How to build an Autonomic data storage Hub on IBM Z and LinuxONE using Spectrum Scale

Wilhelm Mild
IBM Executive IT Architect
IBM R & D Lab Germany

2018 VM Workshop

June 28. – June 30. 2018

Greensboro, NC

IBM

IBM Spectrum Storage options for IBM Z

IBM Spectrum Control	Analytics-driven data management to reduce costs by up to 50 percent	
IBM Spectrum Protect	Optimized data protection to reduce backup costs by up to 53 percent	
IBM Spectrum Archive	Fast data retention that reduces TCO for active archive data by up to 90%	
IBM Spectrum Virtualize	Virtualization of mixed environments stores up to 5x more data	
IBM Spectrum Accelerate	Enterprise storage for cloud deployed in minutes instead of months	
IBM Spectrum Scale	High-performance, highly scalable storage for all data	
IBM Spectrum Access	A certified storage solution for container environments for integration with cloud	

FlashSystem

Any Storage

Private, Public or Hybrid Cloud

IBM Z

IBM LinuxONE

IBM Z the I/O monster

Scalability, Security, reliability
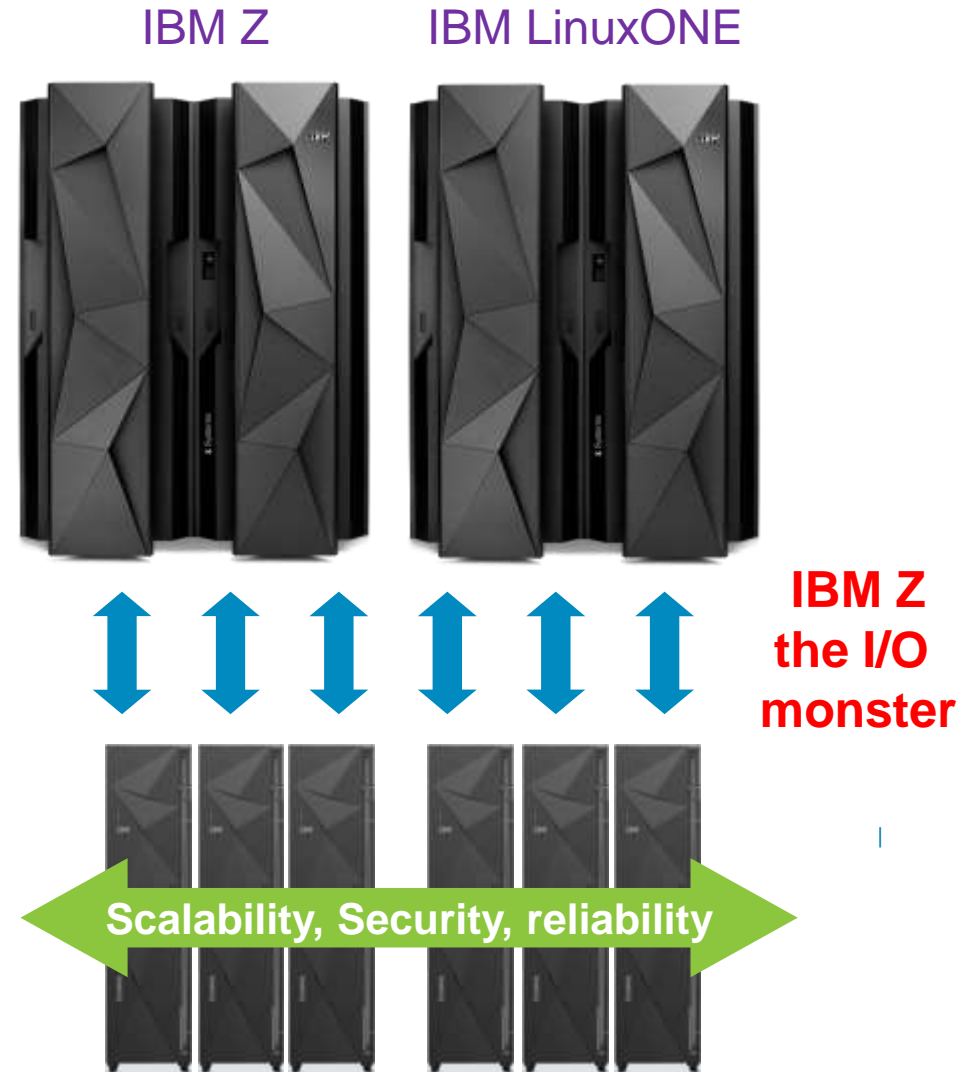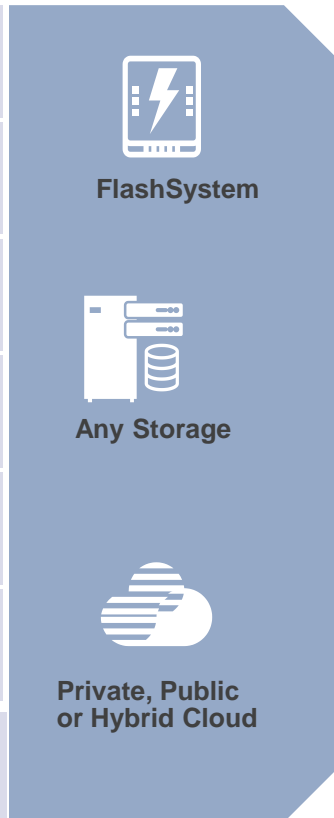
What if …

- ..you can use your applications in parallel and don't care of simultaneous data access

- ..your files will relocate automatic based on heat map, usage or importance

- ..your applications can use the data transparent from multiple plattforms

- ..you have automatic HA / failover

- ..you can even Backup / Archive the data automatically

**This is reality and you can implement it NOW!
with
The Autonomic data storage hub on Linux on Z !**

IBM Z   IBM LinuxONE

**IBM Z
the I/O
monster**

Scalability, Security, reliability

The most important component - IBM Spectrum Scale
– for Cluster Data Management at Scale, Automation and Secure

Many ways to access data

Policy based data migration

Highly scalable

High Performance

Multi-Cluster Support

Backup / Restore
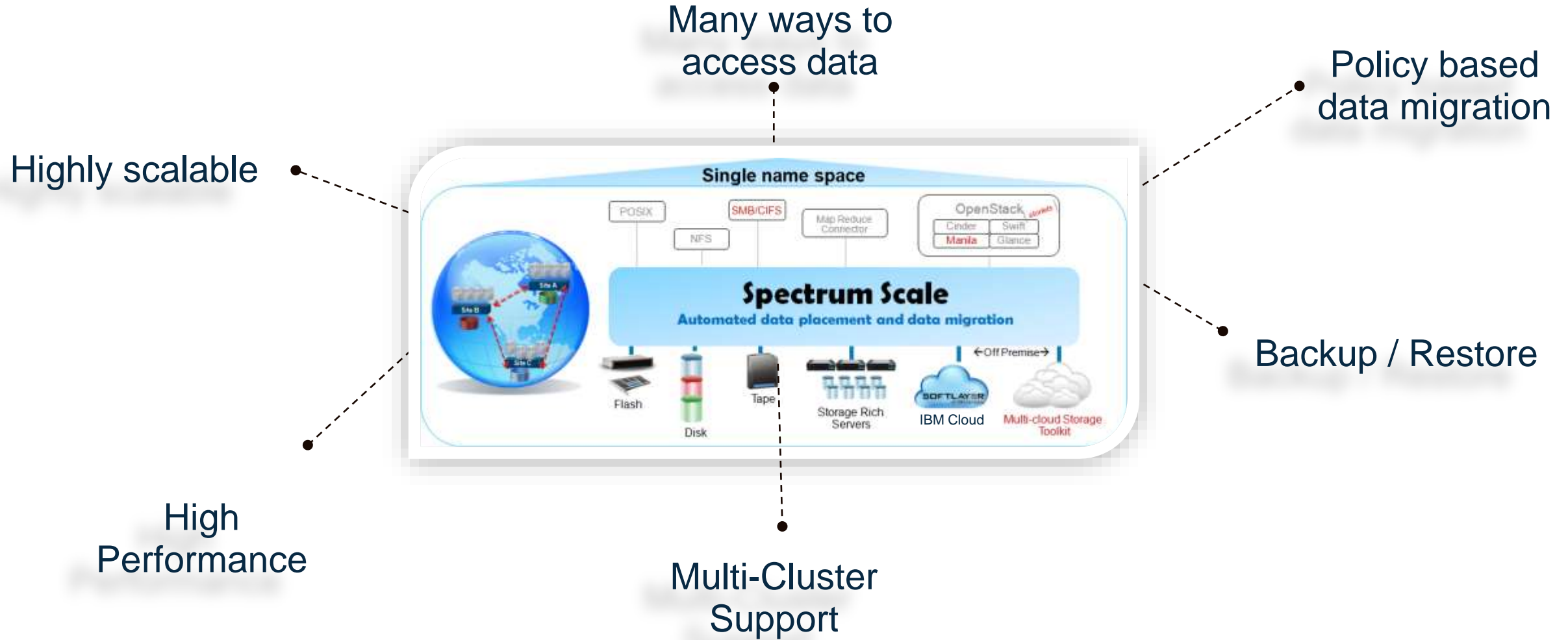
... enhanced with a complete storage integration stack

A Blueprint for IT Simplification with built in investment protection
"Your recipe-book for a complete private cloud"

IBM Spectrum Access Blueprint

IBM Spectrum Access

Data management for operating private cloud platforms with hybrid, multi-cloud services

IBM Cloud public

other public clouds

Modern
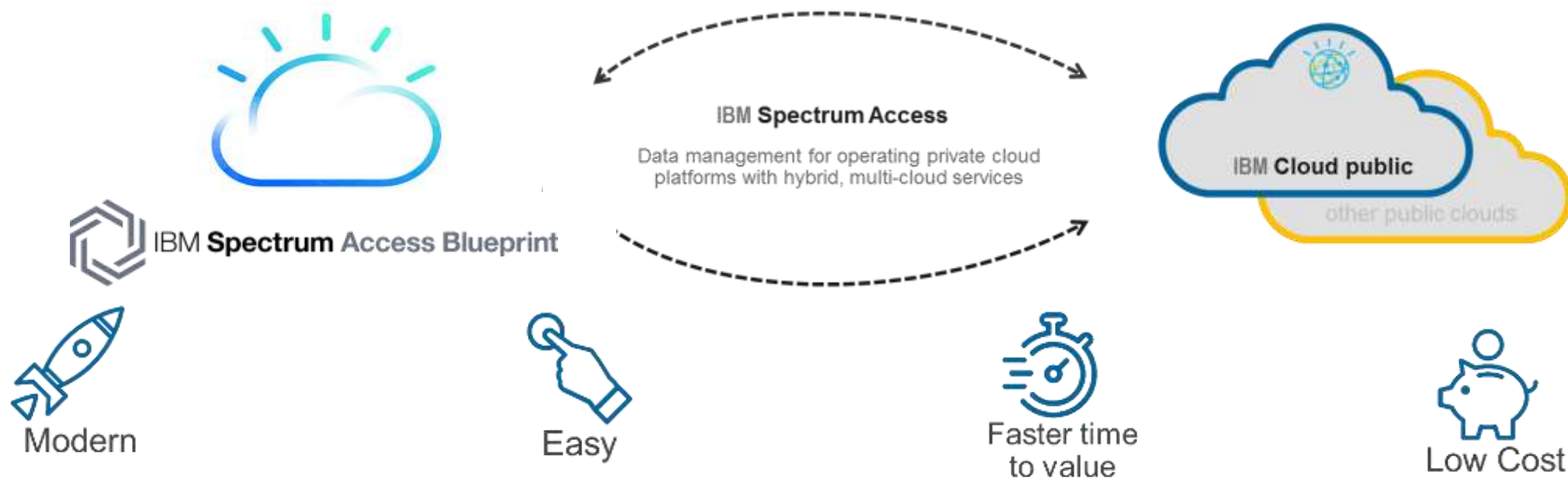
- Personal On-premises Cloud environment
- **Fully integrated IaaS + PaaS capabilities**
- Enterprise Secure

Easy

- Pre-tested, easy to deploy, configure and manage
- Quickly Create, Run and Manage cloud apps
- **Automation and provisioning** with Docker containers and Kubernetes orchestration

Faster time to value

- **Validated & blueprinted** for quick deployment
- Dynamically provision, manage and protect your Cloud Private data

