-test-dir-1/iorfile
        access             = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= no tasks offsets
        clients            = 12 (1 per node)
        repetitions        = 3
        xfersize           = 16 MiB
        blocksize          = 4064 GiB
        aggregate filesize = 48768 GiB

Max Write: 34507.52 MiB/sec (36183.76 MB/sec)
Max Read:  41420.56 MiB/sec (43432.61 MB/sec)

Summary of all tests:
Operation   Max(MiB)   Min(MiB)  Mean(MiB)     StdDev    Mean(s) Test# #Tasks tPN reps fPP reord reordoff reordrand seed segcnt blksiz xsize aggsize API
RefNum
write       34507.52   34321.65   34418.00      76.04 1450.94658 0 12 1 3 1 0 1 0 0 1 4363686772736 16777216 52364241272832 POSIX 0
read        41420.56   41210.13   41340.40      92.93 1207.98744 0 12 1 3 1 0 1 0 0 1 4363686772736 16777216 52364241272832 POSIX 0

Finished: Fri Oct 27 04:05:07 2017
```





http://files.gpfsug.org/presentations/2017/SC17/SC17-UG-CORAL_V3.pdf

# IBM Elastic Storage Server (ESS)
## Integrated scale out data management for file and object data

**Optimal building block for high-performance, scalable, reliable enterprise storage**
- Faster data access with choice to scale-up or out
- Easy to deploy clusters with unified system GUI
- Simplified storage administration with IBM Spectrum Control integration

**One solution for all your data needs**
- Single repository of data with unified file and object support
- Anywhere access with multi-protocol support:
  NFS 4.0, SMB, OpenStack Swift, Cinder, and Manila
- Ideal for Big Data Analytics with full Hadoop transparency with 4.2

**Ready for business critical data**
- Disaster recovery with synchronous or asynchronous replication
- Ensure reliability and fast rebuild times using Spectrum Scale RAID's dispersed data and erasure code

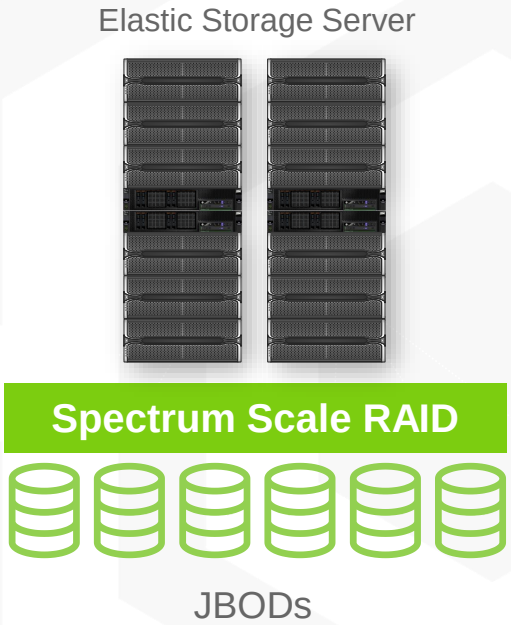
# Advantages of Spectrum Scale RAID

**Use of standard and inexpensive disk drives**
- Erasure Code software implemented in Spectrum Scale

**Faster rebuild times**
- More disks are involved during rebuild
- Approx. 3.5 times faster than RAID-5

**Minimal impact of rebuild on system performance**
- Rebuild is done by many disks
- Rebuilds can be deferred with sufficient protection

**Better fault tolerance**
- End to end checksum
- Much higher mean-time-to-data-loss (MTTDL)
  - 8+2P: ~ 200 Years
  - 8+3P: ~ 200 Million Years



# A new level of storage performance and efficiency

**Dramatic improvements in I/O performance**

**Support for newest low-latency, high bandwidth hardware such as NVMe**

- Significantly reduced communication latency between nodes

**Improved performance, space efficiency for mixed workloads**

- Small and large block size workloads running simultaneously in same file system
- Optimize large block performance via new 4MB default block
- Simultaneously optimize small file space efficiency with variable sub-block size

**Improved IOP/sec and metadata performance**
**IOP/s can improve 3x to 5x over previous releases***

*Performance numbers are estimates based on IBM internal lab tests and are subject to verification

# Spectrum Scale User Group

The Spectrum Scale User Group is free to join and open to all using, interested in or integrating Spectrum Scale.

Independent of IBM. IBM supports with speaker.

Join the User Group activities to meet your peers and get access to experts from partners and IBM.

