IBM Spectrum Scale for Linux on IBM z Systems

*– Introduction to **Standard Edition v4.1.1***

Session ID:17777

Wilhelm Mild
IBM Executive IT Architect
IBM Laboratory Germany
wilhelm.mild@de.ibm.com

IBM Spectrum Storage portfolio

- IBM Spectrum Scale™ is industrial strength, highly scalable software defined storage that enables global shared access to data with extreme scalability and agility for cloud and analytics
- IBM Spectrum Accelerate™ offers grid-scale block storage with rapid deployment IBM Spectrum Scale for Linux on z Systems
- IBM Spectrum Virtualize™ software is at the heart of IBM SAN Volume Controller and IBM Storwize family . It enables these systems to deliver industry-leading virtualization that enhances storage to improve resource utilization and productivity
- IBM Spectrum Control™ provides efficient infrastructure management for virtualized, cloud and software-defined storage to simplify and automate storage provisioning
- IBM Spectrum Protect™ enables reliable, efficient data protection and resiliency for software defined, virtual, physical and cloud environments.
- IBM Spectrum Archive™ enables you to automatically move infrequently accessed data from disk to tape

Introduction to Spectrum Scale:
http://public.dhe.ibm.com/common/ssi/ecm/dc/en/dcw03057usen/DCW03057USEN.PDF

IBM Spectrum Scale (former IBM Global Parallel File System - GPFS)

What is it and how is it delivered:

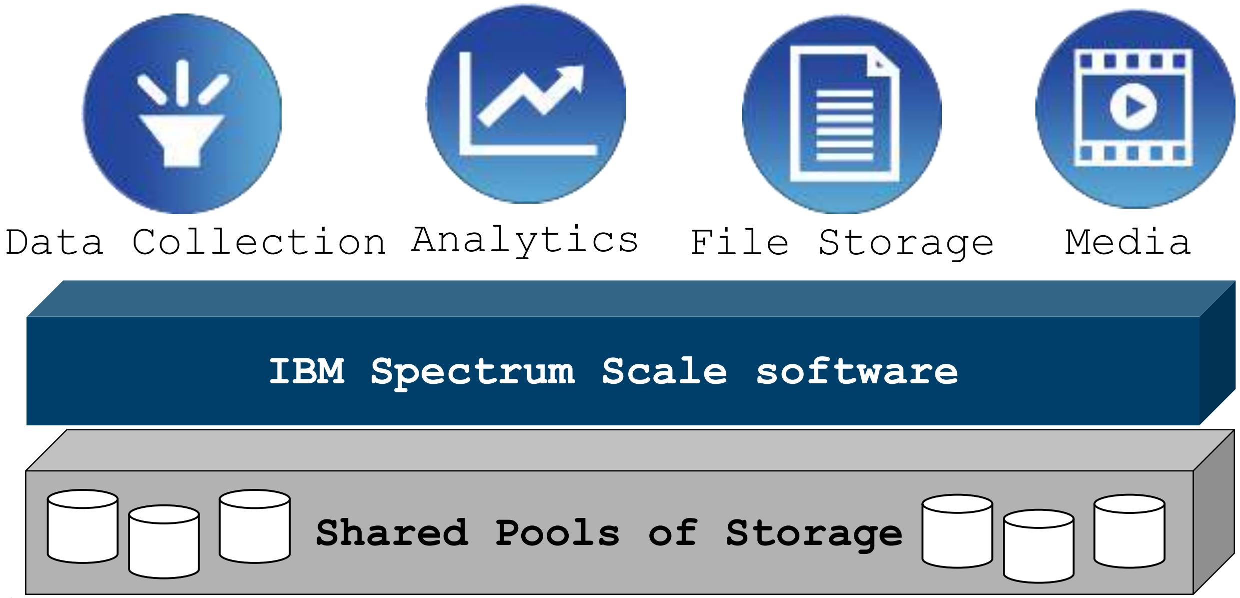
As a Software-only solution: runs on many hardware platforms and supports almost any block storage device.

As an integrated IBM Elastic Storage Server solution: IBM Elastic Storage Server is an optimized storage solution bundled with IBM hardware and Spectrum Scale software.

As a Cloud service: IBM Spectrum Scale delivered as a service, bringing high performance, scalable storage and integrated data governance.

When to use:

For fast data access and simple, cost effective data management



Data Collection Analytics File Storage Media

IBM Spectrum Scale software

Shared Pools of Storage

A clustered file system with:

- high-performance
- highly scalable
- high availability
- parallel file access read and write

Clustered and Distributed File Systems

Clustered file systems

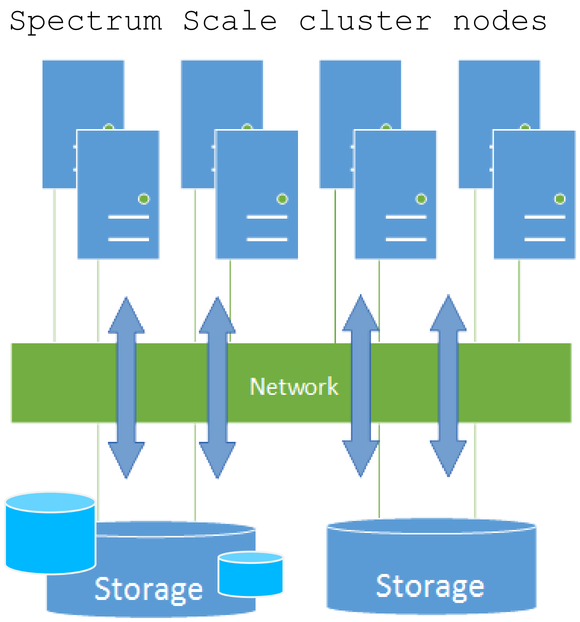
- **Shared File system being simultaneously mounted on multiple servers accessing the same storage read/write**
- Clustered file systems can provide features like location-independent addressing and redundancy which improve reliability or reduce the complexity of the other parts of the cluster.
- Examples: IBM Spectrum Scale (formerly IBM GPFS™), Oracle Cluster File System (OCFS2), Global File System (GFS2)