Low Cost

- Scale up, down or out
- **Easily fits your budget**

Parallel access, Scalability and Performance = IBM Spectrum Scale

- IBM Spectrum Scale based on GPFS technology

- Parallel access from all clients to all storage: end to end

- High performance access to small and large files

- High BW access to single file
  - Wide striping
  - Large block size
  - Parallel access to files from multiple nodes
  - Efficient deep pre-fetching: read ahead, write behind
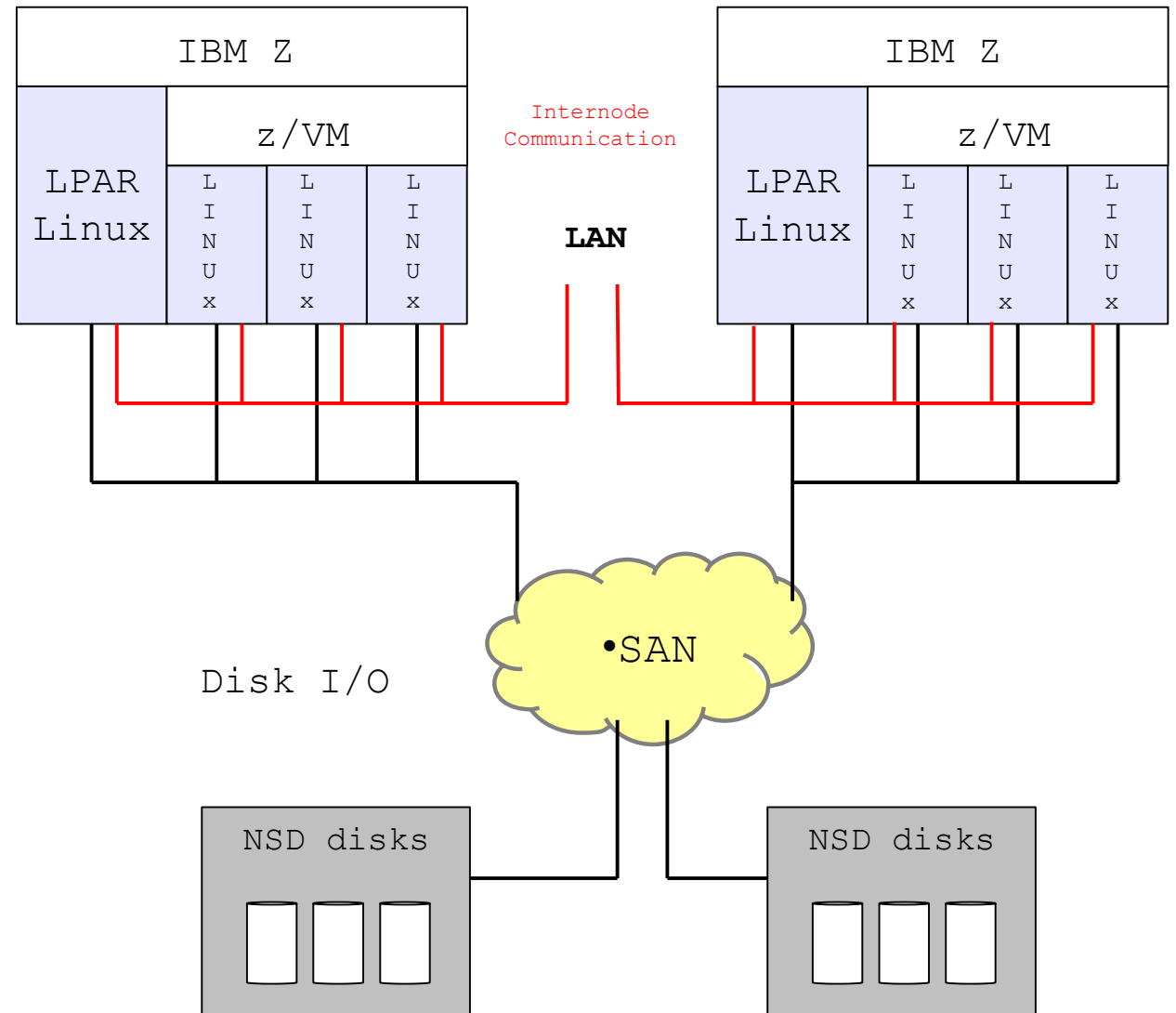 - Highly multi-threaded daemon

IBM Z   IBM LinuxONE

Parallel Access

IBM Z the I/O monster

Single Name Space
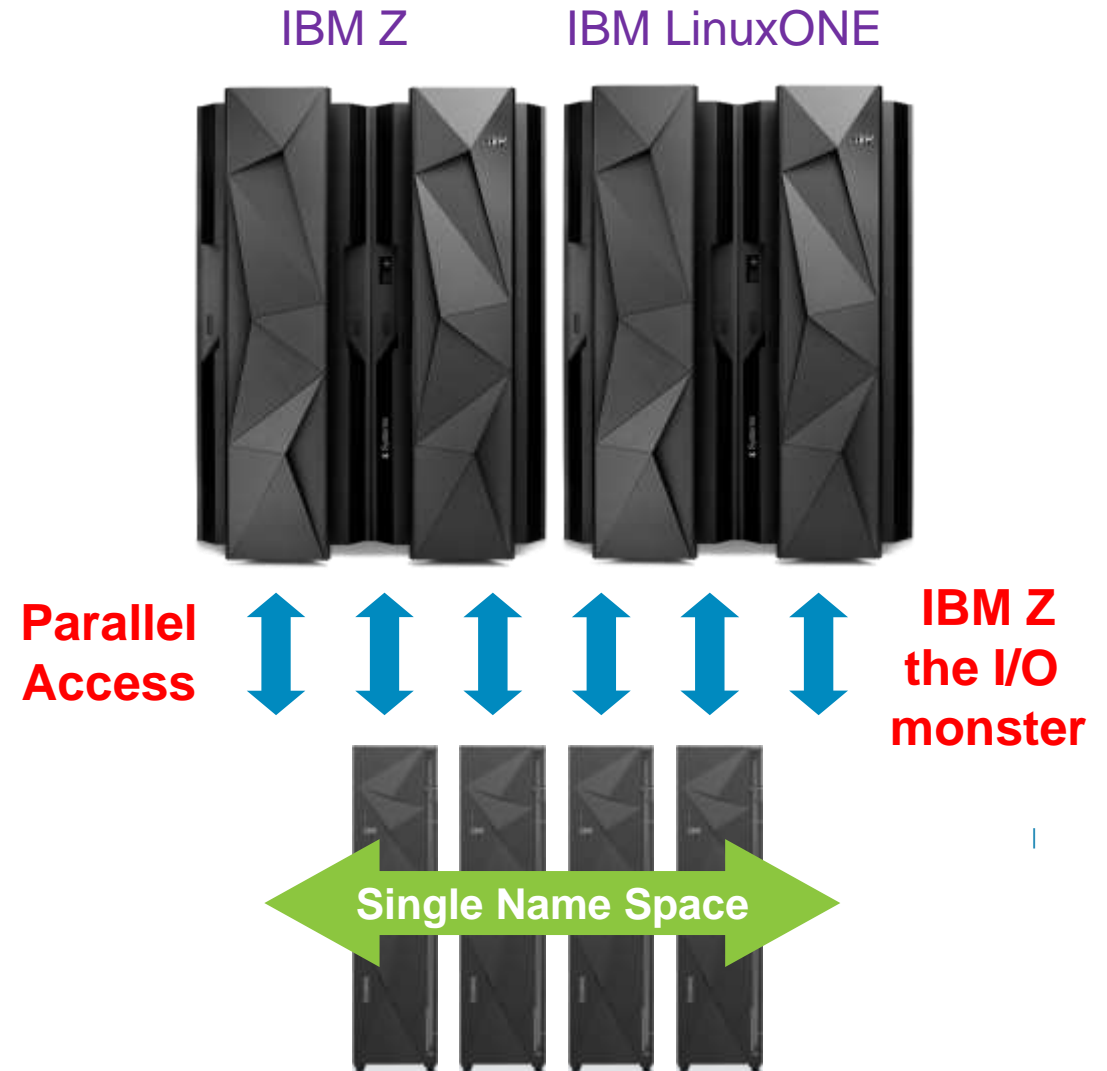
IBM Spectrum Scale cluster - overview

- IBM's shared disk, parallel cluster file system
- Cluster: 1 to 16,384* nodes, fast reliable communication, common admin domain
- Shared disk: all data and metadata on storage devices accessible from any node through block I/O interface ("disk": any kind of block storage device)
- Parallel: data and metadata flow from all of the nodes to all of the disks in parallel.

Spectrum Scale cluster nodes

Network
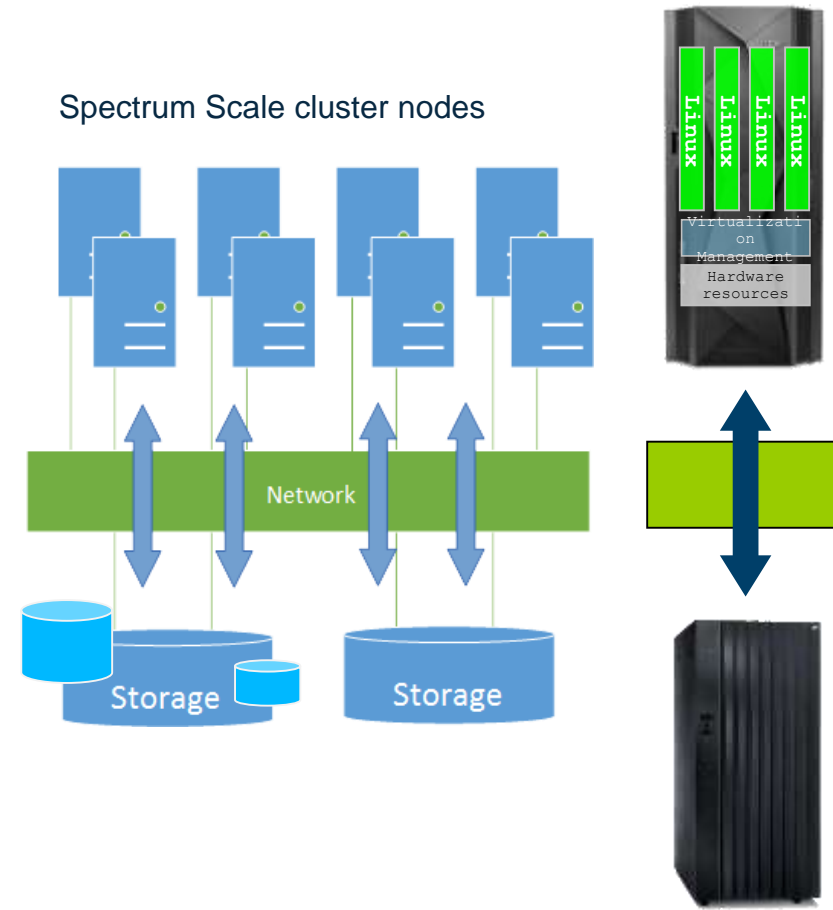
Storage   Storage

# 1) Spectrum Scale cluster – as Shared Disk (SAN) Model

- Every cluster node has direct access to the storage disks through the SAN

- Internode communication (administration communication) is enabled through network connections

- Spectrum Scale does not send disk data from one cluster node to another using the network

- This architectural model can achieves a better performance than the NSD client/server model

IBM Z

z/VM

LPAR Linux

L I N U x

L I N U x

L I N U x

Internode Communication

LAN

IBM Z

z/VM

LPAR Linux

L I N U x

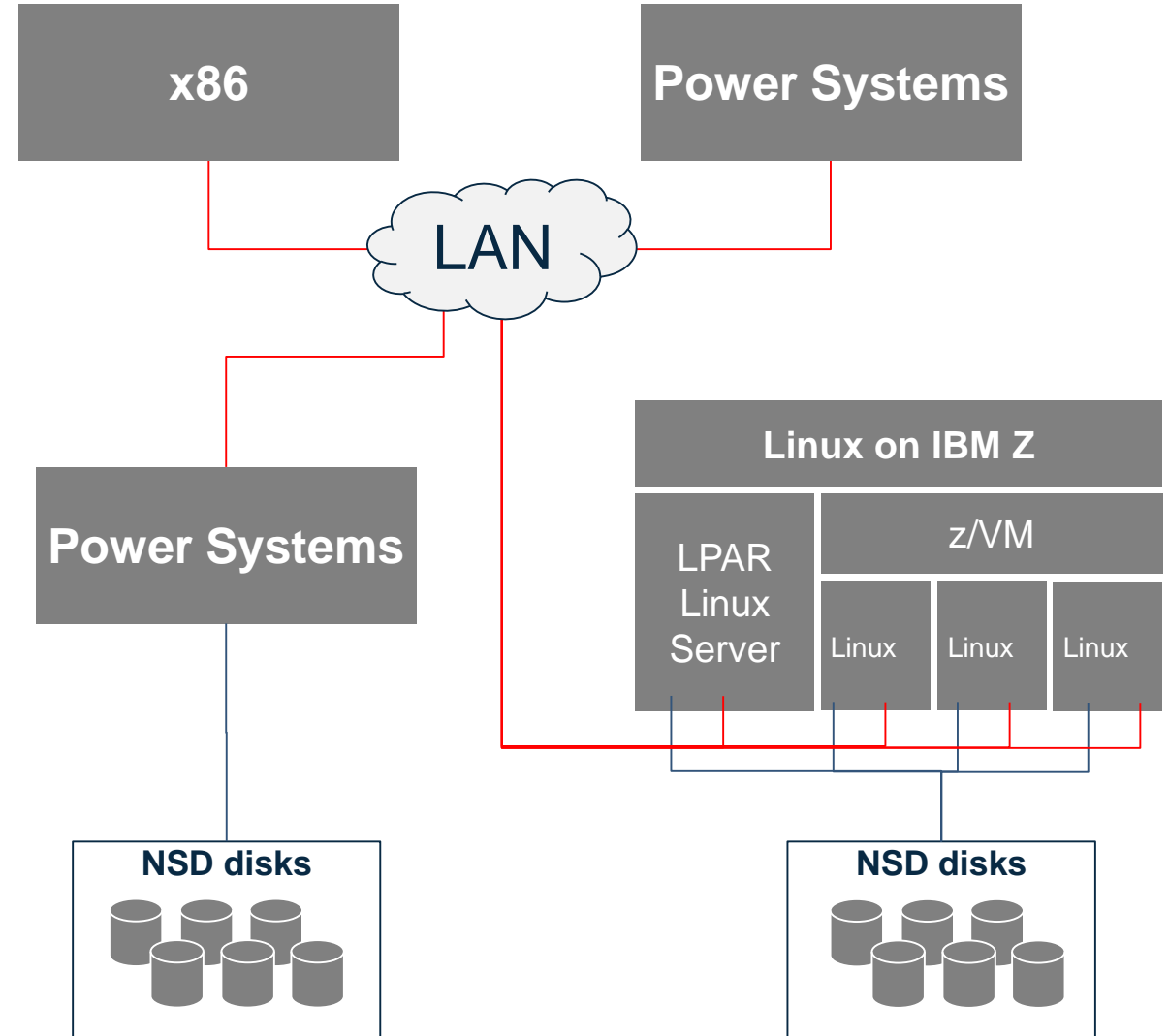L I N U x

L I N U x

Disk I/O

•SAN

NSD disks

NSD disks

2) Spectrum Scale cluster – as Network Shared Disk (NSD) Client/Server Model

- Only a couple of cluster nodes have direct access to the data disks and serve disks to other nodes