http://www.spectrumscale.org/uk-may-2016-group-report/

**Major meetings 2018:**

| | | |
|---|---|---|
| **- German User Meeting** | **Feb 28/Mar 1, 2018** | **Ehningen** |
| **- AP User Group Meeting** | **Mar 26, 2018** | **Singapore** |
| **- UK User Group Meeting** | **Apr 18+19, 2018** | **London** |
| **- US User Group Meeting** | **May 16+17** | **Boston** |
| **- User Group Meeting @ ISC18** | **Jun 25, 2018** | **Frankfurt** |
| **- User Group Meeting @ SC18** | **Nov 11, 2018** | **Dallas** |

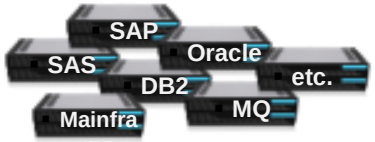
Web page: http://www.spectrumscale.org/
Presentations: http://www.spectrumscale.org/presentations/
Mailing list: http://www.spectrumscale.org/join/
Contact: http://www.spectrumscale.org/committee/

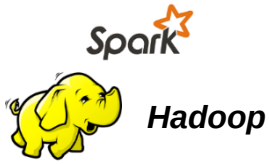
# Outline

1) IBM Spectrum Scale Overview
2) ***Designing an IBM Spectrum Scale Implementation***
3) Composable Infrastructure for Ultra Fast Data Acquisition, Data Analysis and Archiving
4) Best Practices & Discussion

# Unleash New Storage Economics on a Global Scale

Client workstations





Users and applications

## Single name space


POSIX
NFS
SMB/CIFS
HDFS Connector
Object
S3
Swift
OpenStack
Cinder
Glance
Manila
Docker
kubernetes

# IBM Spectrum Scale

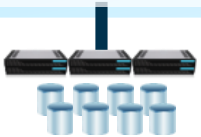
Automated data placement and data migration

*Active = Hot+Warm = "Online"*

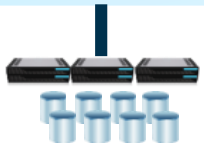
*Archive = Cold = "Offline"*

Flash

Disk



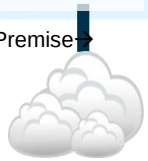
Storage Rich Servers



Object Storage

Tape

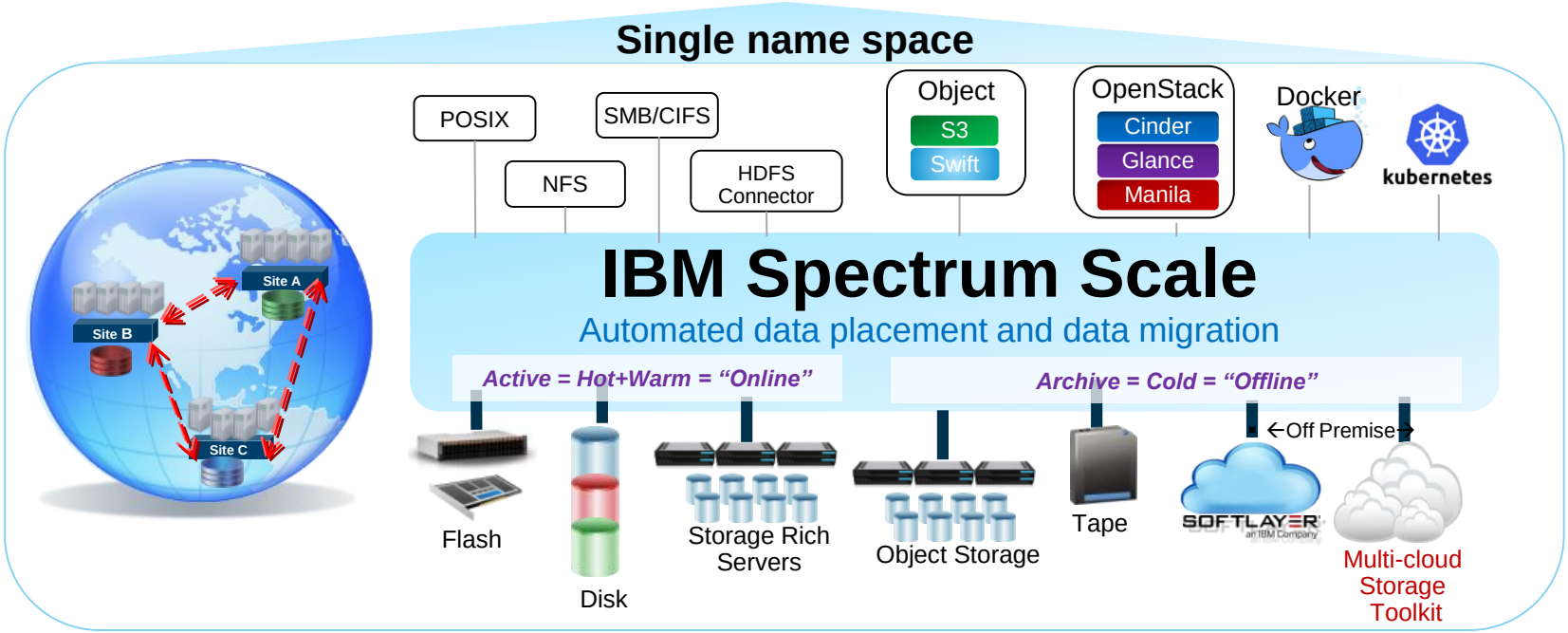
←Off Premise→

Multi-cloud Storage Toolkit

# 1) Understand Use Case

**Core HPC** | **File-based Workflows** | **Application Acceleration** | **Virtual Infrastructure**



# 2) Define the Spectrum Scale Configuration

**Core HPC** | **File-based Workflows** | **Application Acceleration** | **Virtual Infrastructure**

## IBM Spectrum Scale

**File Services / File Storage Virtualization / Application Integration**
**Software-defined Storage Freedom**


*Active = Hot+Warm = "Online"*

*Archive = Cold = "Offline"*

Flash

Disk

Storage Rich Servers

Object Storage

Tape

←Off Premise→

SOFTLAYER an IBM Company

Multi-cloud Storage Toolkit

# 3) Select Best of Breed Storage

**Core HPC** | **File-based Workflows** | **Application Acceleration** | **Virtual Infrastructure**

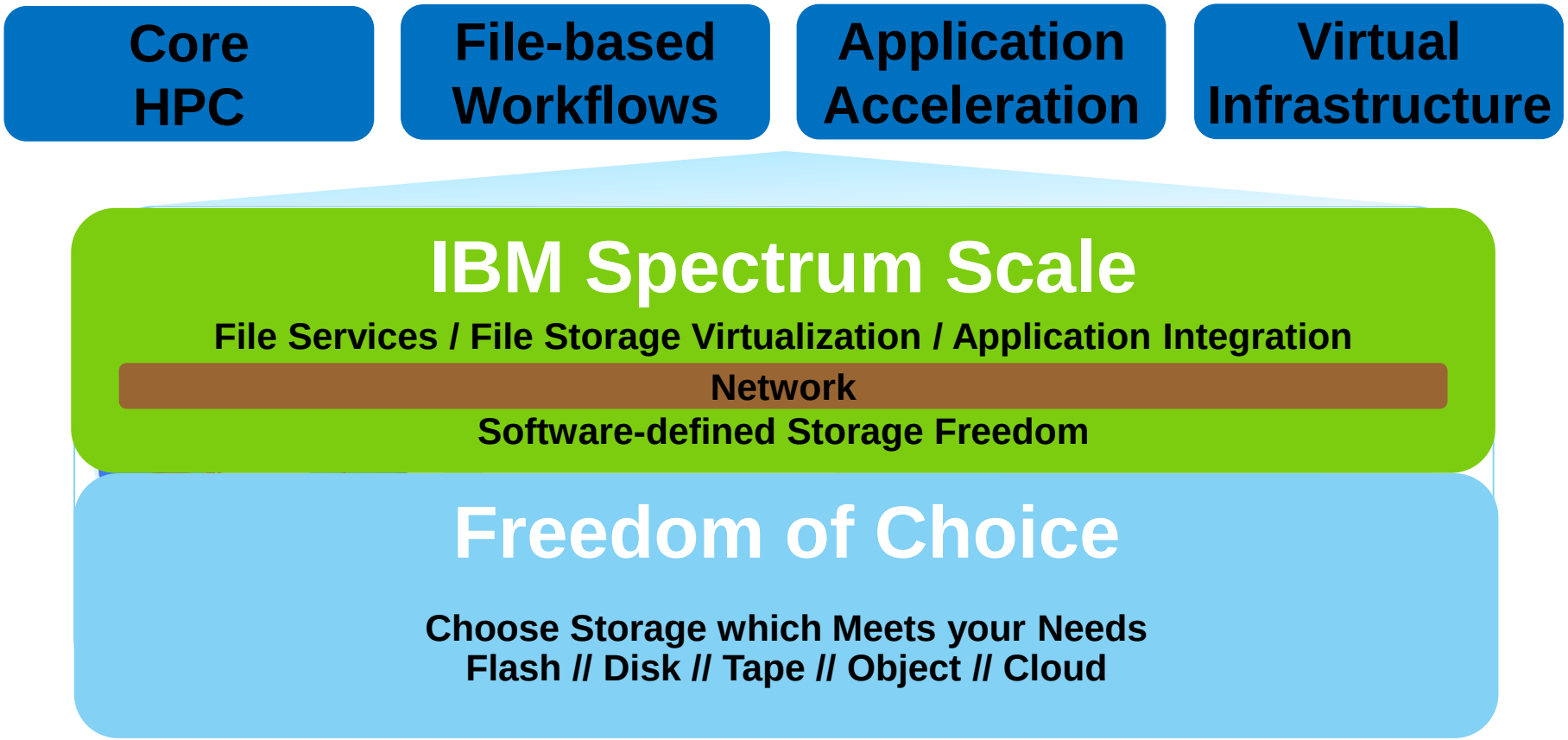
## IBM Spectrum Scale

**File Services / File Storage Virtualization / Application Integration**
**Software-defined Storage Freedom**

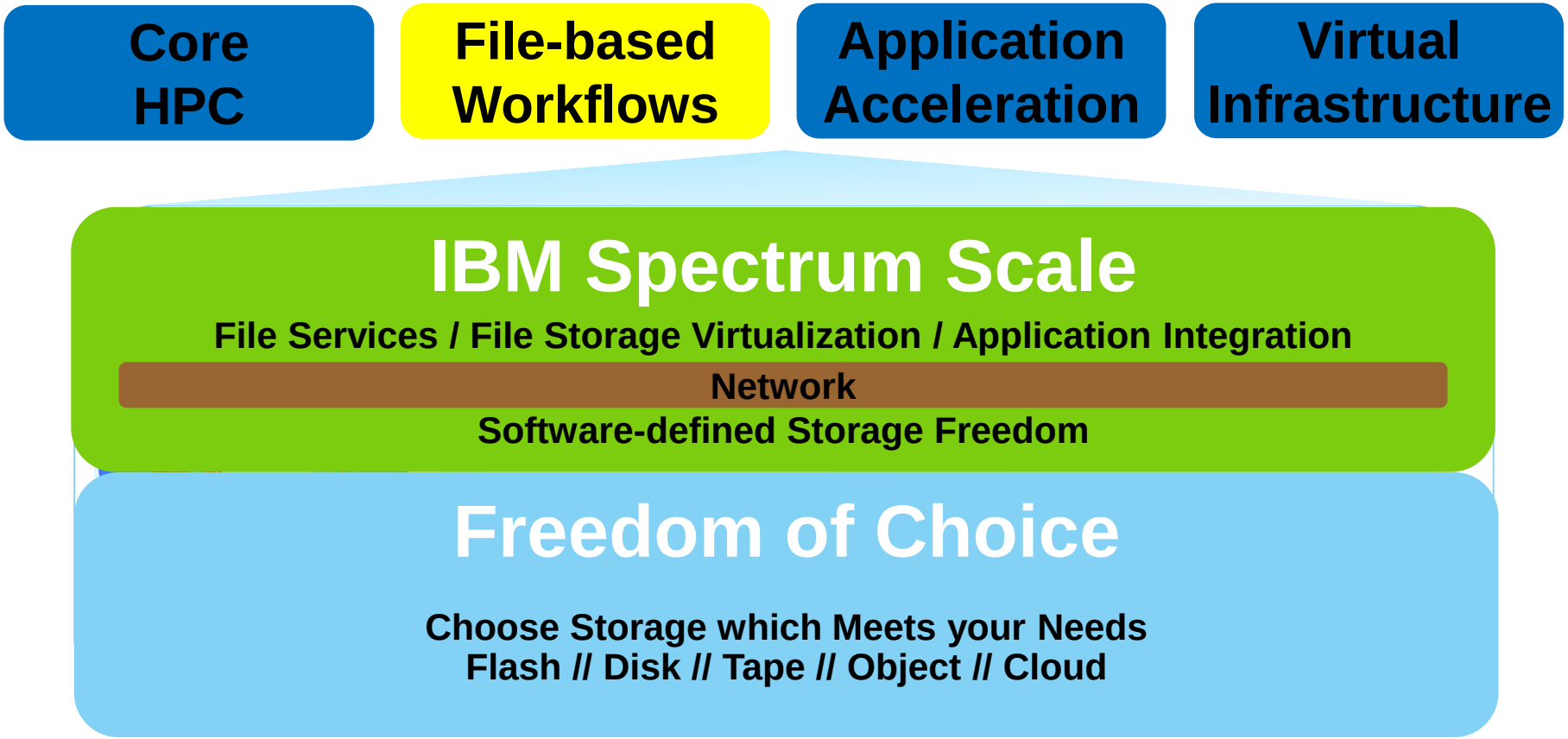
## Freedom of Choice

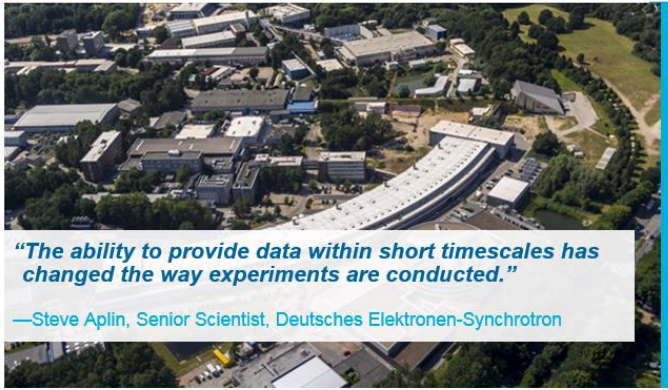
**Choose Storage which Meets your Needs**
**Flash // Disk // Tape // Object // Cloud**

# 4) Network is Integral Part of End-to-end Solution



# Examples – File-based Workflows





*"The ability to provide data within short timescales has changed the way experiments are conducted."*

—Steve Aplin, Senior Scientist, Deutsches Elektronen-Synchrotron

### Business challenge

Research center Deutsches Elektronen-Synchrotron (DESY) found that increasingly resource-intensive experiments was affecting storage system performance, limiting research. How could the organization handle over five gigabytes of data streaming into its computing center every second?