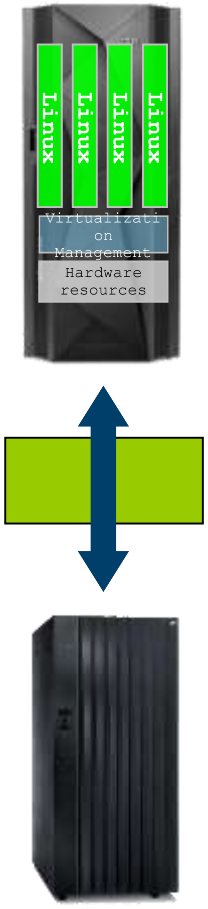
Distributed shared file systems

- **File system is accessed through a network protocol and do not share block level access to the same storage**
- When a user accesses a file on the server, the server sends a copy of the file, which is cached on the user's computer while the data is being processed and is then returned to the server.
- Examples: NFS, OpenAFS, CIFS

IBM Spectrum Scale cluster overview

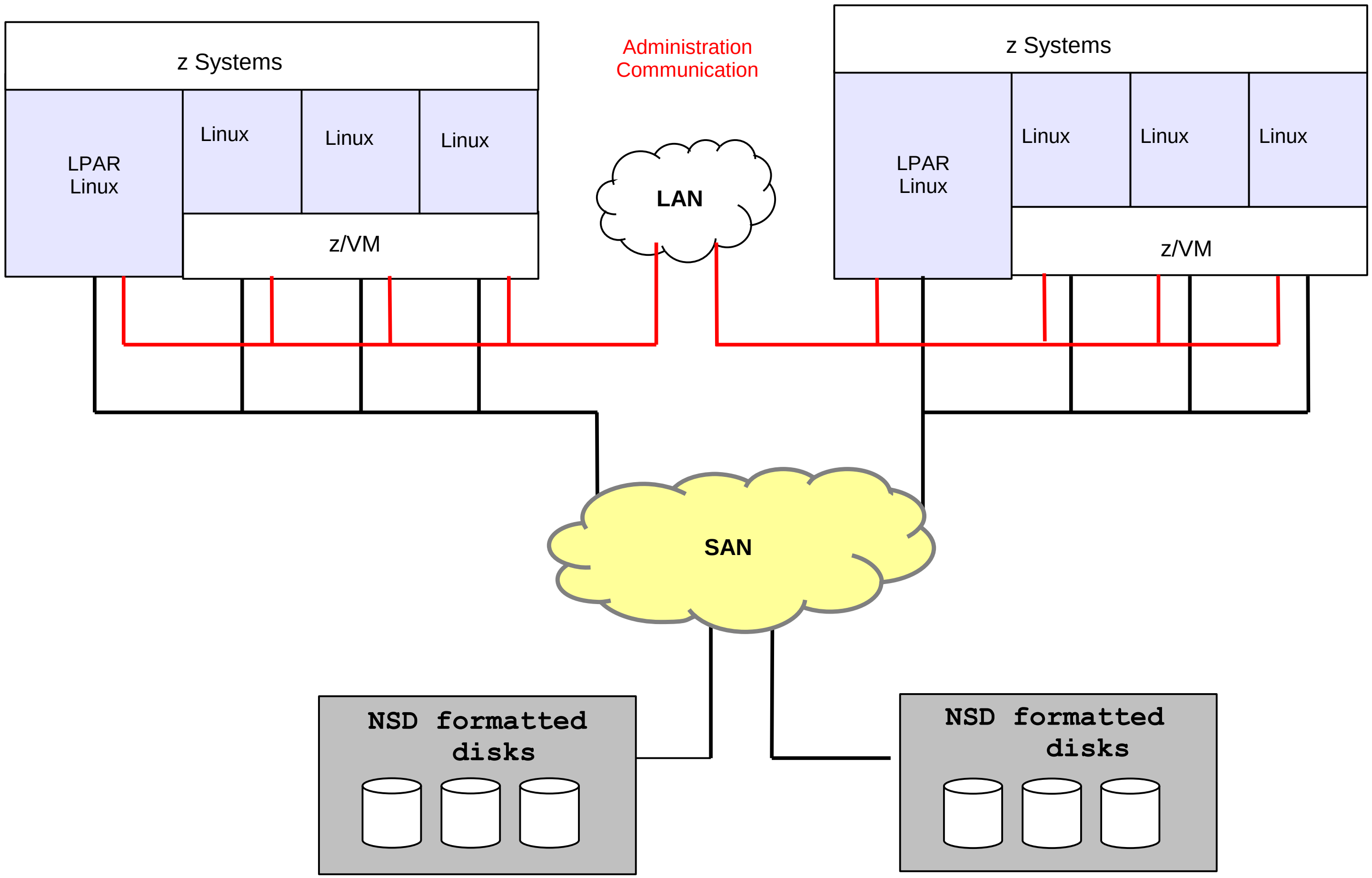
- IBM’s shared disk, parallel cluster file system
- Cluster: 1 to 16,384* nodes, fast reliable communication, common admin domain
- Shared disk: all data and metadata on storage devices accessible from any node through block I/O interface (“disk”: any kind of block storage device)
- Parallel: data and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