- NSD client node data requests are fulfilled via an NSD server node

- This requires a high-speed network with low latency for the best performance

- HA: An NSD client node can be set up with both direct access and access through an NSD server. If direct access is lost, the data access is assured through the available NSD server

- Use disk connectivity on multiple NSD server nodes for each disk to guard against loss of NSD server availability
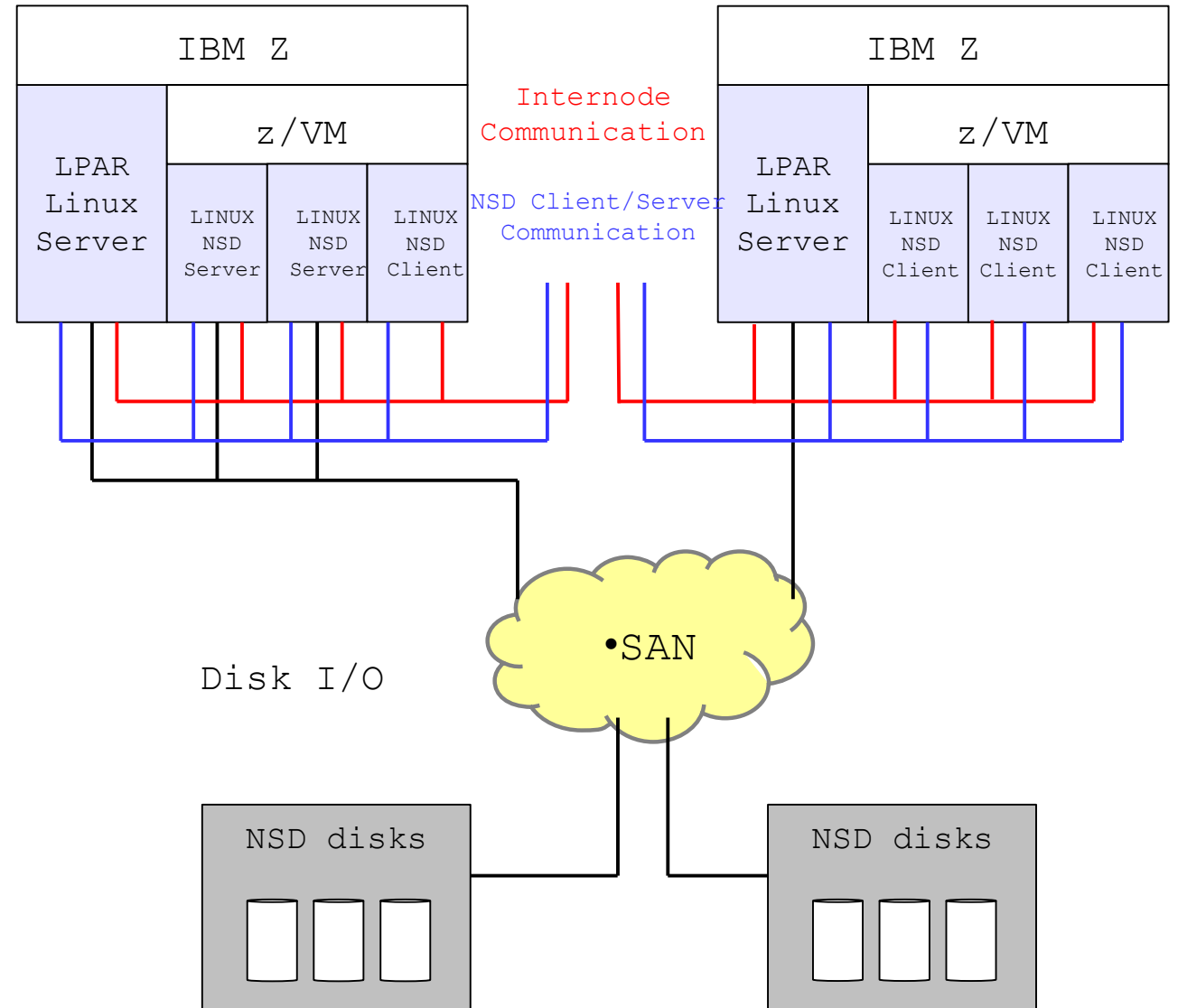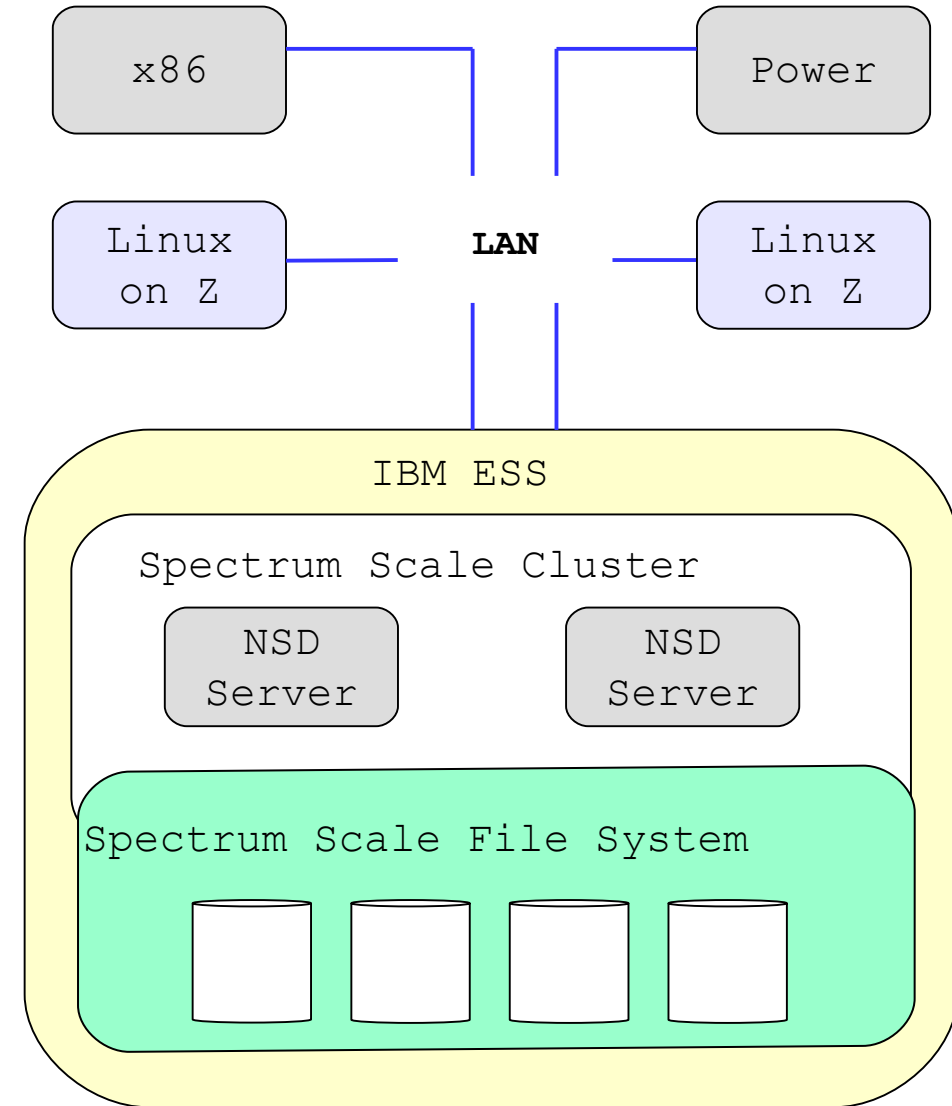
Heterogeneous Cluster Support

- Mix of

  - X86 servers

  - Power servers running RHEL, SLES, AIX, Ubuntu or Debian

  - Linux on Z servers running Linux

x86		Power Systems

LAN

Power Systems		Linux on IBM Z

LPAR Linux Server

z/VM

Linux | Linux | Linux
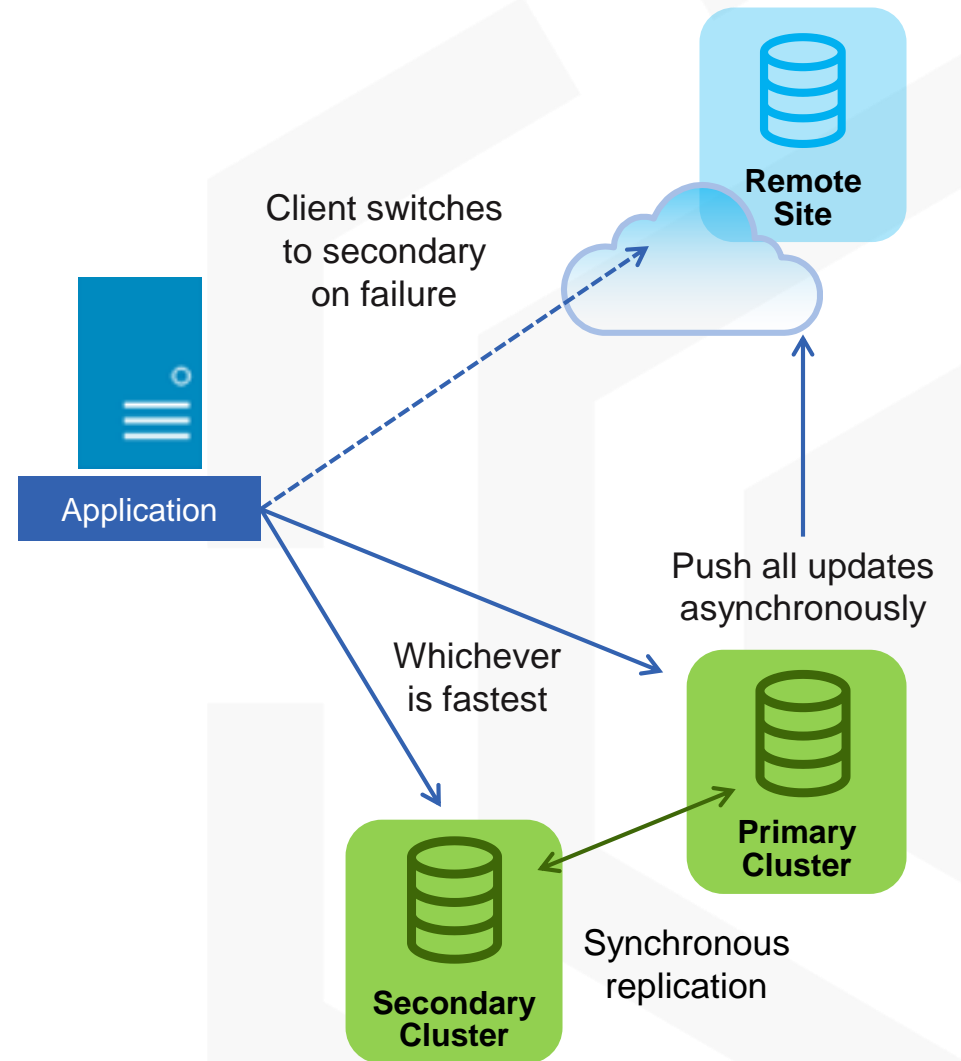
NSD disks

NSD disks

What is an Elastic Storage Server (ESS)

- Elastic Storage Server (ESS) is a (storage) cluster on its own – across architectures

- It is an integrated file based storage solution consisting of:
  - IBM servers
  - IBM storage
  - Spectrum Scale Software including Spectrum Scale RAID software
  - Networking components

- Spectrum Scale for Linux on Z is fully integrated in ESS
  - ESS V4.5.0 and above