### Transformation

With a flexible, high-performance storage solution from IBM, DESY can meet growing demand cost-effectively. Scientists can now start analyzing the data in just a few minutes, instead of days, accelerating ground-breaking research.

## Business benefits:

**Ensures**
DESY can easily maintain a multi-PB library of research data to meet growing demand and remain an attractive research destination

**Rapid**
access to millions of data points accelerates research and helps lead to breakthroughs

**Increases**
administration efficiency with automated data management, improving DESY's service offering

# DESY

## Making the next breakthrough in scientific research possible with the latest in storage innovation

DESY, Deutsches Elektronen-Synchrotron, is a national research center in Germany that operates particle accelerators and photon science facilities used to investigate the structure of matter. DESY is housed in Hamburg and Zeuthen, Germany, and attracts over 3,000 scientists from over 40 countries annually.

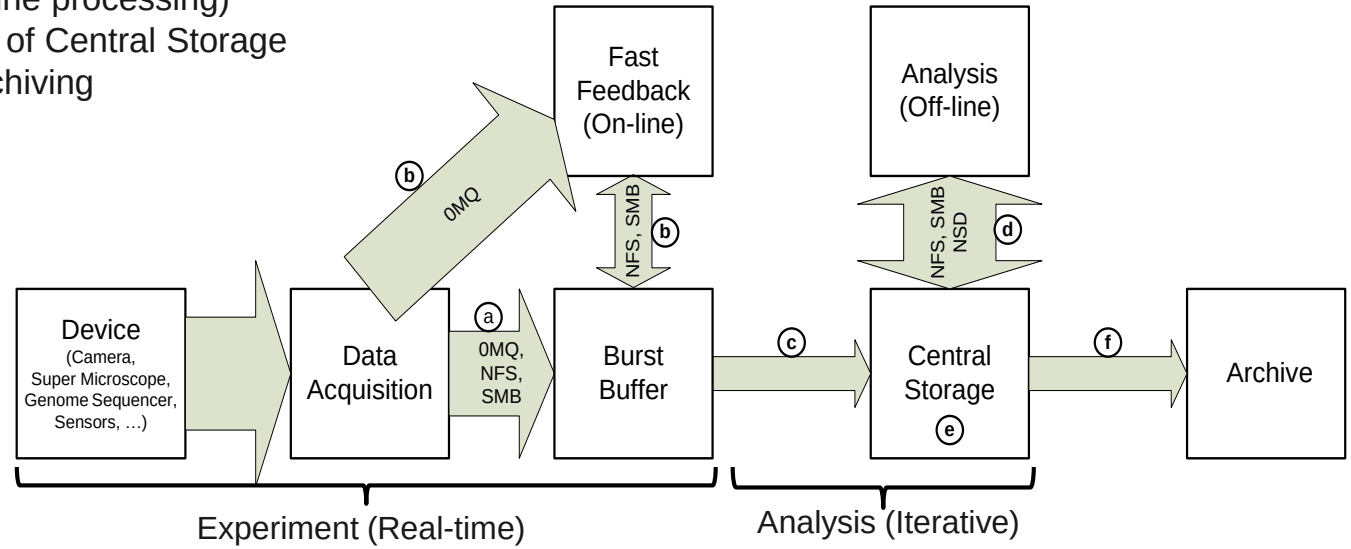
### Solution components

- IBM® Spectrum Scale™
- IBM Spectrum Scale RAID
- IBM Elastic Storage™ Server GS1
- IBM Elastic Storage Server GL4 and GL6
- IBM Power® S822L
- IBM Systems Lab Services

**Share this**

# Typical Workflow for Data Intensive Science

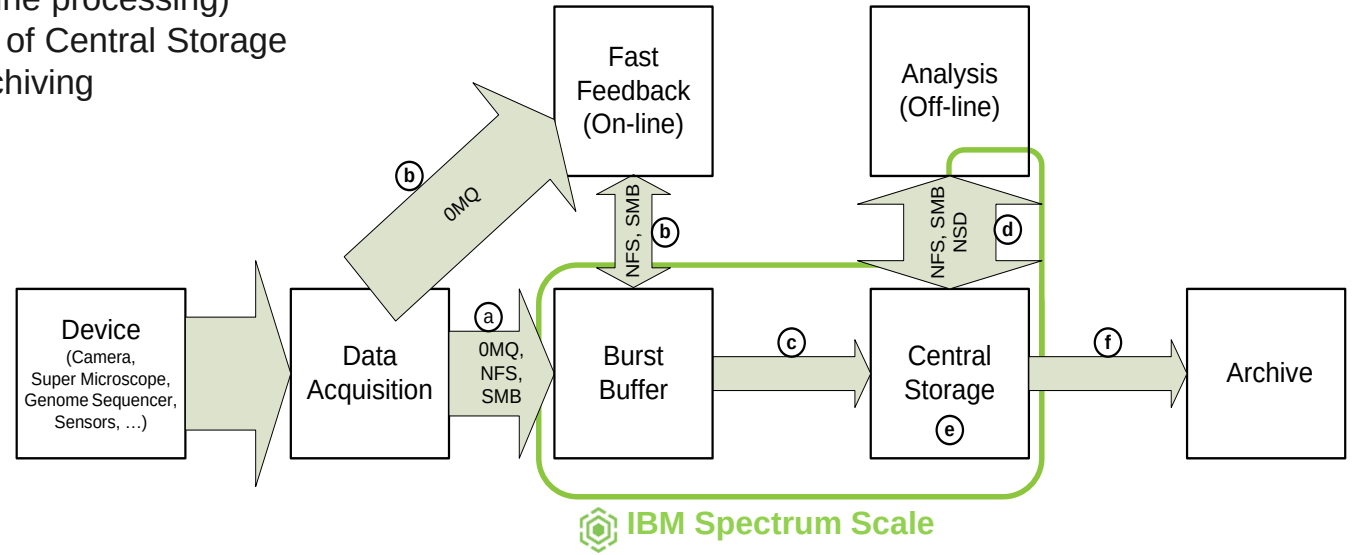
a) Real-time data ingest (data acquisition)
b) Visualization and near real-time analysis (online processing)
c) Data movement from Burst Buffer to Central Storage
d) Deep analysis (offline processing)
e) Data management of Central Storage
f) Long-term data archiving




- Scientists need access to data during each stage of the workflow

# Typical Workflow for Data Intensive Science (continued)

a) Real-time data ingest (data acquisition)
b) Visualization and near real-time analysis (online processing)
c) Data movement from Burst Buffer to Central Storage
d) Deep analysis (offline processing)
e) Data management of Central Storage
f) Long-term data archiving



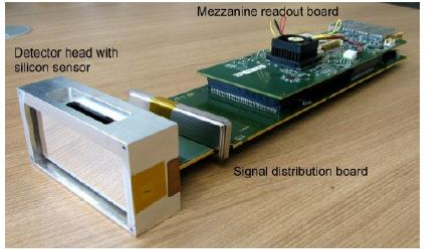

- Scientists need access to data during each stage of the workflow
- IBM Spectrum Scale has proven to support this workflow

# Current and Future Detector Rates

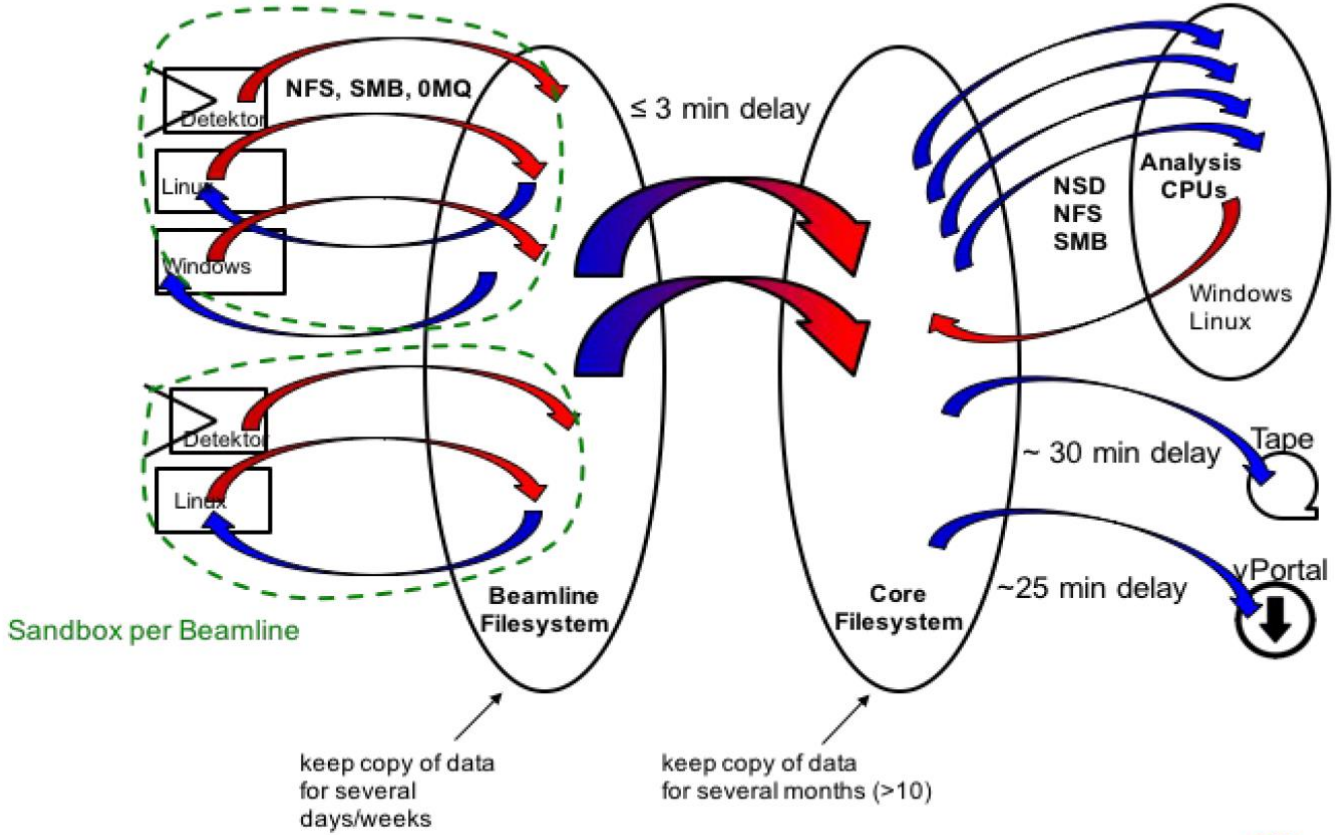
- Detectors exceeded capabilities of prev. system:
  - Pilatus 300k: 1,2 MB Files @ 200 Hz
  - Pilatus 6M: 25 MB files @ 25 Hz
    7 MB files @ 100 Hz
  - PCO Edge: 8 MB files @ 100Hz
  - PerkinElmer: 16 MB + 700 Byte files @ 15 Hz
  - Lambda: 60 Gb/s @ 2000 Hz (Future)
  - Eiger: 30 Gb/s @ 2000 Hz (Future)



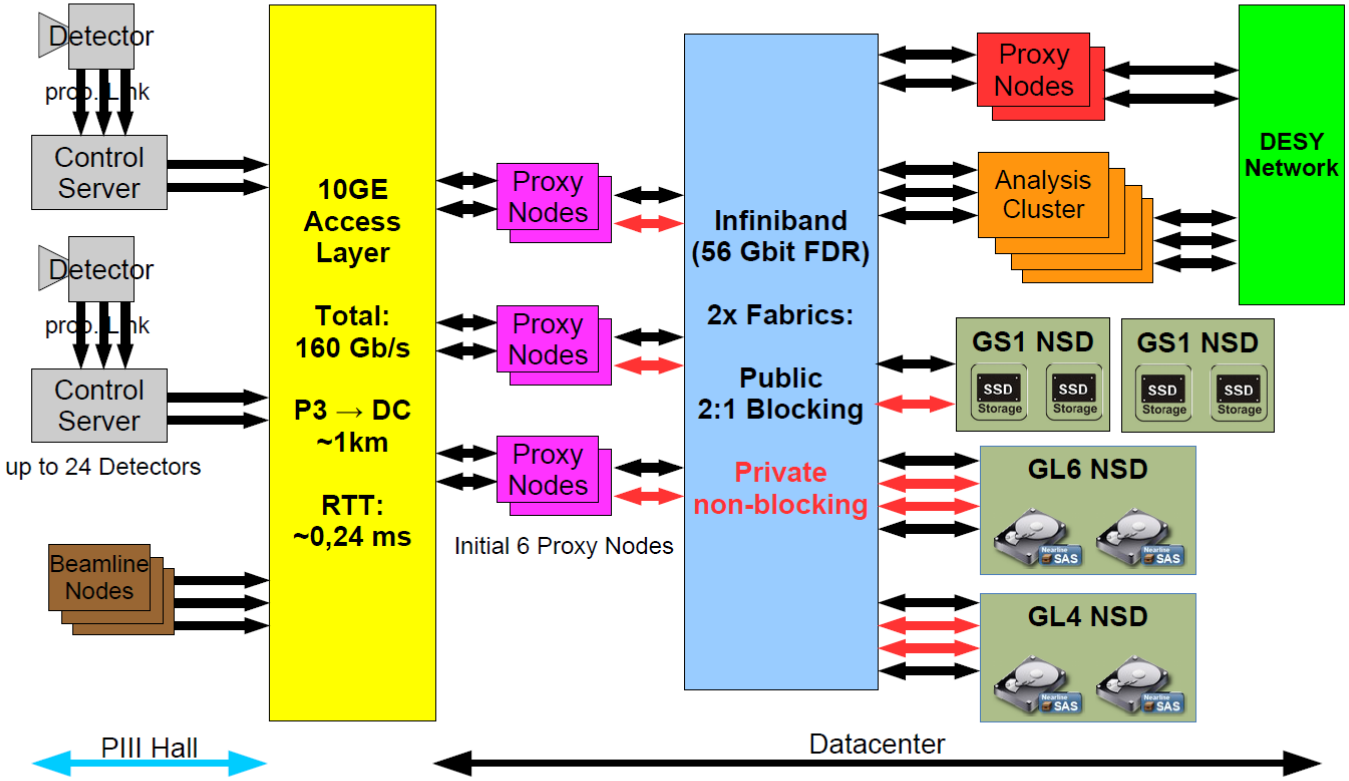
- GPFS is now used to handle those rates
  - SMB/NFS sufficient for current detectors
  - Future detectors need new methods



# from the cradle to the grave



# ASAP³ Architecture

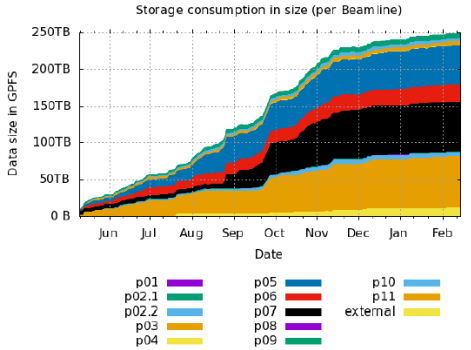


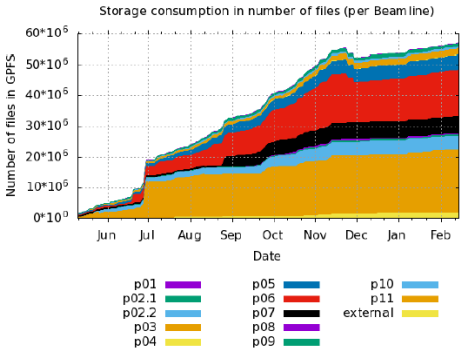