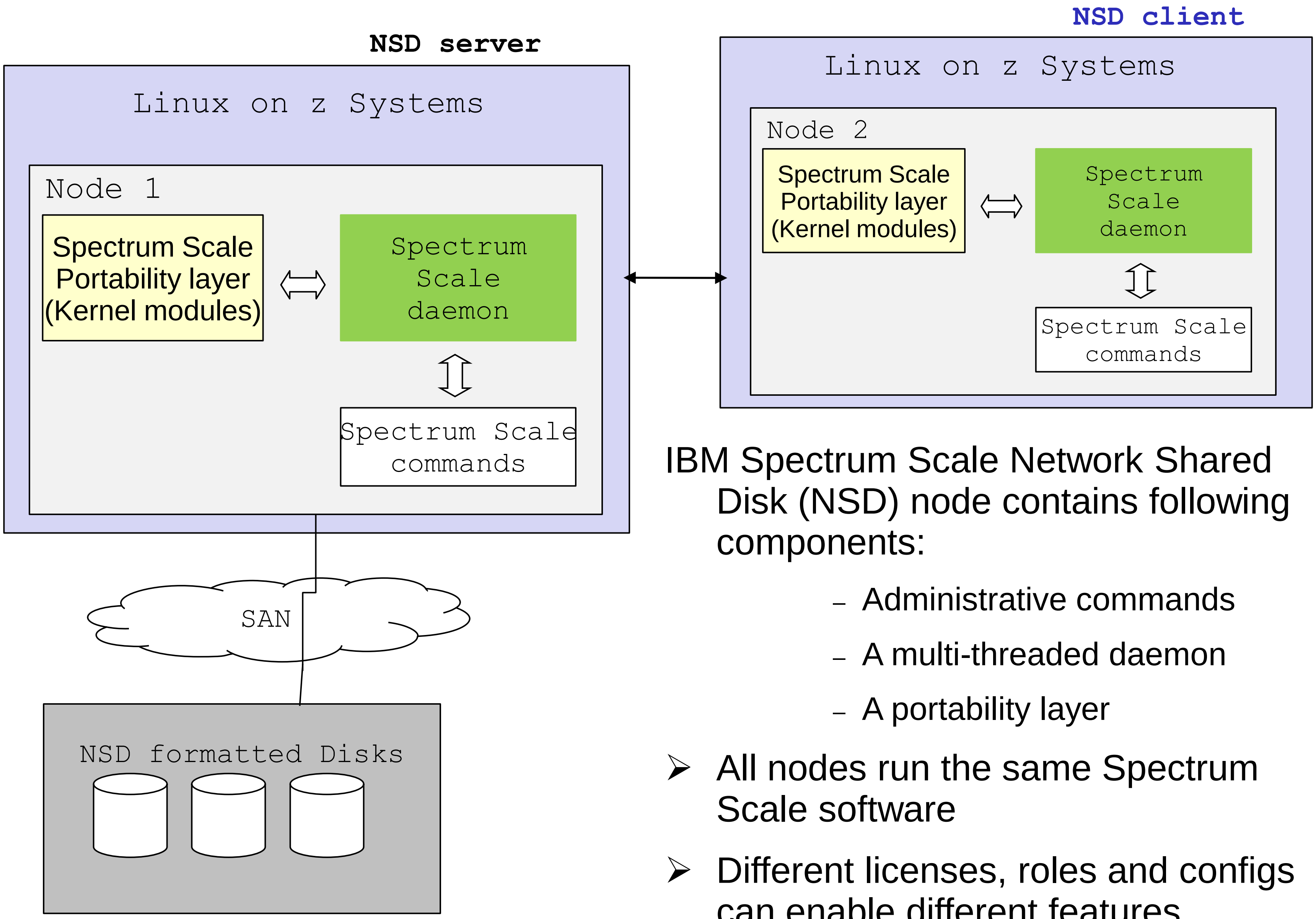


*largest cluster in production as of August 2014
Is LRZ SuperMUC 9400 Nodes of x86_64

IBM Spectrum Scale Nodes overview
- HA scenario with Linux on z Systems



IBM Spectrum Scale Node architecture overview



IBM Spectrum Scale Network Shared Disk (NSD) node contains following components:

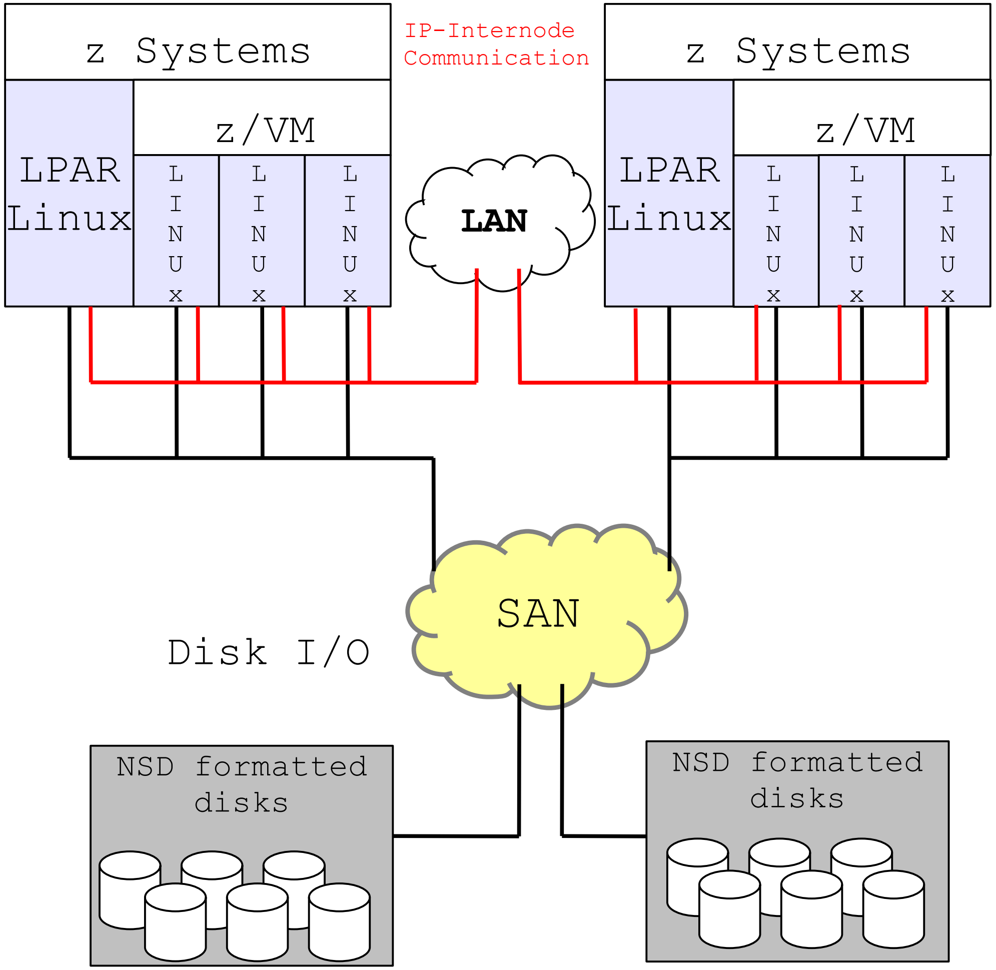
- Administrative commands
- A multi-threaded daemon
- A portability layer

➢ All nodes run the same Spectrum Scale software

➢ Different licenses, roles and configs can enable different features

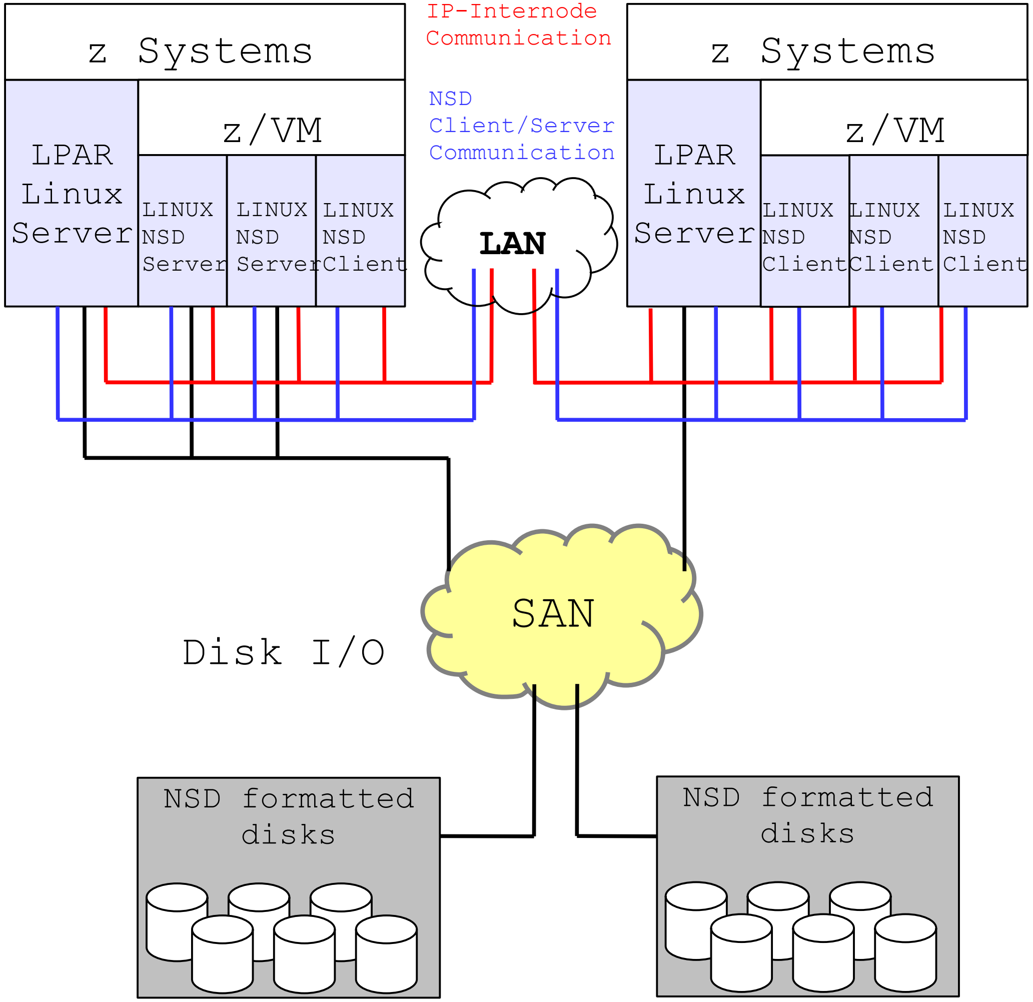
IBM Spectrum Scale - Shared Disk (SAN) Model