New

x86

Power

Linux on Z

LAN

Linux on Z

IBM ESS

Spectrum Scale Cluster

NSD Server

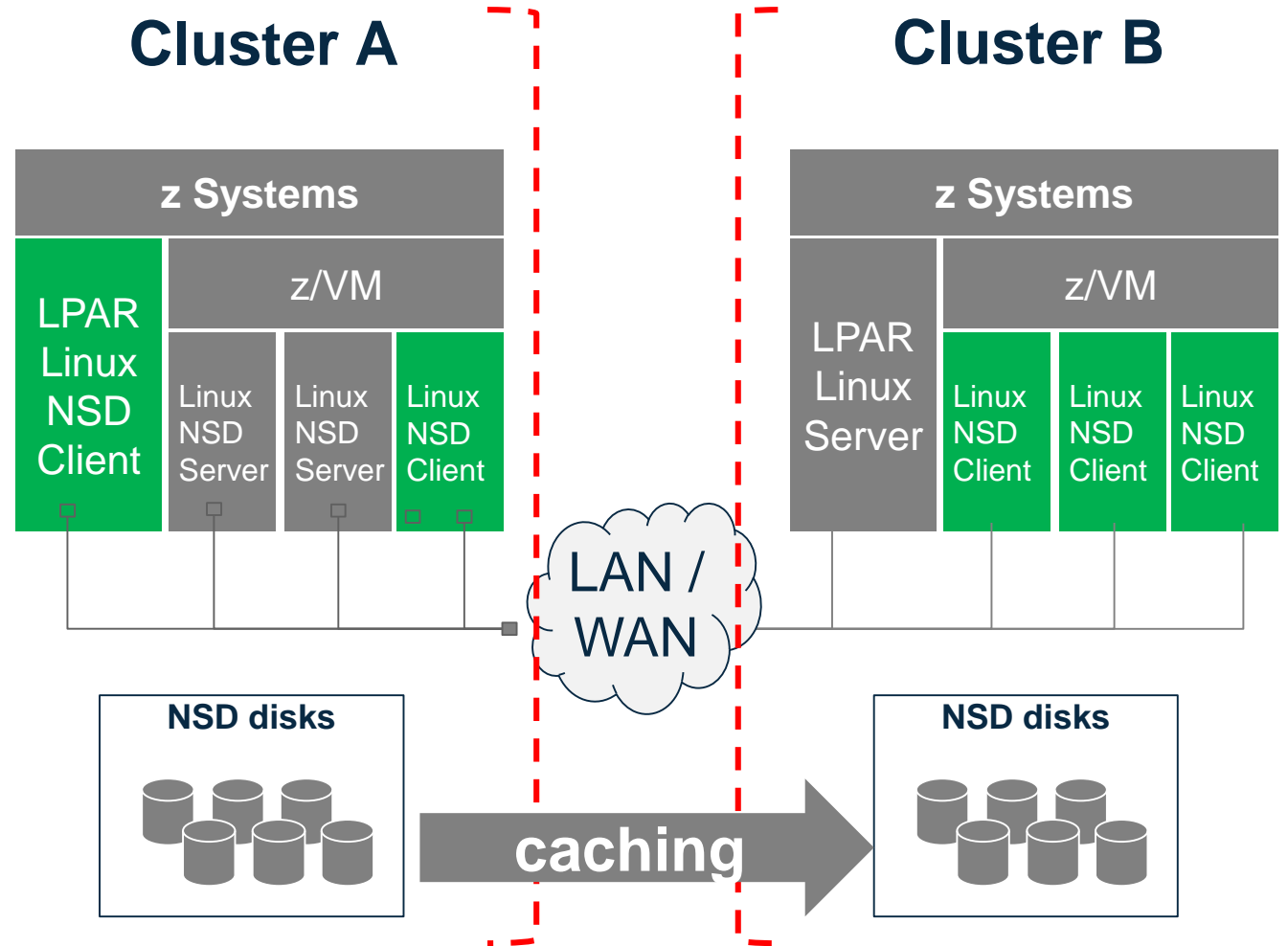NSD Server

Spectrum Scale File System

Replication options and High Availability

- Synchronous replication
  - Active/Active data access
  - DR with automatic failover and seamless file-system recovery
  - Up to 300km

- Asynchronous replication
  - Active/Passive data access
  - Based upon AFM
  - Write now, copy later across network
  - Write to Active, Read from fastest
  - Any storage target, including cloud

- GDPS based replication / failover

Remote Site

Client switches to secondary on failure

Application

Push all updates asynchronously

Whichever is fastest

Primary Cluster

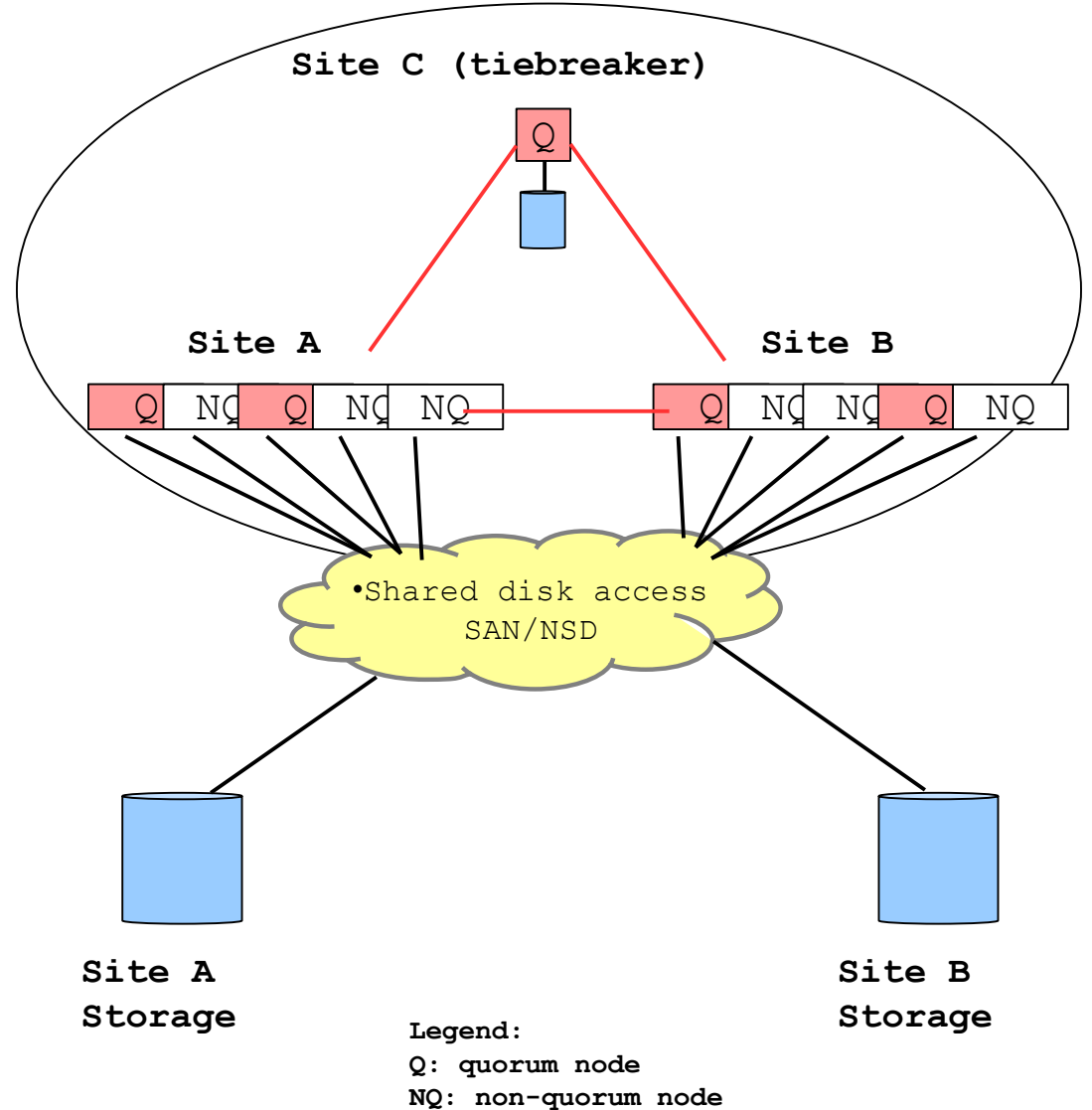Secondary Cluster

Synchronous replication

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 the same data
  - Efficient data transfers over wide area network (WAN)

- Data is copied from cluster A to B
  - Read on demand
  - Asynchronous write-back

Cluster A

z Systems

LPAR Linux NSD Client	z/VM		
	Linux NSD Server	Linux NSD Server	Linux NSD Client

LAN / WAN

NSD disks

caching

Cluster B

z Systems

| LPAR Linux Server	z/VM		
	Linux NSD Client	Linux NSD Client	Linux NSD Client

NSD disks

Synchronous mirroring utilizing GPFS replication

- A stretched cluster is a single IBM Spectrum Scale cluster defined across multiple geographic sites

- The goal of a stretched cluster is to provide high availability against catastrophic hardware failures by replicating of the file system's data to a geographically separated site

- One geographically dispersed cluster ("stretched" cluster)
  - All nodes in either site have SAN/NSD access to the disks
  - Site A storage is duplicated in site B with GPFS replication

Site C (tiebreaker)

Site A

Site B

•Shared disk access SAN/NSD

Site A Storage

Site B Storage

Legend:
Q: quorum node
NQ: non-quorum node

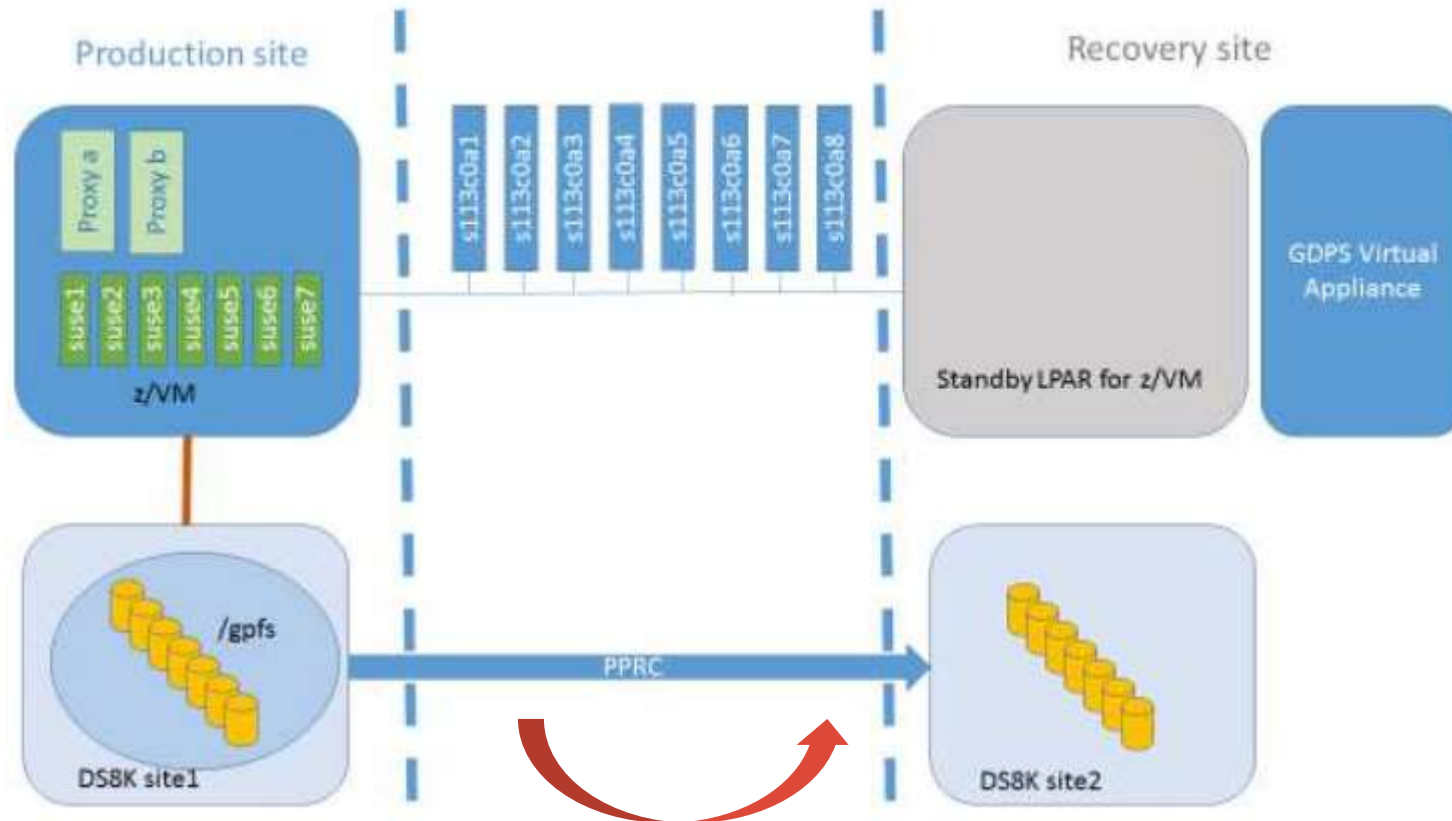# GDPS integration – via GDPS Virtual Appliance

- Data is/can be read from any storage site
- Data write is only successful, when the update is committed from both storage sites (→ synchronous data replication)

- Both data and metadata will be replicated
  - On file system level
- Replication relies on failure groups
- Failure Group: collection of disks that could become unavailable simultaneously, e.g.,
  - Disks attached to the same storage controller
  - Disks served by the same NSD server
- Typically the storage of each site creates one failure group
  - Reason: common point of failure
- Important to set failure groups correctly to have effective file system replication.

https://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102640
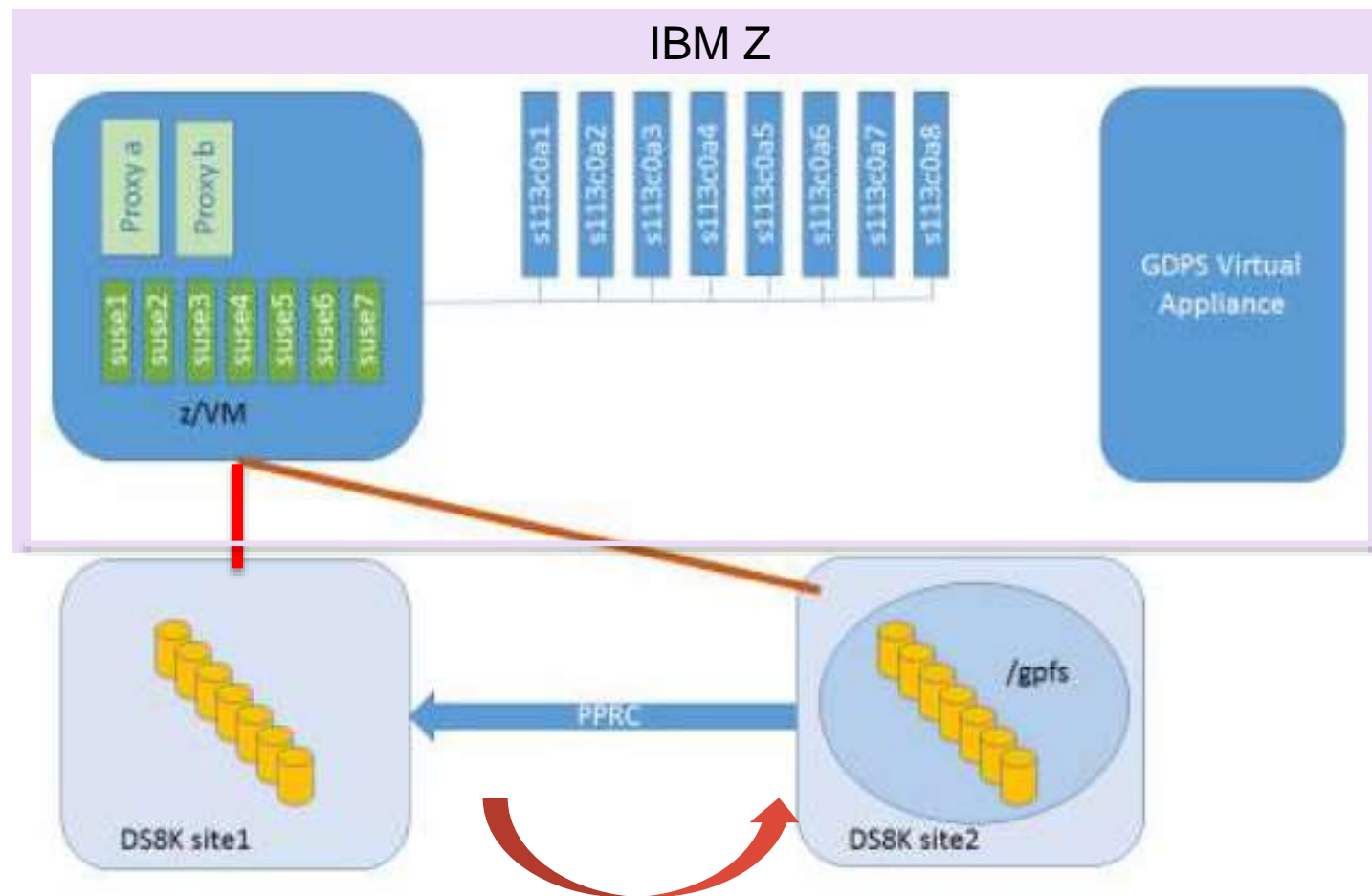
GDPS integration – via GDPS Virtual Appliance