Source: https://www.spxxl.org/sites/default/files/GPFS_for_data_taking_&_analysis_at_next_generation_Light_Source_Experiments.pdf

# Customer Feedback

## Experiences from the first period

- Overall user experience: good!
  - BL scientist more time for experiment, sample preparation and user support
  - No beamtine loss due to lacking space
  - Reconstruction faster and more stable
  - Over reliance: “Runs with blind trust”
- Overall GPFS and ESS stability: good!
  - Good stability and performance
  - Initial MOFED issue on ppc64
  - Connect-IB FW issue
- First detectors of new generation being installed during current shutdown
  - 3x Lambda Modules
  - 1x Eiger 4M



# Resources

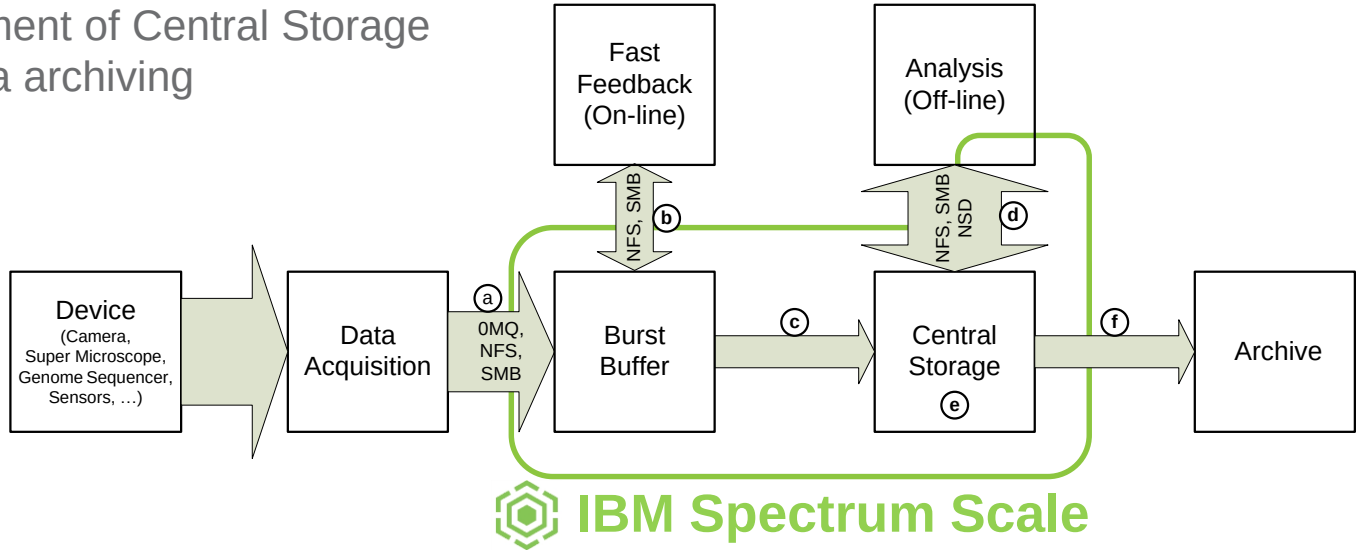
- Detailed whitepaper published by DESY at CHEP2015 http://iopscience.iop.org/article/10.1088/1742-6596/664/4/042053
- DESY presentation at IBM Edge 2015: http://www.slideshare.net/UlfTroppens/desy-ibm-edge2015-technical-computing-for-photon-science-20150520v2
- DESY presentation at ALICE, ATLAS, CMS & LHCb Second Joint Workshop on DAQ@LHC: https://indico.cern.ch/event/471309/contributions/1981091/attachments/1257042/1856128/gpfs-for-p3xfel.pdf
- IBM customer story with video and four page pdf: http://ibm.co/1qCIAuL

# Outline

1) IBM Spectrum Scale Overview
2) Designing an IBM Spectrum Scale Implementation
3) ***Composable Infrastructure for Ultra Fast Data Acquisition, Data Analysis and Archiving***
4) Best Practices & Discussion

# Key Use Cases

a) Real-time data ingest (data acquisition)
b) Visualization and near real-time analysis (online processing)
c) Data movement from Burst Buffer to Central Storage
d) Deep analysis (offline processing)
e) Data management of Central Storage
f) Long-term data archiving



Source: https://www.spxxl.org/sites/default/files/GPFS_for_data_taking_&_analysis_at_next_generation_Light_Source_Experiments.pdf

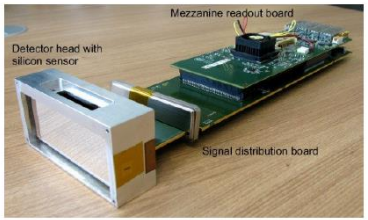
# a) Real-time data ingest

## Current and Future Detector Rates

- Detectors exceeded capabilities of prev. system:
  - Pilatus 300k: 1,2 MB Files @ 200 Hz
  - Pilatus 6M: 25 MB files @ 25 Hz
    7 MB files @ 100 Hz
  - PCO Edge: 8 MB files @ 100Hz
  - PerkinElmer: 16 MB + 700 Byte files @ 15 Hz
  - Lambda: 60 Gb/s @ 2000 Hz (Future)
  - Eiger: 30 Gb/s @ 2000 Hz (Future)



- GPFS is now used to handle those rates
  - SMB/NFS sufficient for current detectors
  - Future detectors need new methods






DESY

# a) Real-time data ingest (data acquisition)

- Characterization
  - Close to instrument (=outside data center)
  - Make measured data persistent
  - Storage must never cause the data taking of an experiment to fail
  - Side effects from other detectors, off-line analysis, etc. are not acceptable
  - Ingest via one or more 10GbE links

- Spectrum Scale capabilities
  - Tunable for high-speed data ingest
  - Parallel architecture allows to create a single file system for all detectors without having side effects