- Every cluster node has direct access to the storage disks through the SAN
- Internode communication (control /administration) is done via IP network connections
- Spectrum Scale does not send disk data from one cluster node to another using the network (just meta data)
- This architectural model can achieve the best performance (better than NSD client/server model)



Network Shared Disk (NSD) Client/Server Model

- Only a couple of cluster nodes (NSD Server) 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
- 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



IBM Spectrum Scale Standard Edition v4.1.1 for Linux on z Systems

- “Stretched cluster” with synchronous mirroring utilizing Spectrum Scale block-level replication with less than 40 km
- Heterogeneous clusters with client nodes without local storage access running on AIX®, Linux on Power® and Linux on x86
- Information Lifecycle Management (ILM)
- Active file management (AFM) for active/active configurations
- Up to 128 cluster nodes with same or mixed Linux distributions/releases
- Support for Linux instances in LPAR and as z/VM® guest, on the same or different z Systems server
  - IBM Spectrum Scale has no dependency on a specific version of z/VM
- Well suited and supported workloads are WebSphere Application Server (WAS), MQ, or similar workload infrastructure environments, Websphere based OLTP workloads, FileNet®P8/ECM5.2.1.

Announcement 05/2015:
http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=ca&infotype=an&appname=iSource&supplier=897&letternum=ENUS215-147

“Stretched” cluster

Spectrum Scale Standard Edition V4.1.1

Introduction to “stretched” cluster

- 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
- A stretched cluster ensures data availability in the event of a total failure of the primary (production) site
- A disaster-resilient IBM Spectrum Scale cluster is made up of three, distinct, geographically-separate hardware sites operating in a coordinated fashion.
  - Two of the sites consist of Spectrum Scale nodes and storage resources holding a complete replica of the file system.
  - The third site consists of a single node and a single disk used as a tiebreaker for GPFS quorum.
- The data is synchronously mirrored from one site to the other

Cluster node roles

In general, IBM Spectrum Scale performs the same functions on all nodes. It handles application requests on the node where the application exists.

There are two important cases where one node provides a management function affecting the operation of multiple nodes. These are nodes acting as:

- The cluster manager
  - There is one cluster manager per cluster. The cluster manager is chosen through an election held among the set of quorum nodes designated for the cluster.
  - Role set: `quorum - nonquorum`
- The file system manager
  - There is one file system manager per file system, which handles all of the nodes using the file system.
  - Role set: `manager - client`

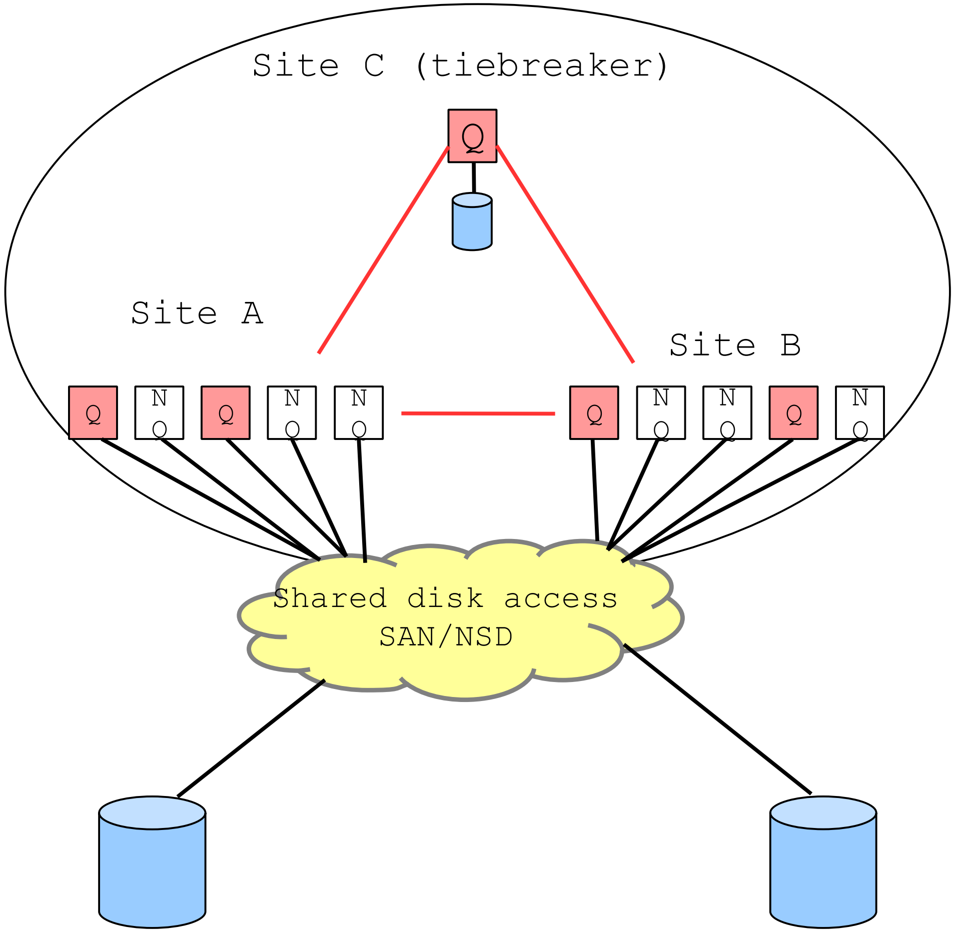
A cluster node can have various roles from different role sets

Many of the roles are dynamically assigned to a node

Depending on the role, different licenses are required.