/gpfs is mounted on all nodes on the production site and the client nodes (s133xxxx).
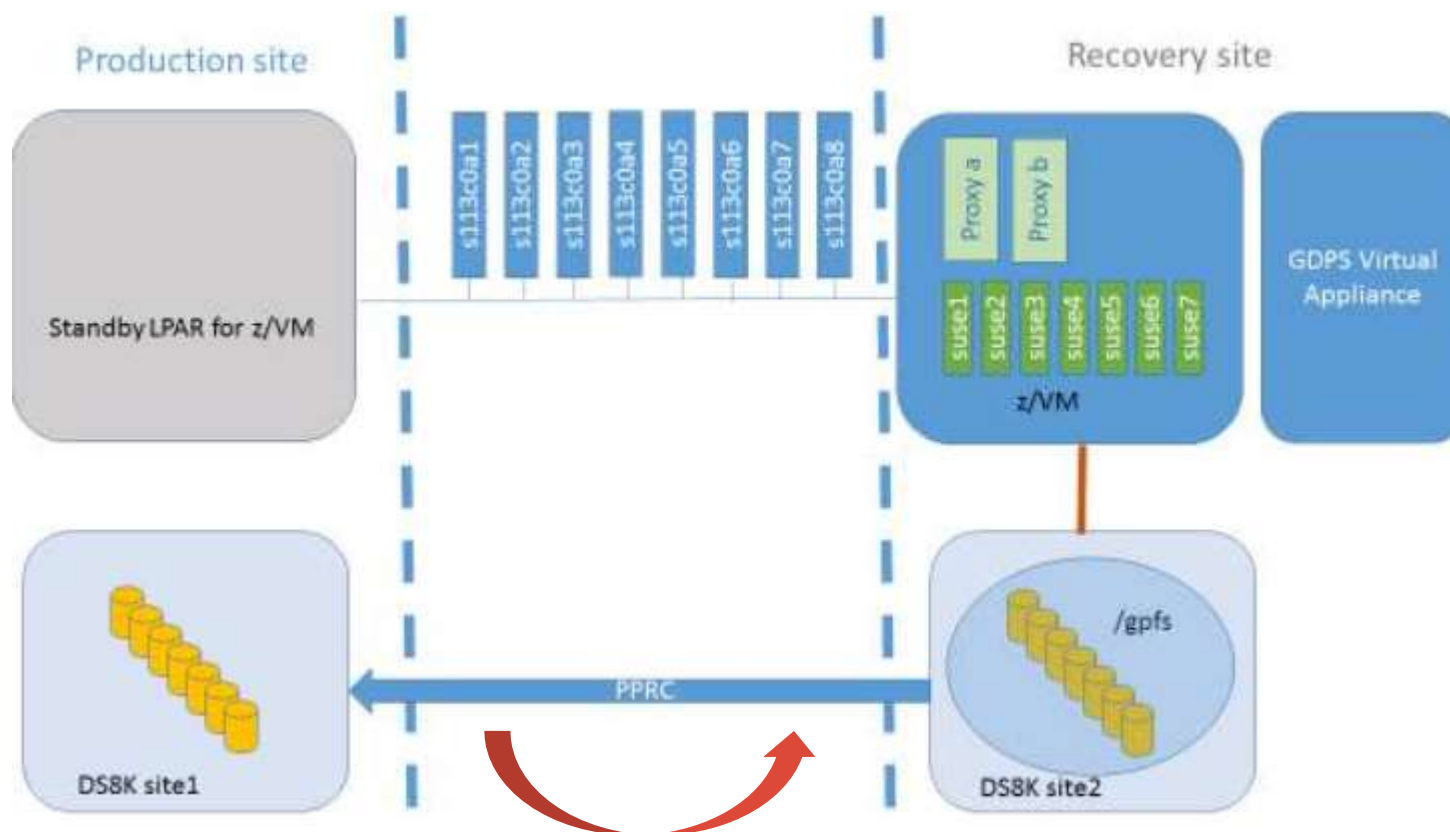NSDs are on the production site.

https://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102640

GDPS integration – via GDPS Virtual Appliance

After the site switch, NSDs and GPFS servers are on the recovery site, /gpfs is remounted on all nodes at startup

Store everywhere. Automatic placement and tiering
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, 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

**System pool
(Flash)**

**Gold pool
(SSD)**

**Silver pool
(NL SAS)**

Tape Library
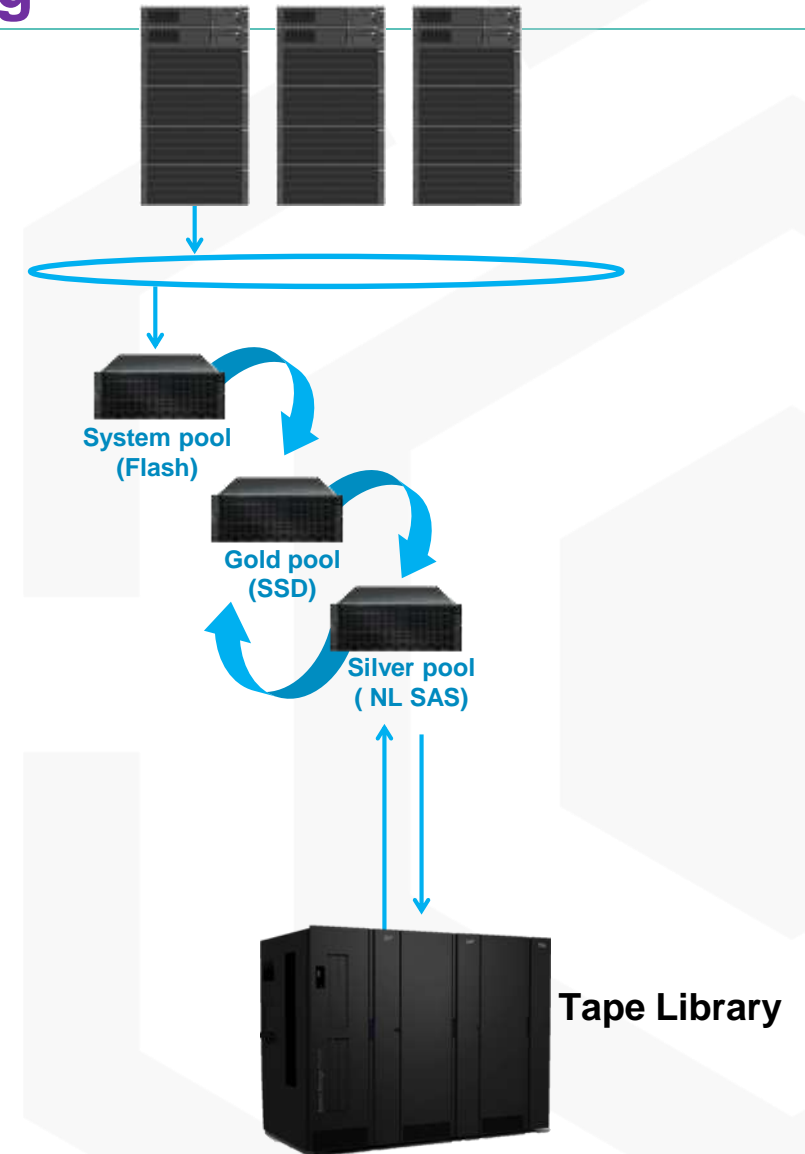
Introduction to Information Lifecycle Management (ILM)

- Optimize costs with ILM, by storing data on the most appropriate storage medium over its life time.

- This starts with the creation of data and continues over the lifecycle until deletion

- There are two key techniques:
  - **Placement:** assures that a file is initially created on the most appropriate storage medium
  - **Migration**: assures that files are migrated to the most appropriate storage medium over their life cycle
    - Integration with Spectrum Protect (formerly TSM)

- **Policy-driven** automated tiered storage management
  - ILM toolkit helps to automate data management for sets of files and pools of storage

- ILM and Spectrum Scale automatically determines where to physically store your data regardless of its placement in the logical directory structure.
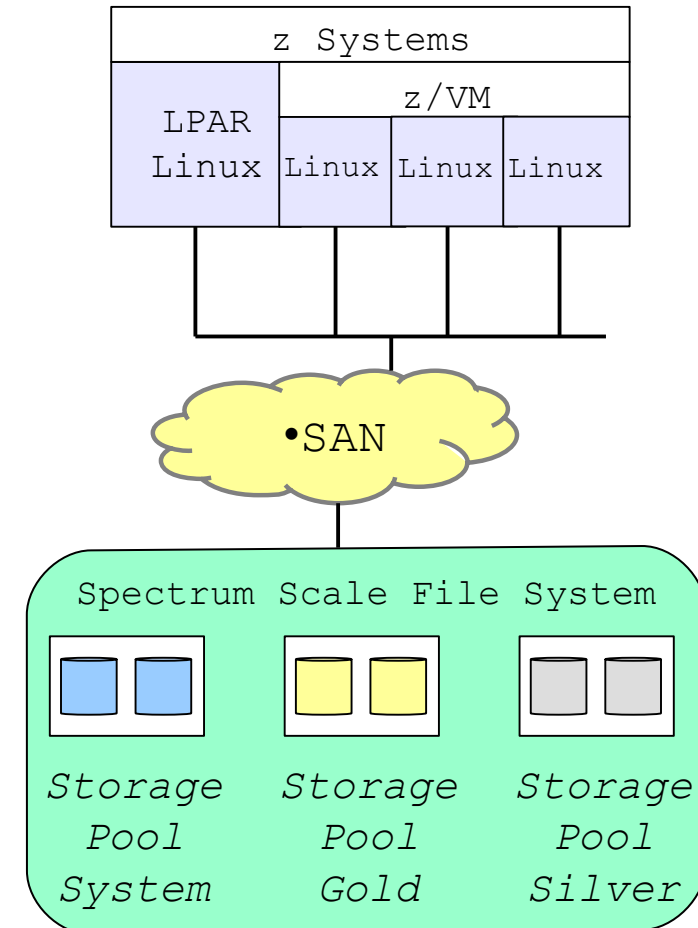
ILM Tools and policies

Storage pools
- Storage pools provide grouping of storage that are managed together

Filesets
- Filesets (named subdirectories) allow you to organize data

File placement policies
- File placement policies assign data to pools on file creation

File management policies
- File management policies automate migration/ deletion/replication/ reporting

Placement Policies

IN

Filesets

Management Policies

Storage Pool

Storage Pool
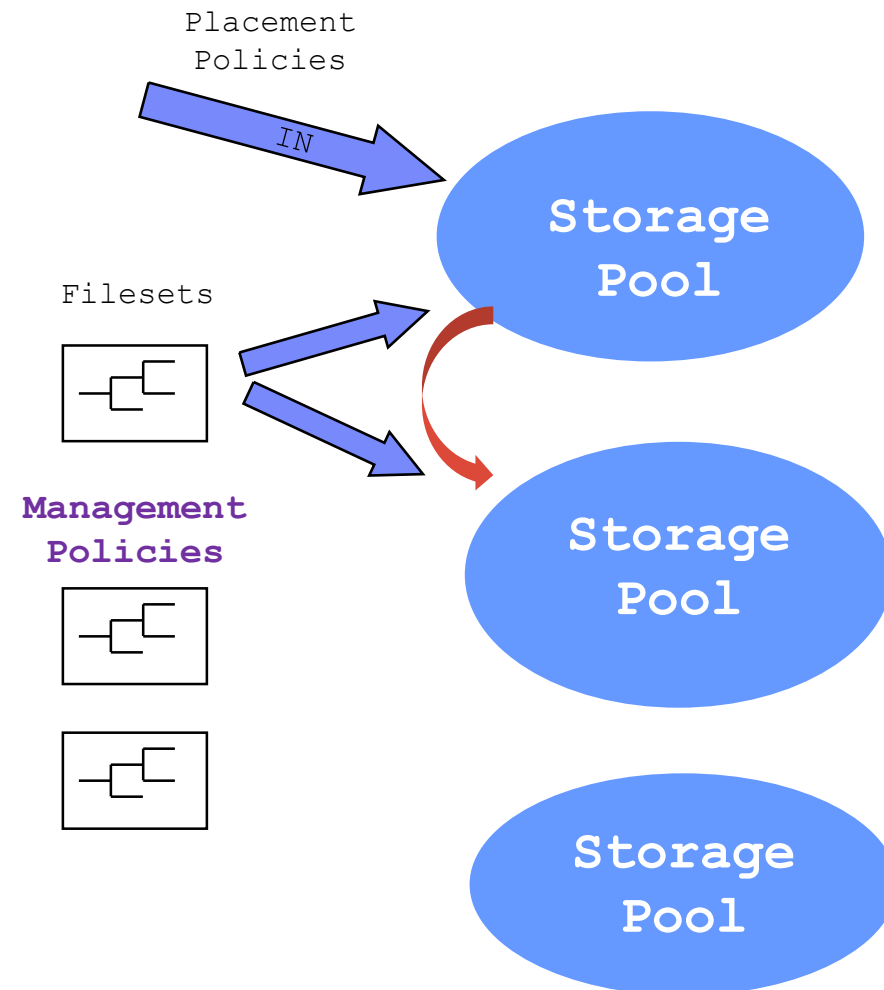
Storage Pool

Storage Pools

- Motivation:
  - Not all storage is the same: some is faster, cheaper, more reliable, ...
  - Not all data are the same: some are more valuable, important, popular, ...