  - Built-in support for NFS and SMB
  - Allows to integrate customer configured ingest methods (e.g. ZeroMQ)
  - GNR technology minimizes impact of RAID rebuilds
  - Allows to integrate SSD and NL SAS in single file system
  - Optimizes costs using placement and migration policies

# Spectrum Scale Parallel Architecture

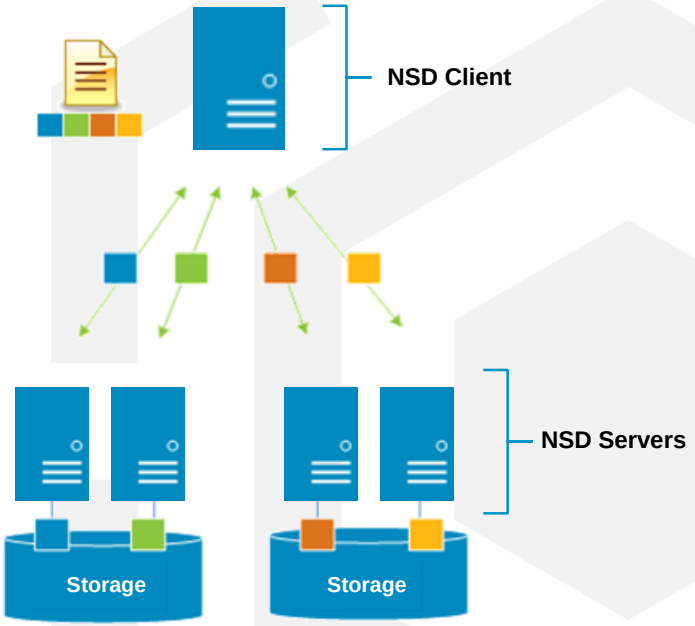
## No Hot Spots

All NSD servers export to all clients in active-active mode

Spectrum Scale stripes files across NSD servers and NSDs in units of file-system block-size

File-system load spread evenly

Easy to scale file-system capacity and performance while keeping the architecture balanced




NSD Client does real-time parallel I/O
to all the NSD servers and storage volumes/NSDs

# Advantages of Spectrum Scale RAID

**Use of standard and inexpensive disk drives**

- Erasure Code software implemented in Spectrum Scale

**Faster rebuild times**

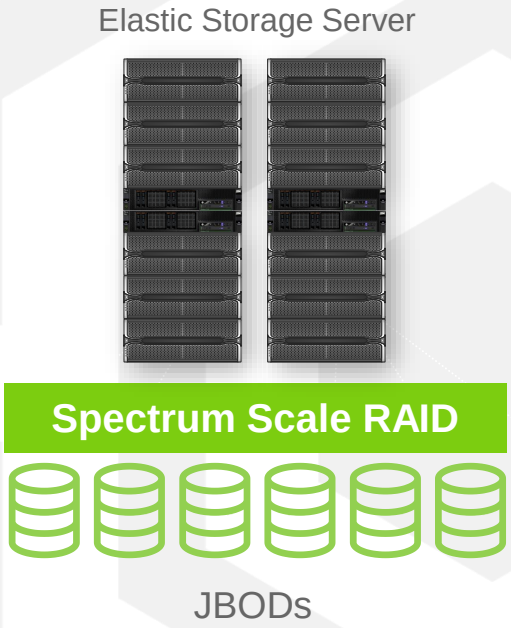
- More disks are involved during rebuild
- Approx. 3.5 times faster than RAID-5

**Minimal impact of rebuild on system performance**

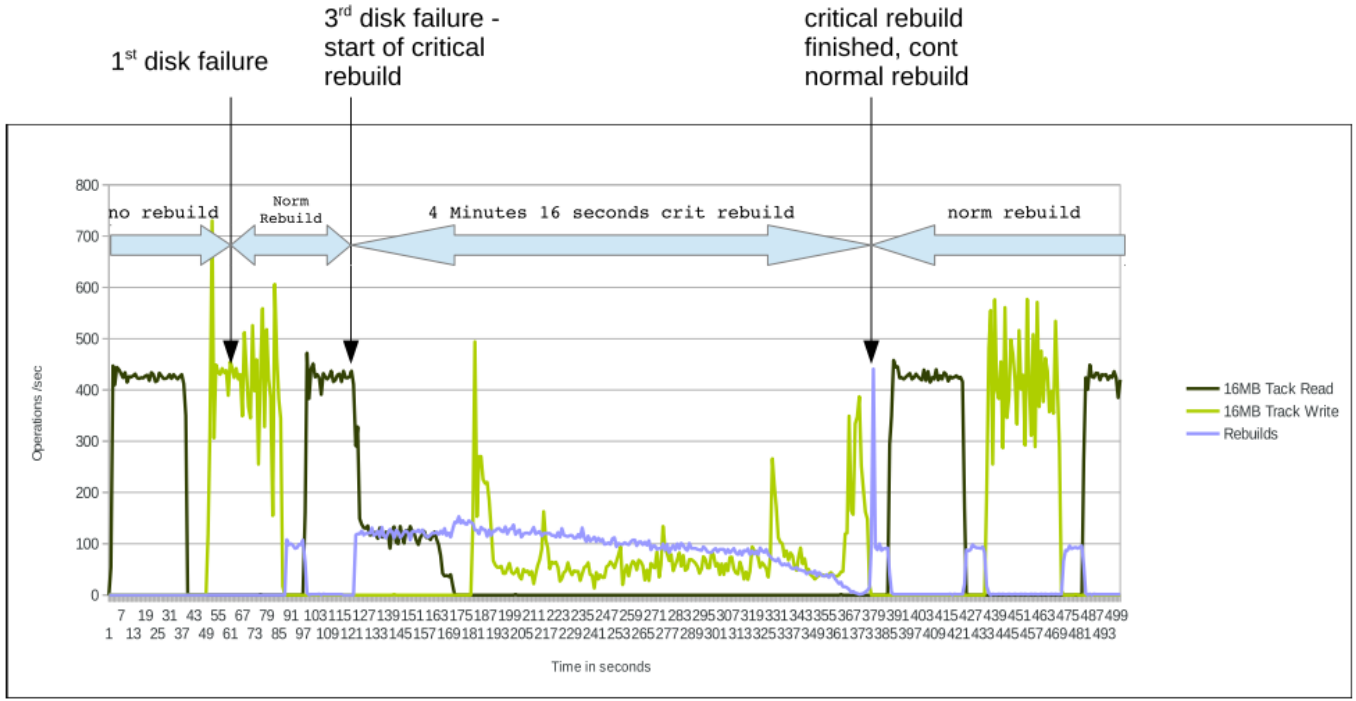
- Rebuild is done by many disks
- Rebuilds can be deferred with sufficient protection

**Better fault tolerance**

- End to end checksum
- Much higher mean-time-to-data-loss (MTTDL)
  - 8+2P: ~ 200 Years
  - 8+3P: ~ 200 Million Years

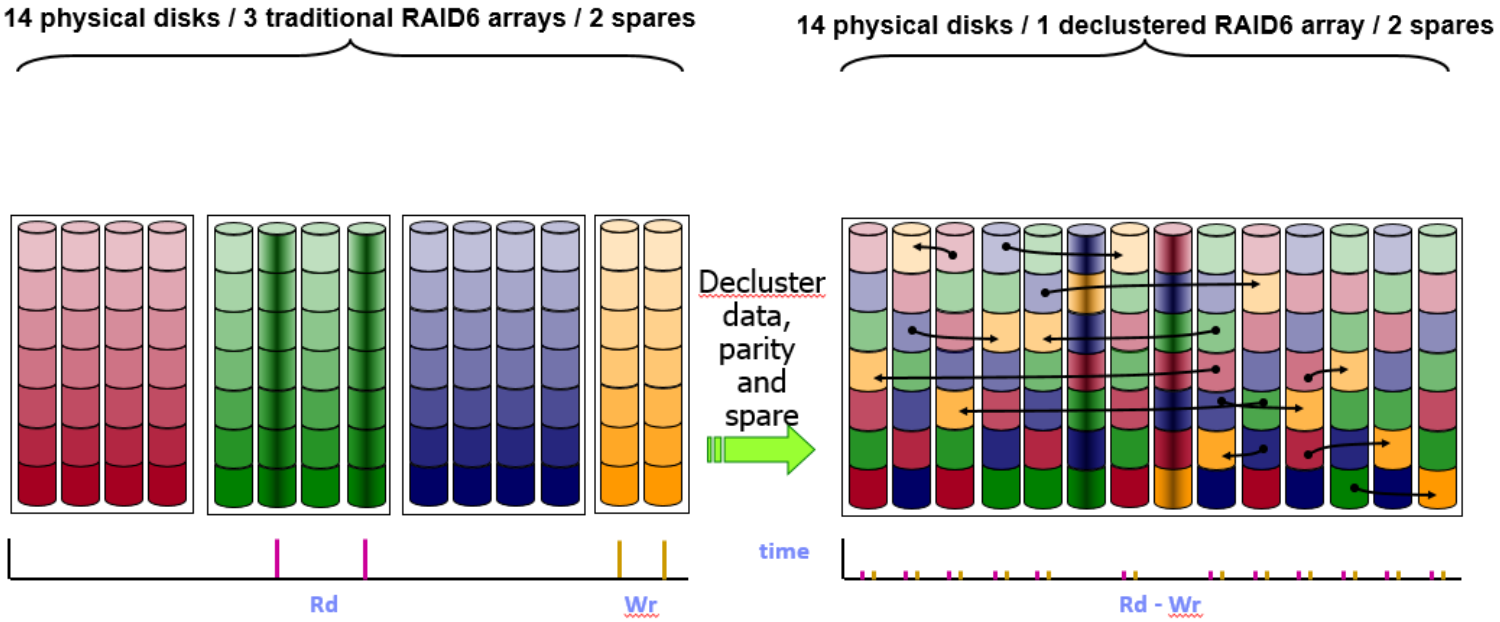


# GNR Technology: Rebuild Test 8+3p on a EL6 with 2TB NL-SAS




As one can see during the critical rebuild impact on workload was high, but as soon as we were back to double parity (+2P) the impact to the customers workload was <5%

# GNR Technology: Distribute rebuild workload on many drives

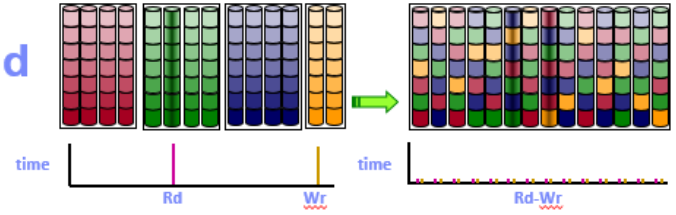



Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs

Rebuild activity spread across many disks, less disruption to user programs

# GNR Technology: Critical Rebuild vs. Rebuild

- **De-clustered RAID: Prioritize Rebuild**
  - Choose 8+2P or 8+3P
  - Failure: One parity left (most common)
    - Rebuild slowly with minimal impact to client workload
  - No parity left: (very rare)
    - Only fraction of stripes have three failures ~ 1%
    - Get back to non-critical (redundant) state in minutes vs. rebuilding all stripes (hours / days) for conventional RAID
- **Optional 2, 3 or 4 way replication**
  - Often used for metadata




Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs

Rebuild activity spread across many disks, less disruption to user programs

# Store everywhere. Run anywhere.

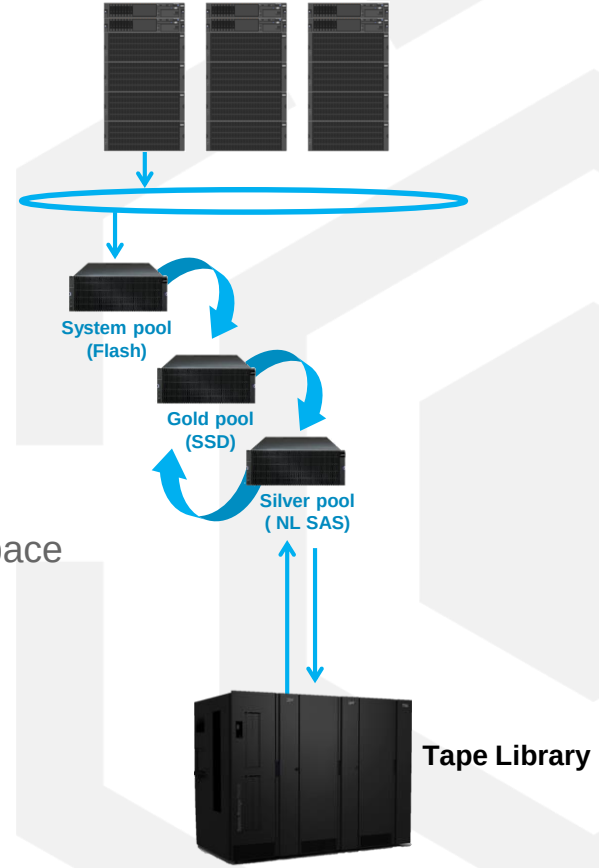
*Optimize Cost and Performance*

## Challenge

- Data growth is outpacing budget
  - Low-cost archive is another storage silo
  - Flash is under utilized because it isn't shared
  - Locally attached disk can't be used with centralized storage
  - Migration overhead is preventing storage upgrades