Synchronous mirroring utilizing Spectrum Scale block-level replication

- One geographically dispersed cluster (“stretched” cluster)
  - All nodes in either site have SAN/NSD access to the disk
  - Site A storage is duplicated in site B with Spectrum Scale replication
  - Simple recovery actions in case of site failure (more involved if you lose tiebreaker site as well)
- Performance implication: Spectrum Scale has no knowledge of a replica's physical locality. There is no way to specify disk access priority (i.e. local storage first)



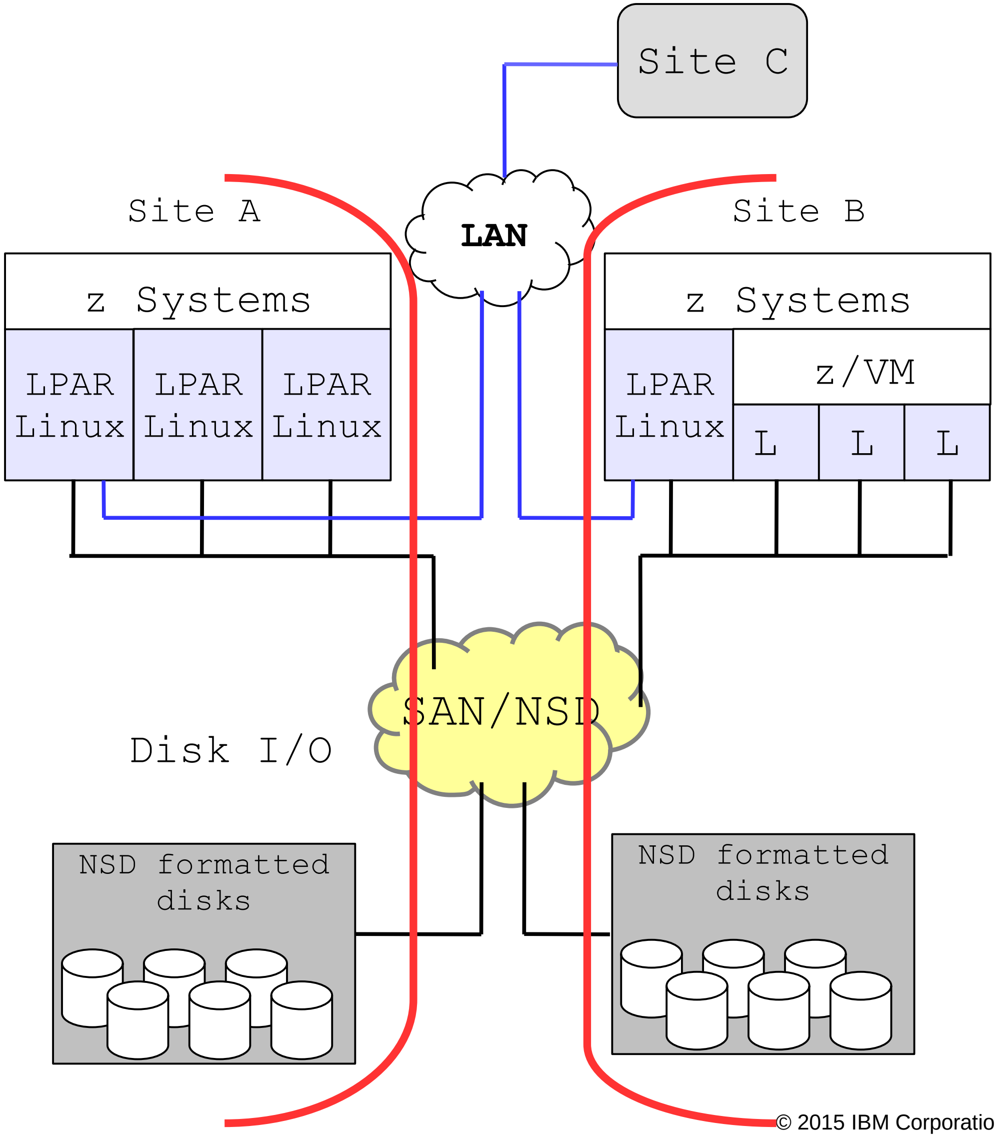
Legend:
Q: quorum node
NQ: non-quorum node

Spectrum Scale block-level 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.