- Storage Pool: A named collection of disks with similar attributes intended to hold similar data
  - System pool: one per file system; holds all metadata
  - Data pools: zero or more: holds only data (up to 7 data pools)
  - External pool: off-line storage (e.g. tape) for rarely accessed data

- Through the use of polices, files may be placed in one of several storage pools according to user-specified rules

```
                    z Systems
                              z/VM
            LPAR
            Linux  Linux  Linux  Linux
```

•SAN

```
      Spectrum Scale File System

      [ ▯ ▯ ]    [ ▯ ▯ ]    [ ▯ ▯ ]

      Storage    Storage    Storage
       Pool       Pool       Pool
      System      Gold      Silver
```

Filesets

- A fileset is a sub-tree of a file system
  - provides a means of partitioning the file system
  - allows to administrate at a finer granularity than the entire file system,
    e.g. disk space limits, user/group quota, snapshots,
  - That can be used to refer to a collection of files in policy rules
  - In many ways behaves like an independent file system

- After creation a fileset has to be linked to an arbitrary point within the file system

- Once linked, a fileset can be populated via normal means, that is, by copying and creating files.

/user1
/user1/appl1
/user1/appl2/dir_a
/user1/appl3/dir_b   ◁ Fileset 1
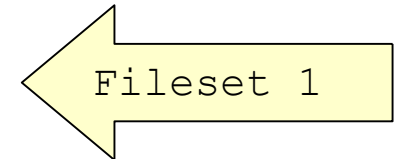
/user2/
/user2/dir1
/user2/dir2   ◁ Fileset 2
/user2/dir3
/user2/dir4
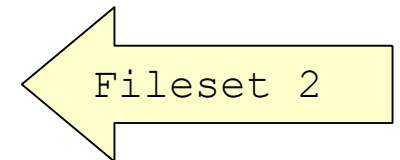
/user4
/user4/data1
/user4/data2
/user4/data3   ◁ Fileset 3
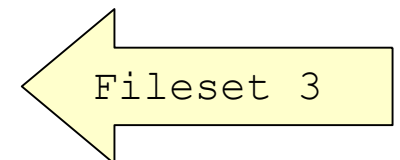
Policies and rules

- **Policy**: A set of user-specified rules that match data to the appropriate pool
  - SQL-like syntax for selecting files based on file attributes
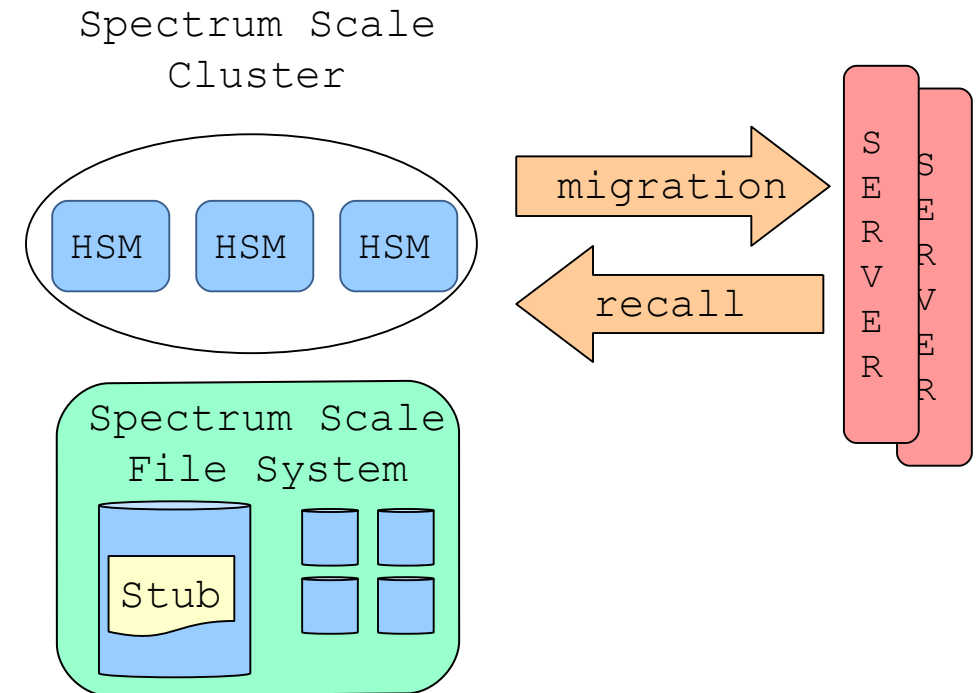
```
RULE 'Clean System' MIGRATE FROM
    POOL 'System' THRESHOLD(85,40)
    WEIGHT(KB_ALLOCATED)
    TO POOL 'Gold'
```

- **File placement policy:**
  - evaluated at file creation time, determines initial file placement and replication

- **File management policy:**
  - used to manage files during their life cycle, can move data between pools, can change replication status and can delete data

Default

System

Clean System

Upgrade To Gold

Gold

Clean Gold

Silver

Delete Silver
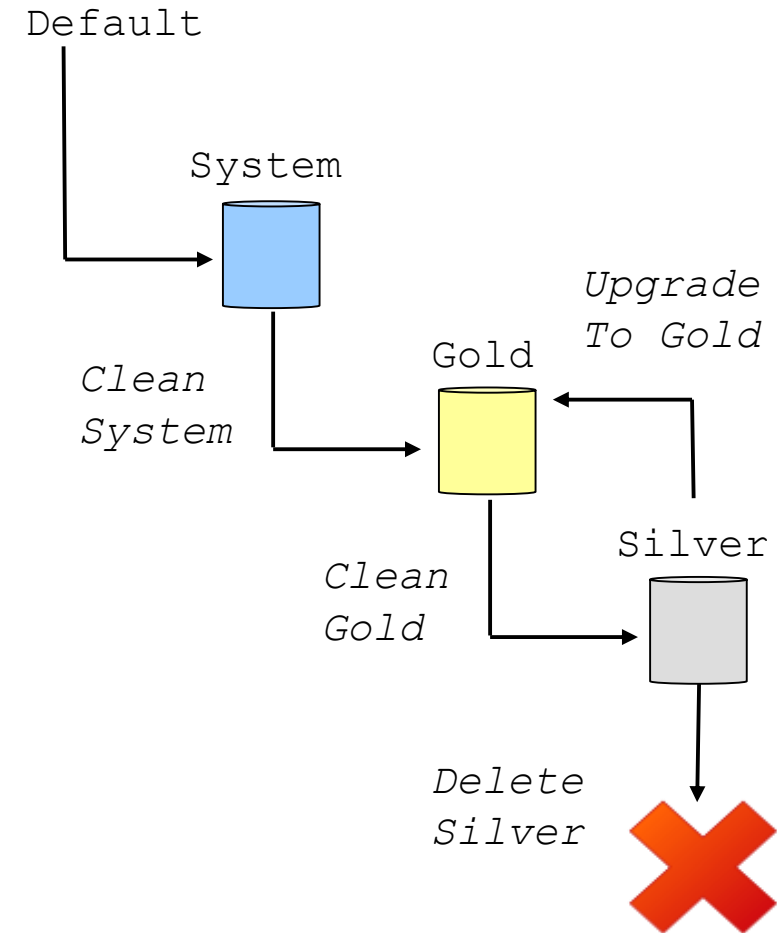
IBM Spectrum Protect (aka TSM) HSM client

- HSM automatically migrates files from on-line to less expensive near-line storage, based on policy settings

- Small stub files are retained on disk, appearing as the original files, thus ensuring transparency of HSM for user applications

- Stub files contain information pointing to corresponding entry in Spectrum Protect server database

- Migrated files are recalled automatically back to disk as required

- Massive parallel migration and cluster wide distributed parallel recall of files, if needed, tape optimized recall for mass recalls in batch mode

Spectrum Scale Cluster

HSM HSM HSM

migration

recall

SERVER

SERVER

Spectrum Scale File System

Stub

Legend:
HSM: Spectrum Protect HSM (Hierarchical Storage Management) Client
Server: Spectrum Protect Server

Ease of Use: The GUI for Spectrum Scale tasks

- Simplified Management
  - Manage filesets, quotas
  - Manage Life Cycle

- Monitoring Dashboard
  - health, events, capacity in a single customizable dashboard

- Easy performance management
  - GUI for Performance visibility with graphs, metrics

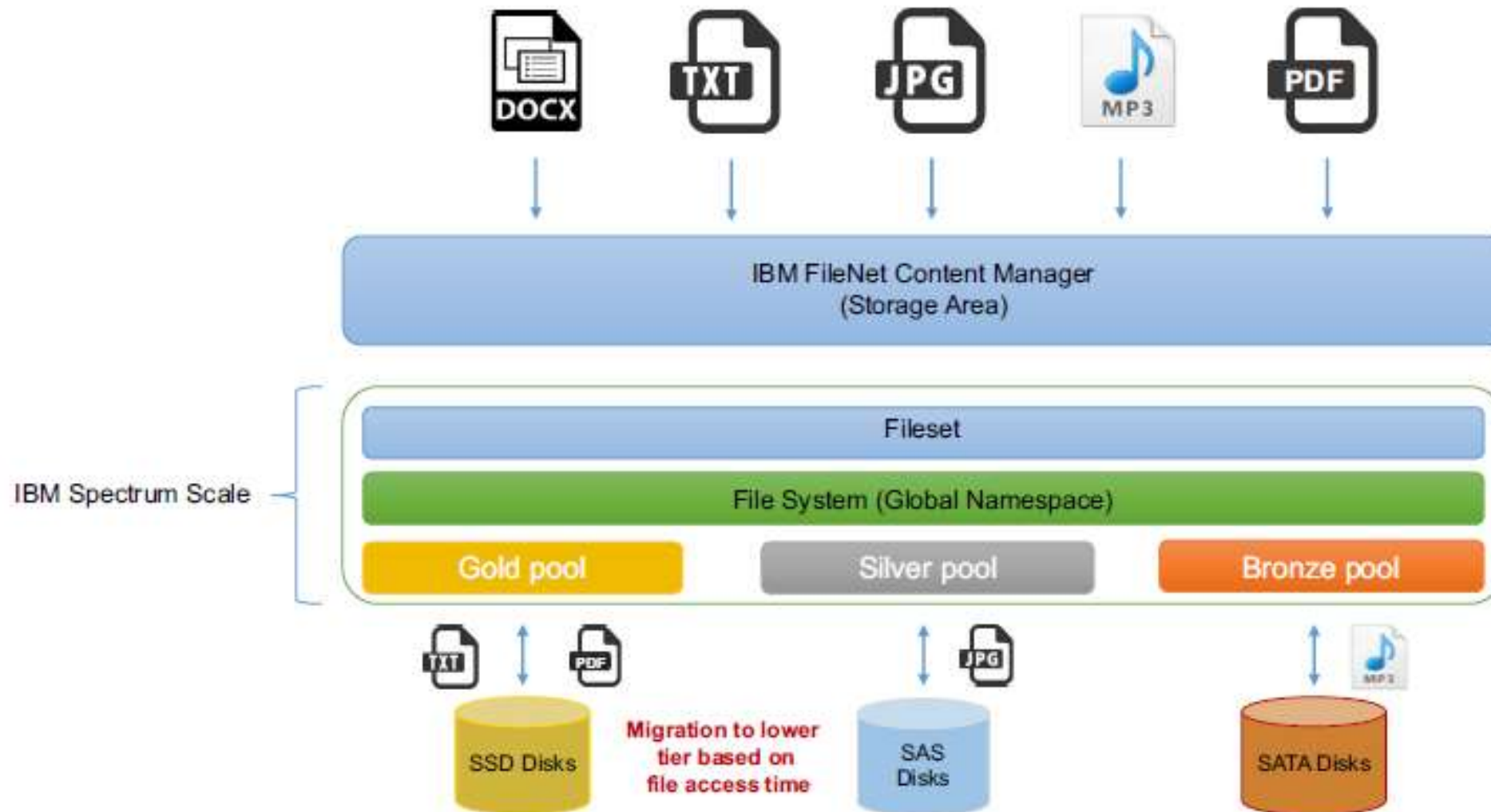- Spectrum Scale GUI running on Linux on Z nodes.
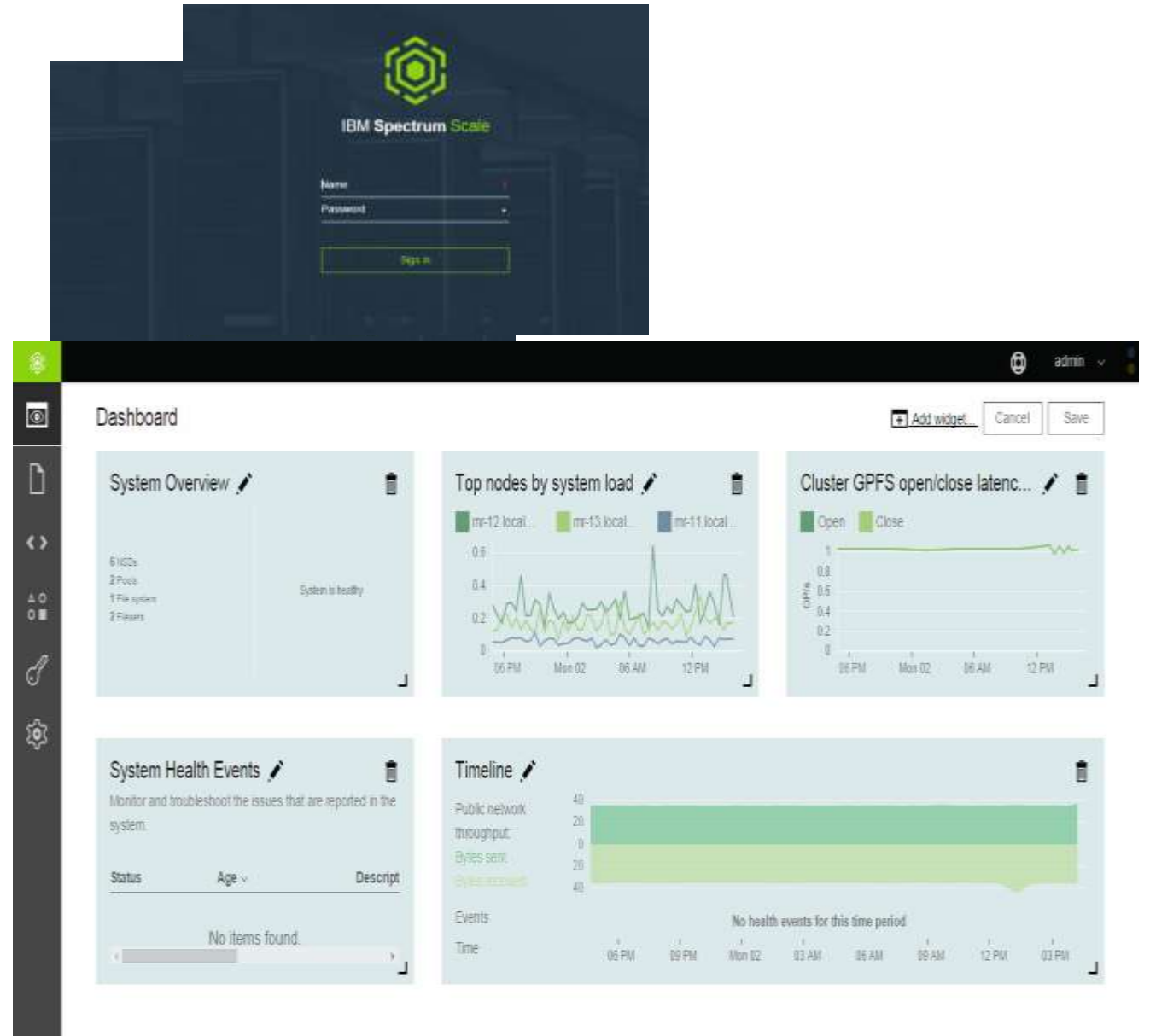
IBM Autonomic Data Storage solution