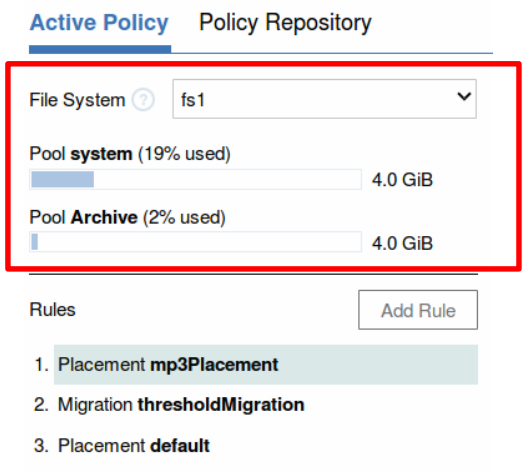
## Automated data placement

- Span entire storage portfolio, including DAS, with a single namespace
- Policy driven data placement & data migration
- Share storage, even low-latency flash
- Automatic failover and seamless file-system recovery
- Lower TCO



# Placement and migration policies (ILM)

- File placement rules enable the automatic placement of new files in a specific storage pool
- File management rules enable the automatic file management during their life cycle, for instance by moving them from one storage pools to another, copying them to archival storage, changing their replication status, compressing or deleting them.
- Example: One filesystem (fs1) with two storage pools (system, Archive)




```
RULE 'mp3Placement'
    SET POOL 'Archive'
        WHERE (LOWER(NAME) LIKE '%.mp3')
```

```
RULE 'thresholdMigration'
    MIGRATE
        FROM POOL 'system'
        THRESHOLD(80,60)
        WEIGHT(100000 -
DAYS(CURRENT_TIMESTAMP) - DAYS(ACCESS_TIME))
    TO POOL 'Archive'
        WHERE USER_ID != 26
```

https://www.ibm.com/developerworks/community/blogs/storageneers/entry/Using_IBM_Spectrum_Scale_Policies_for_Automated_Information_Lifecycle_Management?lang=en_us

# b) Visualization and near real-time analysis (online processing)

- Characterization
  - Control experiment and plan next steps
  - Close to instrument (=outside data center)
  - Requires immediate access to latest measured data

- Spectrum Scale capabilities
  - High-performance, scalable file system
  - Can be mounted on hundreds of nodes and be accessed like a local filesystem

- Caveats
  - Read access impacts data ingest
  - Puts burden on media, busses and controller which are needed for high-speed data ingest
  - May trigger Spectrum Scale Token Activity if directories and files are ingested and accessed on multiple nodes
  - Don’t do polling on directory with many files like: `while true; do ls –lt /incoming | head; sleep 1; done`
  - Consider to duplicate data stream in Data Acquisition (e.g., by using ZeroMQ) to avoid reading from file system

# c) Data movement from Burst Buffer to Central Storage

- Characterization
  - Move data from Burst Buffer close to instrument to Core File system in central data centers
  - Typically over Campus 10GbE, National Research and Science Network or Internet
  - Needs to tolerate network outages
  - Sometimes: Provide access on both sides

- Spectrum Scale capabilities
  - Approach options:
    - Active File Management
    - ‘Homegrown’ based on Policies and Extended Attributes
  - Both approach options allow multiple sites to work on same data and to recover from network failures
  - AFM provides built-in automation and recovery, but must be used as-is
  - ‘Homegrown’ solution has higher initial implementation effort, but allows more customization (e.g., process data and metadata during copying)