“Stretched” Cluster (Synchronous Mirroring w/ Data Replication)

- A single cluster defined across multiple geographic sites
- It ensures data availability in the event of a total failure of the primary (production) site
- Two of the sites consist of cluster nodes and storage (A+B)
- This data is synchronously mirrored from one site to the other
- A third site (C) is used as tiebreaker site
- Supported:
  - up to 40 km distance
  - synchronous mirroring with Spectrum Scale data replication



Cluster Considerations

Cluster considerations / variations

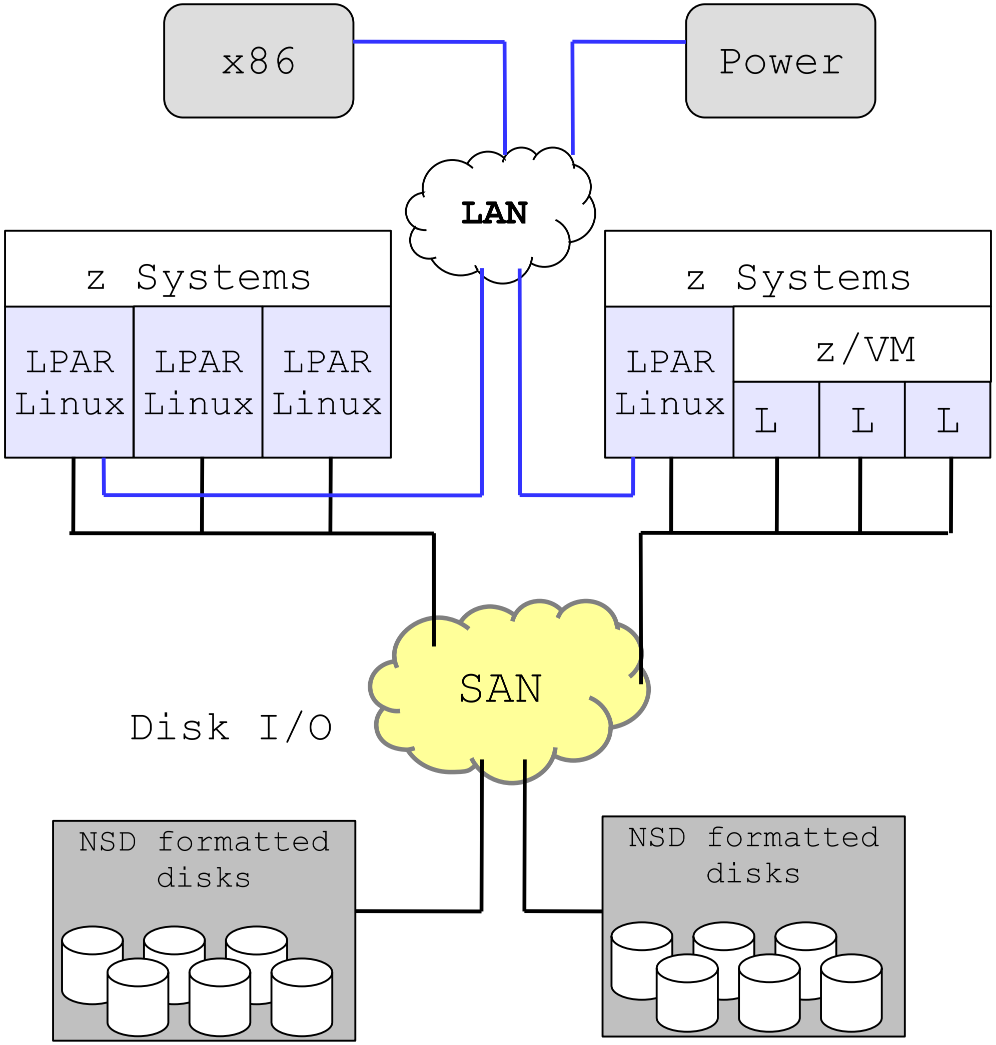
IBM Spectrum Scale for Linux on z Systems supports the following cluster configurations:

- Shared-disk model: single operating system images can access a set of disks directly
- Network Shared Disk (NSD) client/server model: the shared-disk model is extended by mixing direct SAN access with network-attached cluster nodes
- “Stretched” cluster: A single cluster is spread across multiple sites
- Heterogeneous cluster (various platforms)

Heterogeneous cluster across Architectures

Heterogeneous clusters are can include:

- X86 servers, Power servers running RHEL, SLES or AIX
- Must be defined as NSD clients
  - no local storage access from distributed servers
- Do not share storage among different platforms



IBM Spectrum Scale Standard Edition v4.1.1 for Linux on z Systems

- “Stretched cluster” with synchronous mirroring utilizing Spectrum Scale block-level replication with less than 40 km
- Heterogeneous clusters with client nodes without local storage access running on AIX®, Linux on Power® and Linux on x86
- Information Lifecycle Management (ILM)
- Active file management (AFM) for active/active configurations
- Up to 128 cluster nodes with same or mixed Linux distributions/releases
- Support for Linux instances in LPAR and as z/VM® guest, on the same or different z Systems server
  - IBM Spectrum Scale has no dependency on a specific version of z/VM
- In addition to WebSphere®Application Server (WAS), MQ, or similar workload infrastructure environments, Websphere based OLTP workloads, FileNet®P8/ECM5.2.1 supported.