- Powerful policy engine
  - Information Lifecycle Management
  - Fast metadata 'scanning' and data movement
  - Automated data migration to based on threshold
- Users not affected by data migration
  - Single namespace

- Example:
  Online storage reaches 90% full then move all 1GB or larger files that are 60 days old to offline to free up space

- Integrated with Spectrum Archive

- Complies with NIST SP 800-131A and is FIPS 140-2 certified

- Supports HIPAA, Sarbanes-Oxley

Spectrum Scale FS

Linux on z Systems
IBM Spectrum Scale

Small files last accessed > 30 days

System pool (Flash)

Gold pool (SSD)

last accessed > 60days

accessed today and file size is <1G

Silver pool (NL SAS)

Silver pool is >60% full drain it to 20%

Tape Library in IBM managed cloud

secured

Tape Library

Spectrum Archive

Send it back to Silver pool when accessed
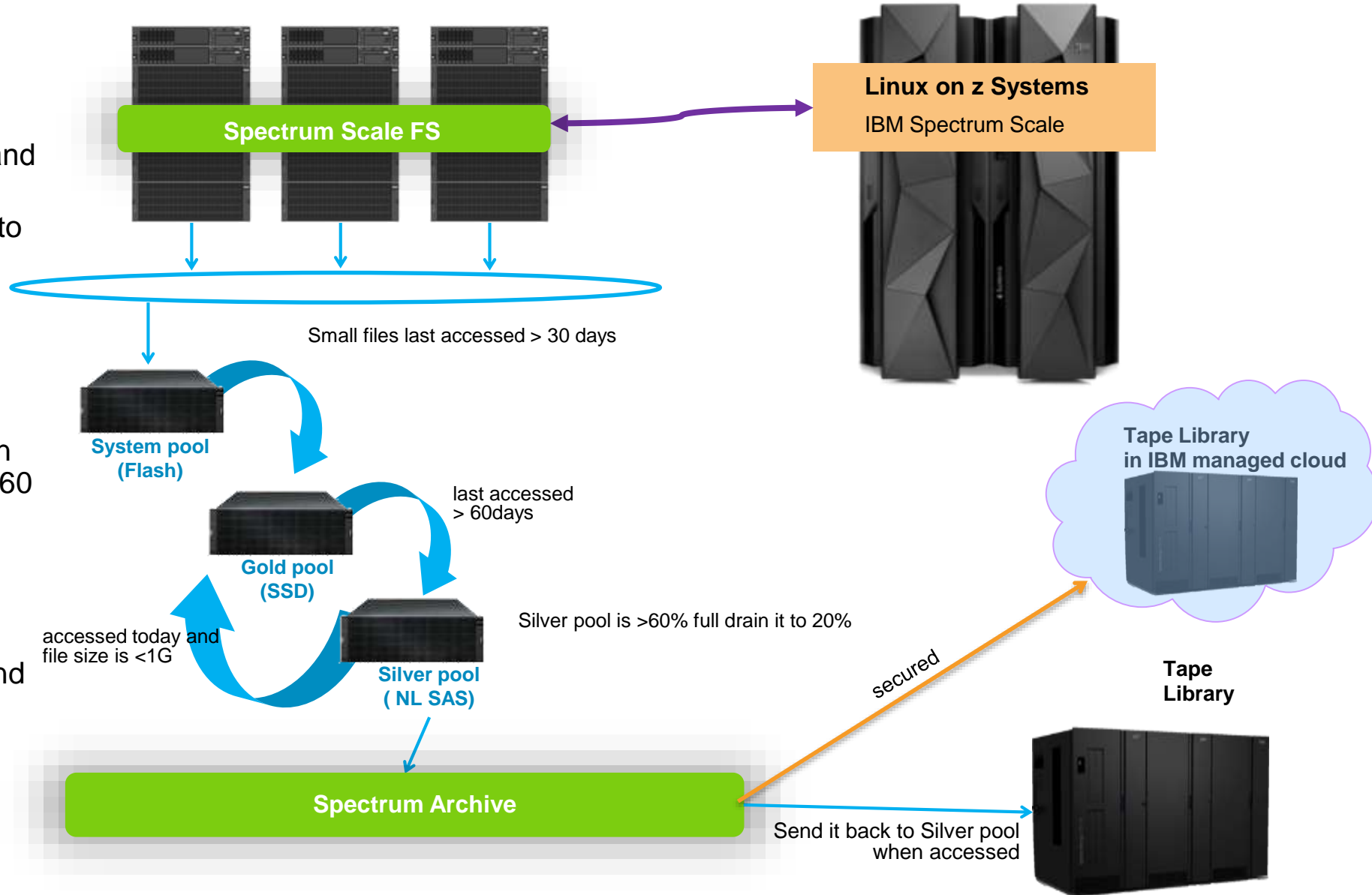
Effective exploitation for performance and storage flexibility IBM CMS