# Store everywhere. Run anywhere.

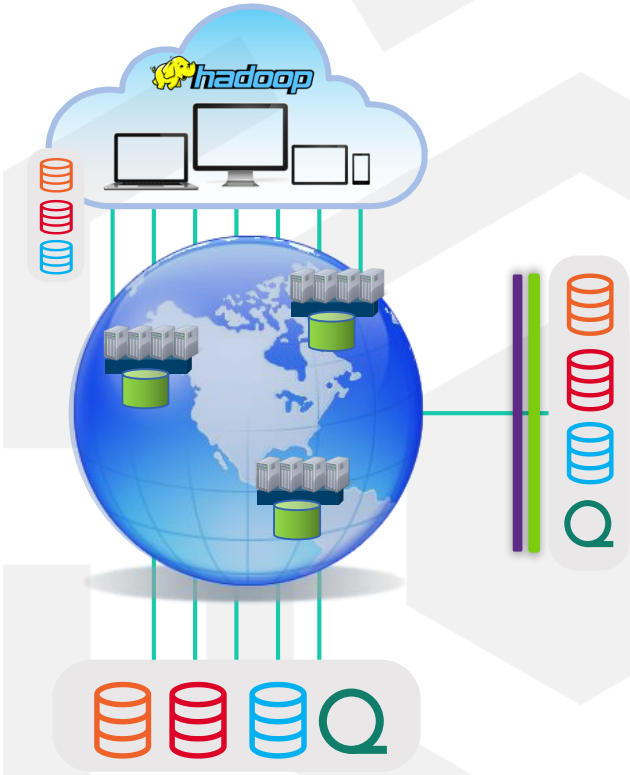
*Enable Global Collaboration*

**Challenge**

- Multiple sites working on same data
  - Remote access is slower than local
  - Consistent metadata & data locking
  - Support for mission critical transactional replication
  - Manage unreliable, remote sites

**Advanced File Management, Routing & Caching**

- Global namespace with fast, consistent metadata
- Latency aware
- Multi-writer and multi-reader
- Automatic failover and seamless file-system recovery



# Spectrum Scale Advanced File Management (AFM)

## Spans geographic distance and unreliable networks

- Caches local ‘copies’ of data distributed to one or more Spectrum Scale clusters
- Low latency ‘local’ read and write performance
- As data is written or modified at one location, all other locations see that same data
- Efficient data transfers over wide area network (WAN)

## Speeds data access to collaborators and resources around the world

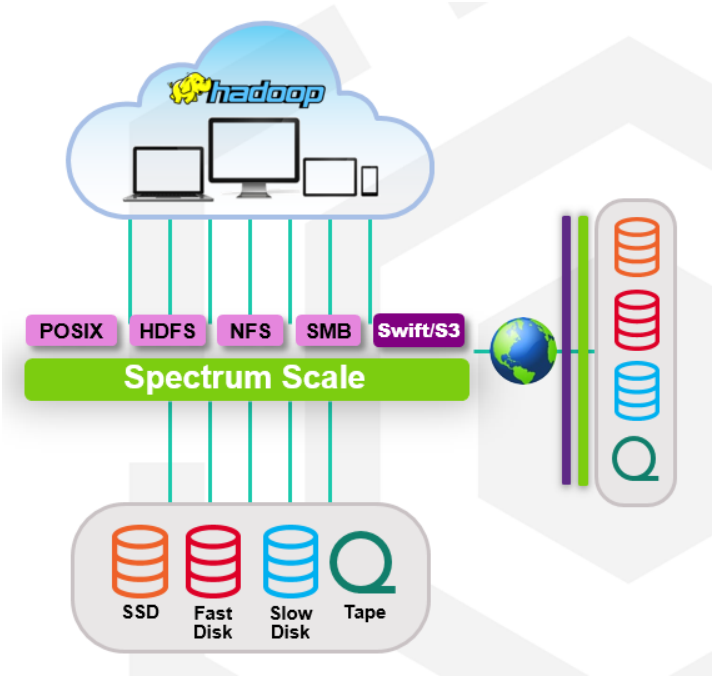
- Unifies heterogeneous remote storage

## Asynchronous DR is a special case of AFM

- Bidirectional awareness for Fail-over & Fail-back with data integrity
- Recovery Point Objectives for volume & application consistency

# d) Deep analysis (offline processing)

- This is solved!
- InfiniBand network between heavyweight Spectrum Scale nodes provides best performance
- External data sources and sinks can still be connected via 10GbE



# e) Data management of Central Storage

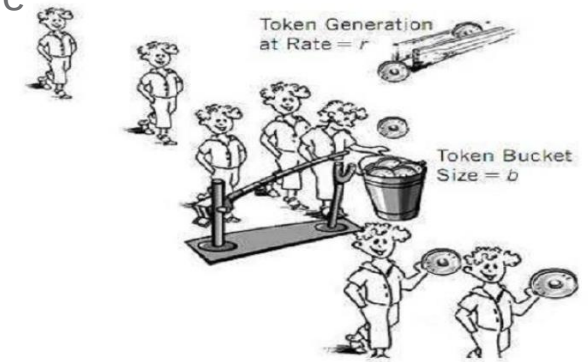
- Characterization
  - Provide and operate petascale file system with hundreds of millions files
  - Reduce the costs for storing and accessing huge amounts of files

- Spectrum Scale capabilities
  - High performance scale-out file system
  - Single name space to simplify access and data management
  - Storage Pools and ILM for cost-effective integration of various media types
  - Policy Engine and Extended Attributes to automate workflows (e.g., find new files and trigger update in iRODS DB)
  - Quality of Service to meet SLA
  - TSM Integration for Backup

# Quality of Service

- Spectrum Scale has great performance, efficiency, etc, etc, but …
- Before QOS – we had no way to control performance of competing tasks/jobs:
  - Restripe, backup, policy scan/ILM/HSM, rcopy and other maintenance tasks – *versus*
  - Real Work: near-real-time decision support, datacollection and crunching

- Spectrum Scale 4.2 introduced QOS for IO operations in 4Q2015
- Multiple token buckets, one token bucket for each combination of:
  - disk pool,
  - QOS class,
  - node



# Quality of Service

- Valid for all filesystem traffic, but need to configure on the cluster that owns the file system
- Currently supported classes: ‘maintenance’ and ‘other’
  - May be used to prevent maintenance tasks from "dominating“ file system performance
  - As of this writing, the following commands are treated as long running Spectrum Scale commands: *mmadddisk, mmapplypolicy, mmcheckquota, mmdefragfs, mmdeldisk, mmdelfileset, mmdelsnapshot, mmdf, mmfileid, mmfsck, mmfsctl/tsreclaim, mmlssnapshot, mmrestripefs, mmrpldisk*
- It is perfectly okay to issue mmchqos at any time. It will affect IO completion rates but it will not “break anything”.

```
mmchqos <fsname> {enable|disable}
```

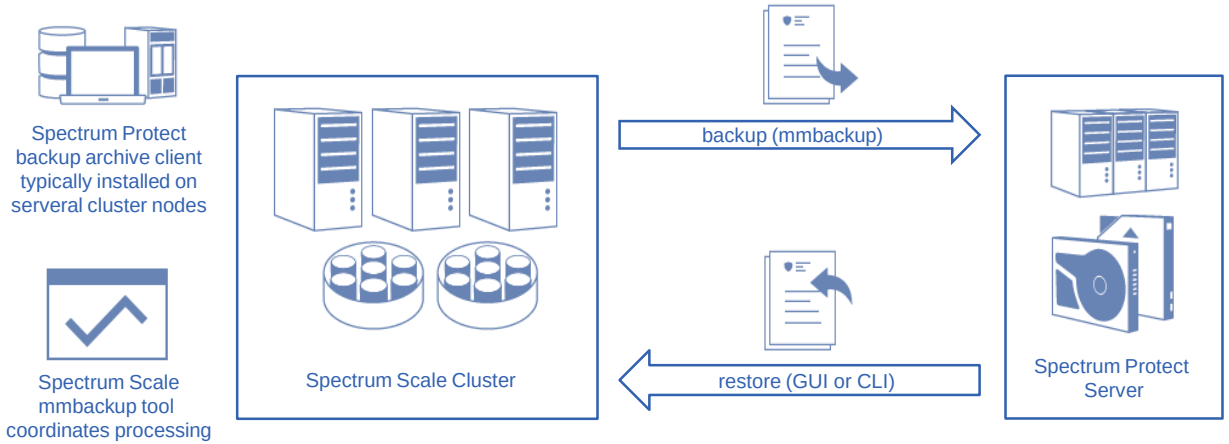
- To cap GPFS maintenance to 300 IOPs:

```
mmchqos <fsname> enable maintenance=300iops,other=unlimited,pool=*
```

- To check quotas

```
mmcheckquota [-v] [-N {Node[,Node...] | NodeFile | NodeClass}]
             [--qos QosClass] {-a | Device [Device ...]}
mmcheckquota {-u UserQuotaFile | -g GroupQuotaFile | -j FilesetQuotaFile}
             [--qos QosClass] Device
mmcheckquota --backup backupDir Device
```

# Backup Of Large Spectrum Scale File Systems




## Function

- Massive parallel filesystem backup processing
- Spectrum Scale mmbackup creates local shadow of Spectrum Protect DB and uses policy engine to identify files for backup
- Spectrum Protect backup archive client is used under the hood to backup files to Spectrum Protect Server
- Spectrum Protect restore (CLI or GUI) can be used to restore files

➔ Use any backup program to backup file, object and Hadoop data
➔ Use Spectrum Protect to benefit from mmbackup and SOBAR to backup and restore huge amounts of data

## Level of file and file system recovery using IBM Spectrum Protect

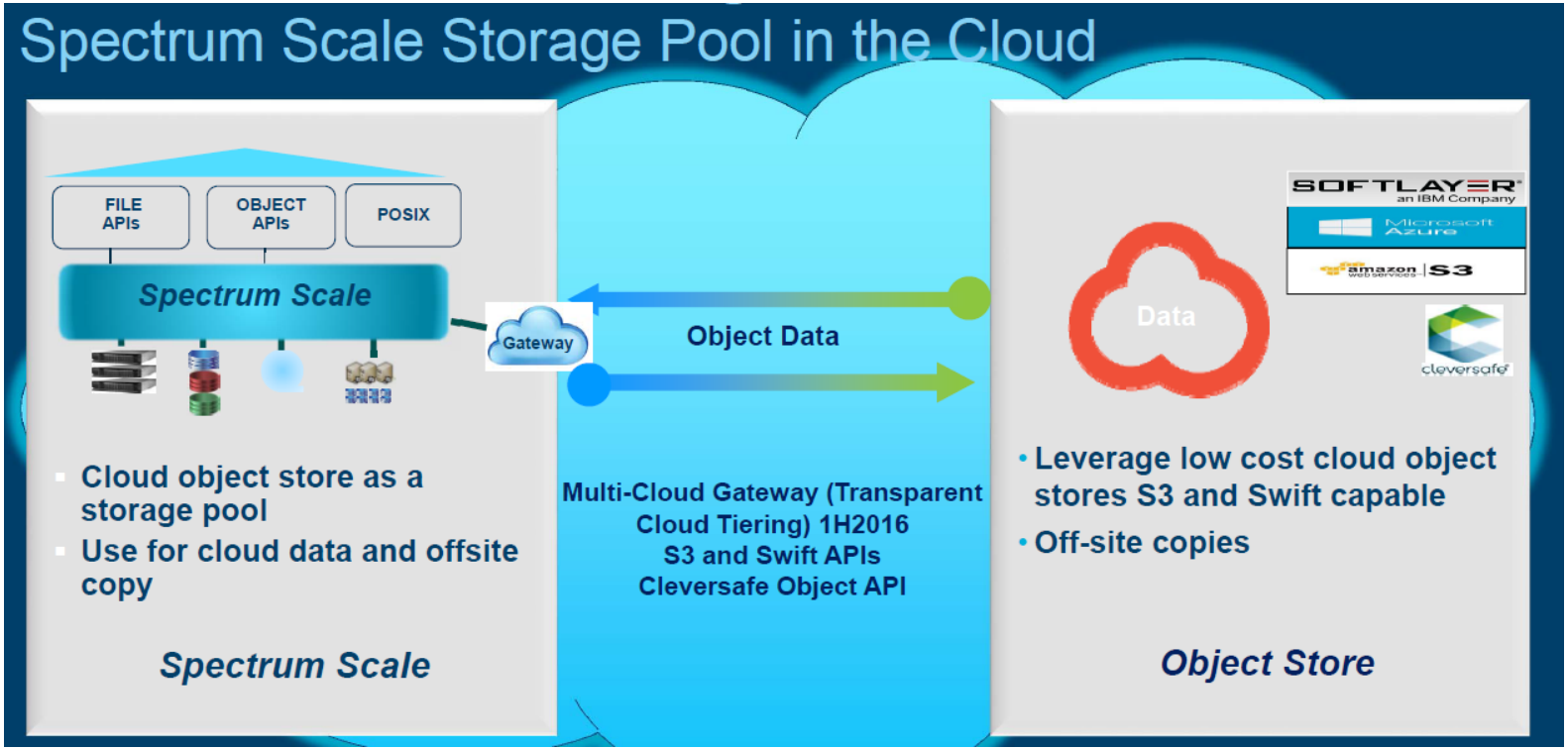
| Restore Type | Restored Content | Comments |
|---|---|---|
| Full Restore | • directories<br>• file data<br>• file metadata (POSIX/ACL/EA) | • full data transfer over network |
| Stub Restore | • directories<br>• file metadata (POSIX) | • quick<br>• next backup requires recall of files<br>• reset of ACL/EA required |
| SOBAR Restore | • directories<br>• file metadata (POSIX/ACL/EA) | • fast<br>• scalable |

# f) Long-term data archiving

- Characterization
  - 10 years and more
  - Tape and cloud is most cost effective
    - Remember: Ingest in cloud is cheap
  - Needs to integrate in existing archive solution

- Spectrum Scale capabilities
  - Integration with a various products
    - High Performance Storage System (HPSS)
    - IBM Spectrum Protect (a.k.a. TSM) HSM and Archive
    - IBM Spectrum Archive (LTFS EE)
    - Transparent Cloud Tiering
    - dCache

# Transparent Cloud Tiering



# Transparent Cloud Tiering – Lightweight Events (LWE)

- Light-weight events are policy rules that are evaluated at specific points in time (open, close, destroy, etc)
  - Rules installed in Spectrum Scale via mmchpolicy
- If the caller or the file's attributes match the policy rule's WHERE clause then the rule's ACTION clause is executed