In-flight and on-rest Security and Privacy support – exploits IBM z14 capabilities
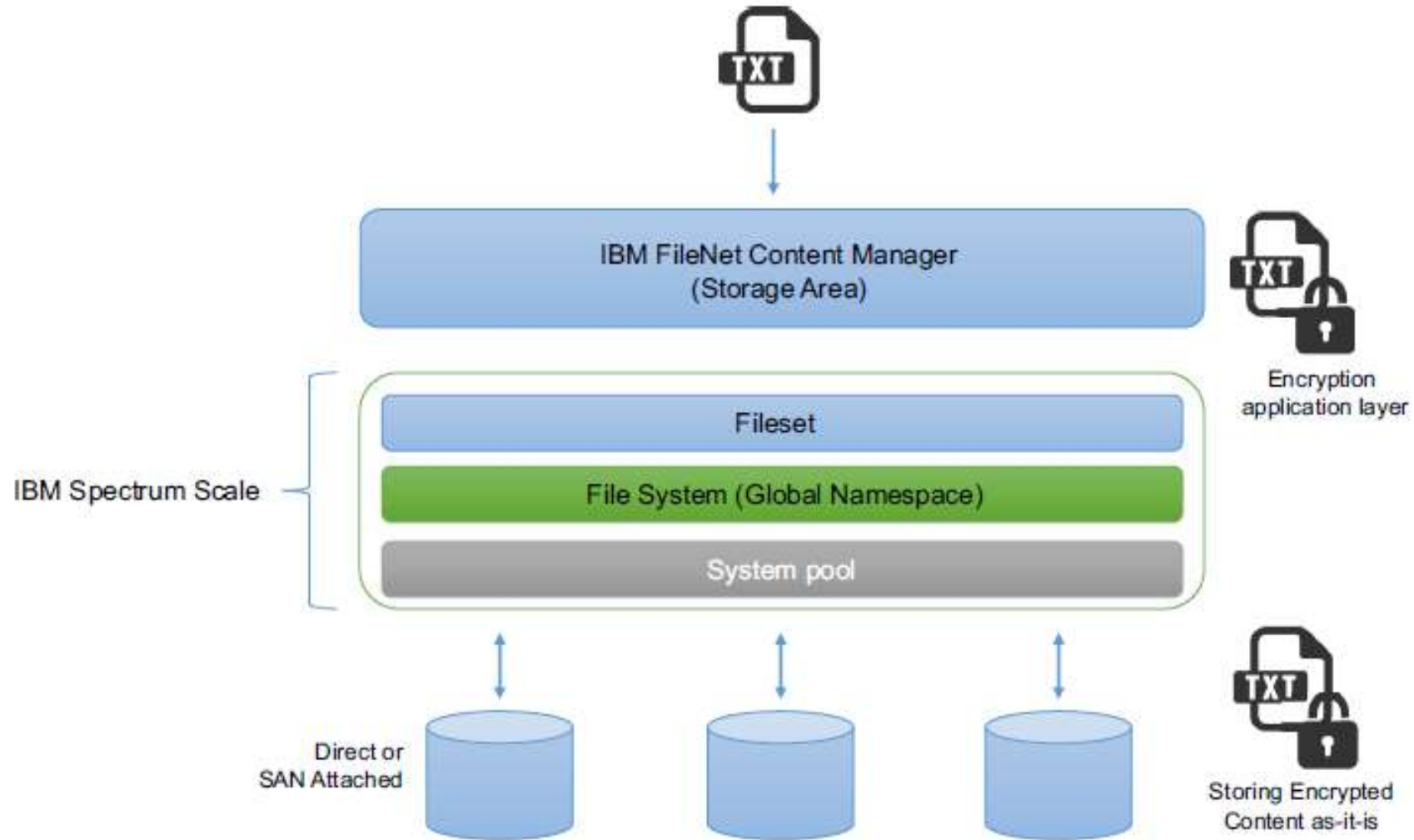
ECM multi nodes implementation with Spectrum Scale
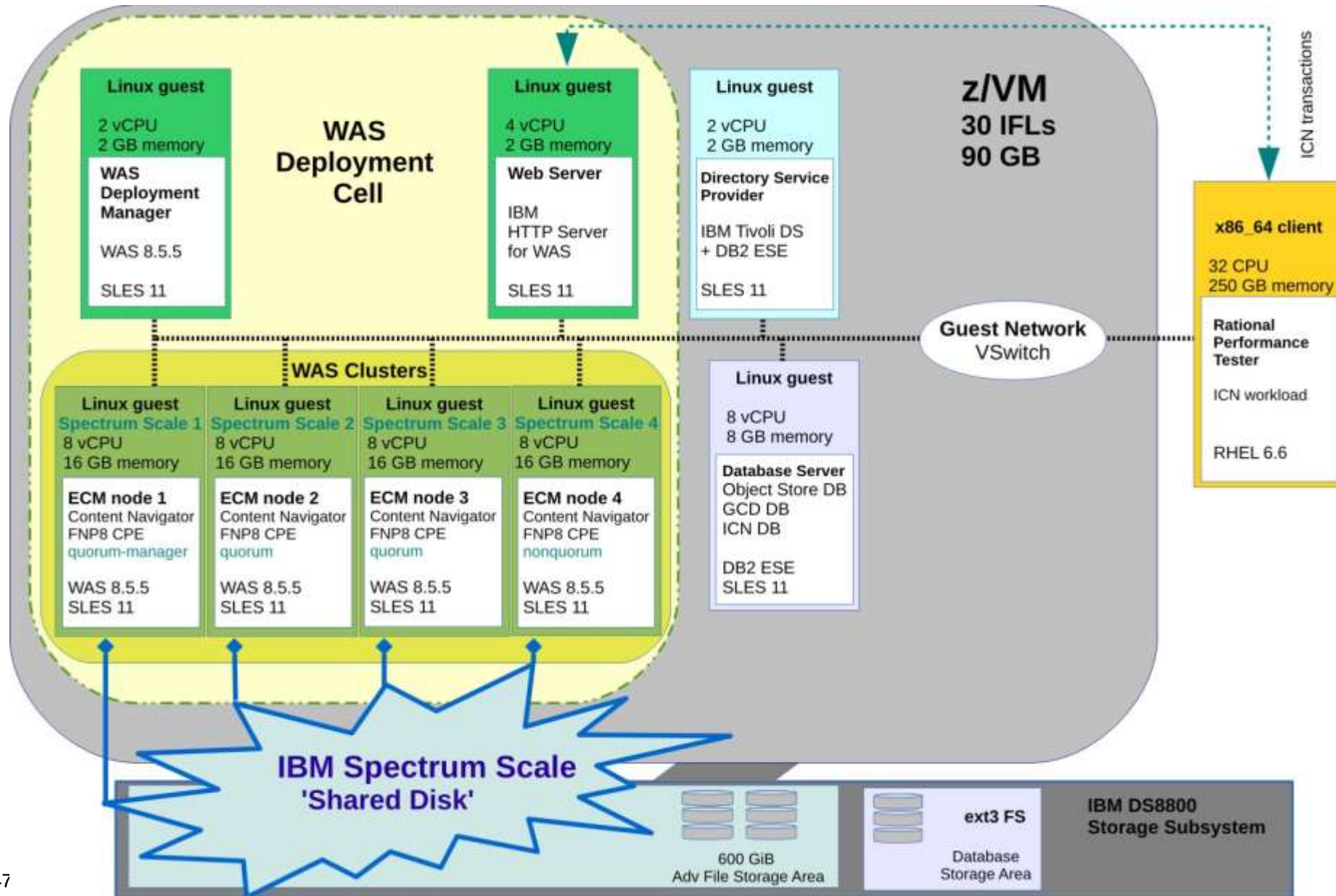
ECM user scaling – average transaction response time and throughput
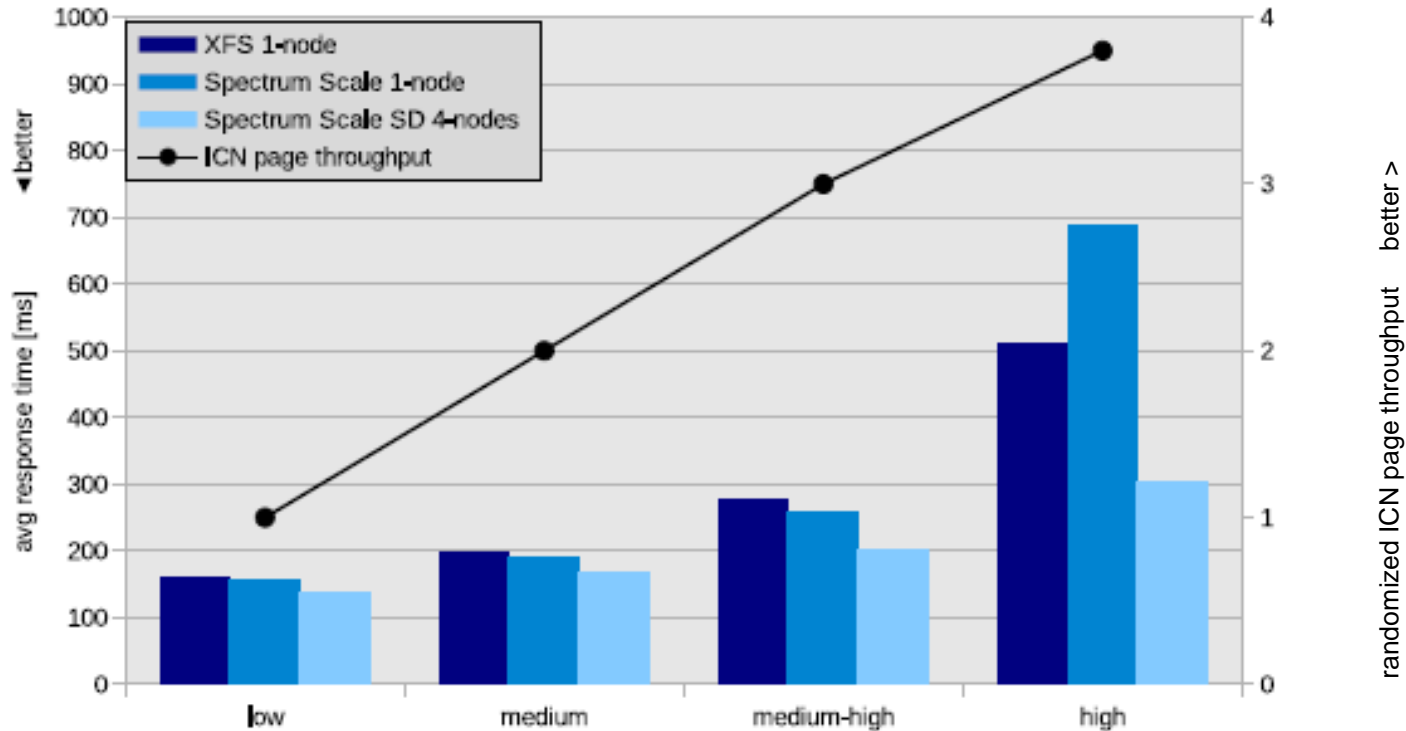
The throughput rate showed an almost linear characteristics correspondent to the increasing SUT load level (number of virtual ICN users), starting at a low load level (250 users) and increasing up to a high load level (1000 users).
The average ICN transaction response time started at 150 ms for the low load level, and increased up to 700 ms for the high load level for a single ECM node with Spectrum Scale. Up to the medium-high load level the setups with Spectrum Scale showed better response times than a single ECM node with XFS.

Spectrum Scale for IBM ECM with FileNet – by the numbers

Ingest	Retrieve	Scale
2M	**41K**	**5.78 B**
Documents added / day	Business Users	Objects stored

- 5M peak per day
- Triggers generation of links & custom objects
- 8M stored objects/day

- 19,000 unique logons/day
- 80+ consumer systems
- 350+ vendor & agency systems

- 2.75 B documents
- 6 repositories
- 12 physical servers
- 47 OS partitions

Larger US based financial services provider

Spectrum scale security

- Security of data in transit
- Security of data at rest
- Authentication
- Authorization
- Hadoop security
- Immutability
- Secure administration
- Audit logging
- Security for transparent cloud tiering (TCT)
- Security for OpenStack drivers

Redbooks
ibm.com/redbooks

IBM Spectrum Scale Security

Felipe Knop
Sandeep R. Patil
Larry Coyne
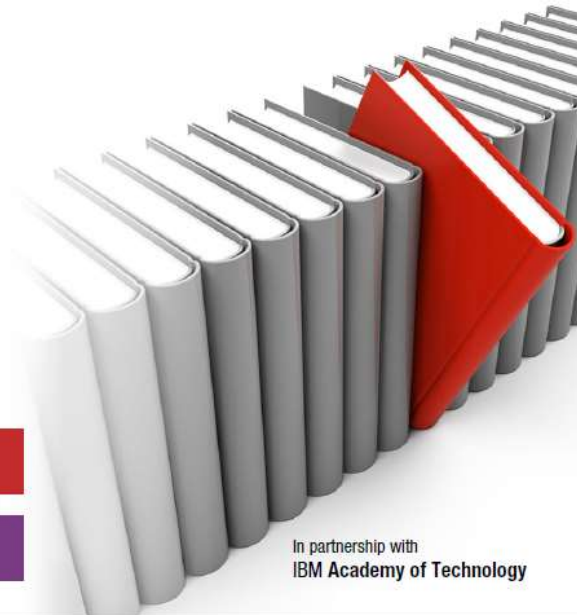
Cloud

Storage
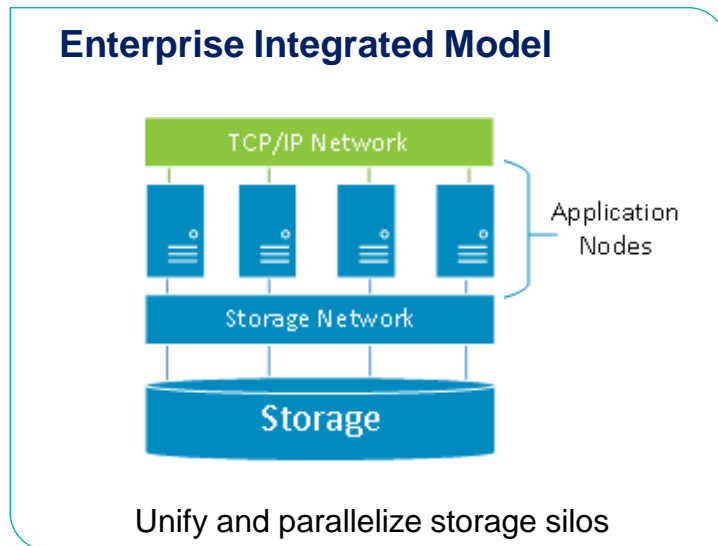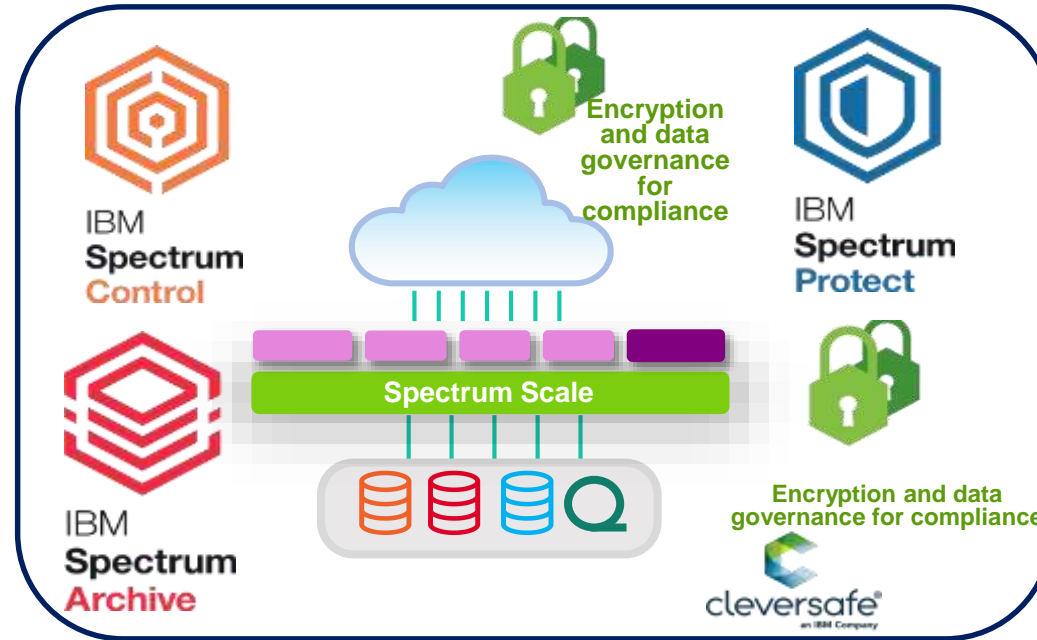
In partnership with
IBM Academy of Technology

IBM

Redpaper

Summary: The Autonomic Data Storage Hub and Ecosystem

- Automated data life cycle, from storing to backup and archive
- Store everywhere, secured, self tiered, compressed, clustered using GPFS
- Automated, policy-driven storage management, highly scalable, massive parallel access of Unified File, Object & HDFS
- Advanced GUI & health monitoring
- Complies with NIST SP 800-131A and is FIPS 140-2 certified
- Supports HIPAA, Sarbanes-Oxley

Encryption and data governance for compliance

IBM **Spectrum Control**

IBM **Spectrum Protect**

IBM **Spectrum Archive**

Spectrum Scale

Encryption and data governance for compliance

cleversafe
an IBM Company

Enterprise Integrated Model

TCP/IP Network

Application Nodes

Storage Network

Storage

Unify and parallelize storage silos

Storage HA Model

Remote Site

Client switches to secondary on failure

Application

Push all updates asynchronously

Whichever is fastest

Primary Cluster

Secondary Cluster

Synchronous replication

Network Shared Disk (NSD) Model

Application Nodes

TCP/IP or Infinband Network

NSD Servers

Storage   Storage

Modular High-Performance Scaling

Enterprise Content Hub additional information

Implementation details and tips:

➢ **IBM Enterprise Content Management for Linux on z Systems Scale-Out Case Study (Part 1)**
https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=ZSW03340USEN&

➢ **IBM Enterprise Content Management for Linux on z Systems Scale-Out Case Study (Part 2):**
https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=ZSW03345USEN&
http://public.dhe.ibm.com/software/dw/linux390/perf/ZSW03345USEN.pdf

➢ IBM Spectrum Scale and ECM FileNet Content Manager, a Winning Combination
  - Deployment Variations and Value-added Features
http://www.redbooks.ibm.com/redpapers/pdfs/redp5239.pdf

Questions?

Wilhelm Mild

IBM Executive IT Architect

The Open Group
OPen Certified
Distinguished Architect

IT Architecture
Chief/Lead IT Architect

IBM

IBM Deutschland Research
& Development GmbH
Schönaicher Strasse 220
71032 Böblingen, Germany

Office: +49 (0)7031-16-3796
wilhelm.mild@de.ibm.com

Notices and disclaimers

Notices and disclaimers continued

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

.

Please note

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM's sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.