  - WHERE clause can also match file's extended attributes
  - **RULE 'RuleName' EVENT 'OPEN_READ' ACTION(EXEC('/bin/whatever/command ' || varchar(INODE) || ... )) WHERE XATTR('system.dmattr') IS NOT 'RESIDENT'**
- Events are synchronous
  - Calling thread waits for event completion & can be aborted
  - Can mark file as having permanent damage (return E_IO)
- LWEs can co-exit with DMAPI

# Outline

1) IBM Spectrum Scale Overview
2) Designing an IBM Spectrum Scale Implementation
3) Composable Infrastructure for Ultra Fast Data Acquisition, Data Analysis and Archiving
4) ***Best Practices & Discussion***

# Things we hope(d) would work, but...

- Current architecture result of process during last months
- Detectors as native GPFS clients
  - Old operating systems (RHEL 4, Suse 10 etc.)
  - Inhomogeneous network for GPFS: InfiniBand and 10G Ethernet
- Windows as native GPFS client
  - More or less working, but source of pain
- Active file management (AFM) as copy process
  - No control during file transfer
  - Not supported with native GPFS Windows client
  - Cache behaviour not optimal for this use case
- Self-made SSD burst buffer
  - SSDs died very fast, firmware bug according to vendor
- Remote Cluster UID-Mapping




# Best Practices = Available today (1/3)

- Design the boundaries of the Spectrum Scale cluster and the Spectrum Scale filesystems carefully
  - The slowest node determines the stability and performance of Spectrum Scale functions such as Snapshots und underlying token management
  - A node can be slow due to
    - Underlying hardware
    - Old operating system or kernel
    - Extreme CPU load or memory consumption
    - Slow network connectivity
  - Decouple detectors/control workstations and Spectrum Scale by NFS, SMB, ZeroMQ, …
  - Separate ingest from off-line analysis and archive
- Be nice to the storage
  - Convert the I/O pattern of the ingest stream to large blocksize I/O
  - Do not misuse the storage as message passing mechanism
    - Duplicate the ingest data stream and feed one end into real-time analysis before writing data to persistent storage
  - Understand and honor the underlying Spectrum Scale Token traffic
    - Different nodes need to ingest into different directories
    - Opening a new file that is completely written on a different node is OK
    - Using `ls` to poll for new files kills storage performance



# Best Practices = Available today (2/3)

- Leverage the built-in parallelism and linear scaling of Spectrum Scale
  - Spectrum Scale is optimized for aggregated performance, but it works for many single-client, single-stream workloads.
  - For almost all use cases it is better to stripe all workload across all resources instead of configuring dedicated resources for critical workload.
  - Measure end-to-end performance from ingest via network to storage for one building block. Add more building blocks to achieve the required performance.
  - Extreme ingest streams might need to be partitioned and mapped to multiple concurrent building blocks. Each building block needs its own directory to avoid underlying token traffic.
- A good network design is key to success
  - The performance and reliability of Spectrum Scale depends on a high-speed data network.
  - The path from all ingest nodes to all storage nodes needs to be non blocking.
  - Be careful with the mix of multiple network speeds and network technologies.
  - Consider NFS and SMB to integrate servers with slow network connectivity.
  - Spectrum Scale does striping over redundant InfiniBand links.
  - There is limited striping over bonded TCP connections.



# Best Practices = Available today (3/3)

- Access to experts is key to success for demanding workloads
  - Spectrum Scale is a racing car. There are many tuning options.
  - Seek guidance from an expert to learn driving on your own.
  - Involve professional services and talk to your peers.
  - Do an in-depth proof-of-concept to tune for extreme workloads.
  - Spectrum Scale beginners should start with the basic features and defer the use of advanced features until they have more experience with the whole solution including the underlying hardware. Don’t try everything on day one.
  - Establish in-depth end-to-end monitoring to get an expert for your own deployment.



# Discussion points

- 100GByte/s data ingest can be done with todays infrastructure!
- Flash storage (NVMe & NVMe/F) are entering the data center.
- Do burst buffers provide value add?
- How to integrate Spectrum Scale in the detector’s control workstation?
- How to make data in memory of control workstations persistent as HDF5 file in Spectrum Scale?



# References

- Whitepaper by DESY for CHEP2015:
  'ASAP3 - New Data Taking and Analysis Infrastructure for PETRA III'
  http://iopscience.iop.org/article/10.1088/1742-6596/664/4/042053
- DESY talk at SC15 – GPFS in data taking and analysis for new light source (X-Ray) experiments:
  http://files.gpfsug.org/presentations/2015/SC15-DESY.pdf
- DESY talk at SC17 – Data taking and NFS Services:
  http://files.gpfsug.org/presentations/2017/SC17/gpfs-ug-sc17-flat.pdf
- IBM Case Study:
  http://www-03.ibm.com/software/businesscasestudies/us/en/corp?synkey=P699680J02746M12
- Spectrum Scale Trial VM:
  https://www.ibm.com/account/reg/us-en/signup?formid=urx-21182
- Spectrum Scale User Group:
  http://www.spectrumscale.org/
- User Group Mailing List:
  http://www.spectrumscale.org/join/
- Discussion Forum:
  https://www.ibm.com/developerworks/community/forums/html/forum?id=11111111-0000-0000-0000-000000000479
- IBM Spectrum Scale
  http://www.ibm.com/systems/storage/spectrum/scale/