- The Express Edition including base functions of the file system is supported with v4.1.1

Information Lifecycle Management (ILM)

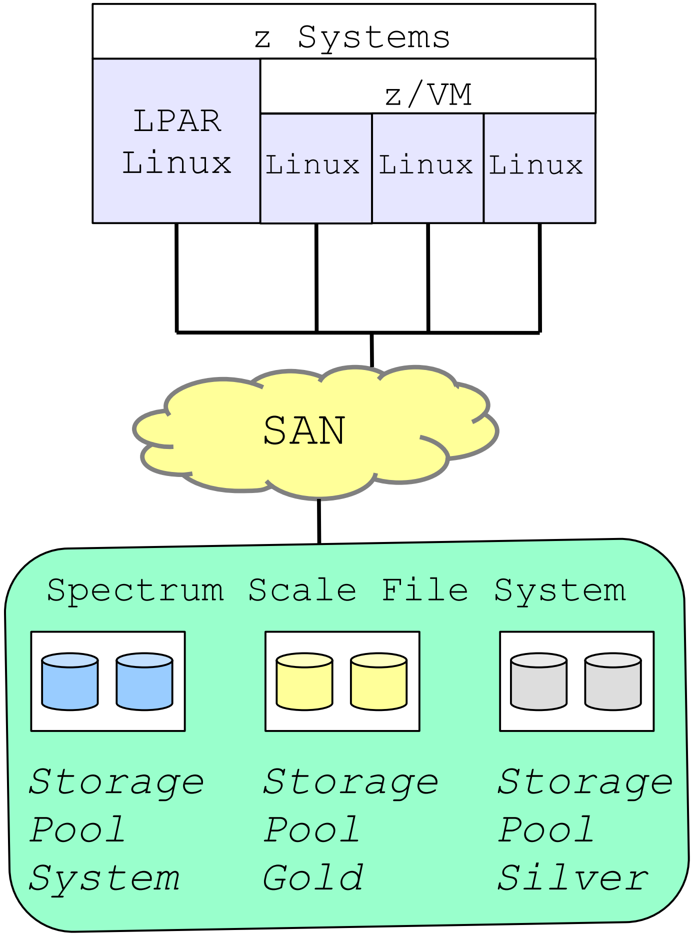
Spectrum Scale Standard Edition V4.1.1

Introduction to Information Lifecycle Management (ILM)

- The goal of ILM is to optimize costs 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
- IBM Spectrum Scale can help you achieve Information Lifecycle Management (ILM) efficiencies through powerful policy-driven automated tiered storage management.
  - The ILM toolkit helps you to manage sets of files and pools of storage, and also enables you to automate the management of file data.
- Using these tools, IBM Spectrum Scale can automatically determine where to physically store your data regardless of its placement in the logical directory structure.

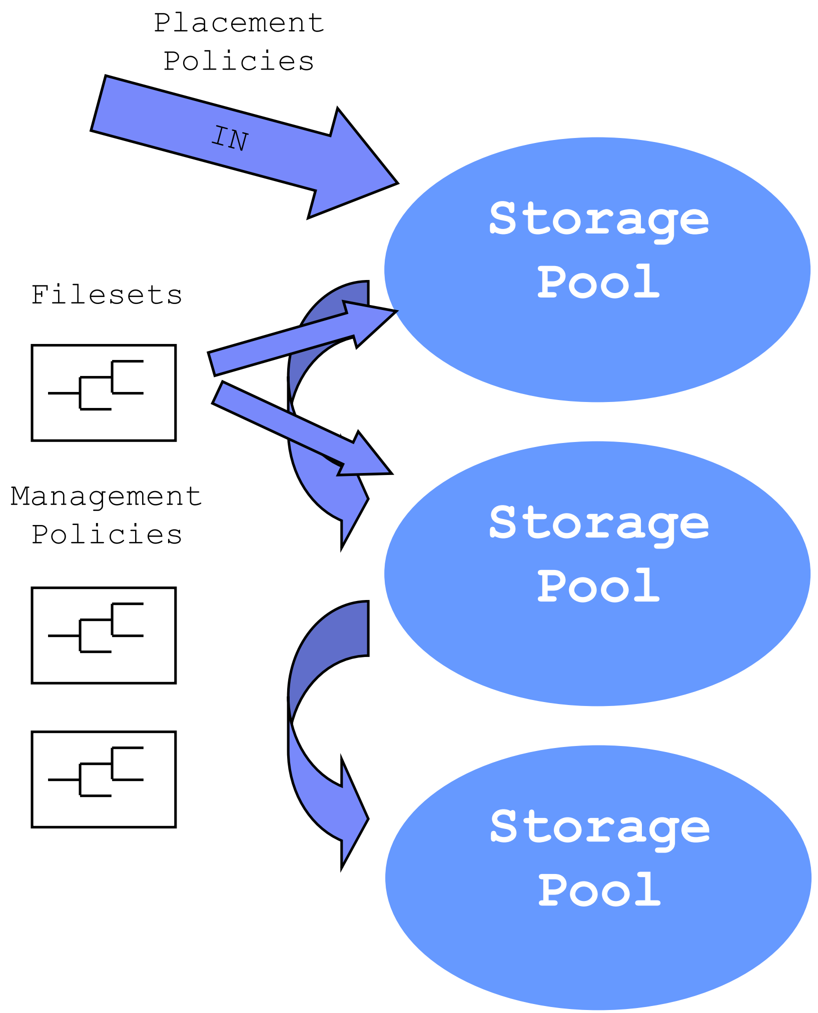
Storage Pools

- 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



ILM Tools

- Storage pools
  - Storage pools provide grouping of storage that are managed together
- File placement policies
  - File placement policies assign data to pools on file creation
- File management policies
  - File management policies automate migration/ deletion/replication/ reporting
- Filesets
  - Filesets (named subdirectories) allow you to organize data

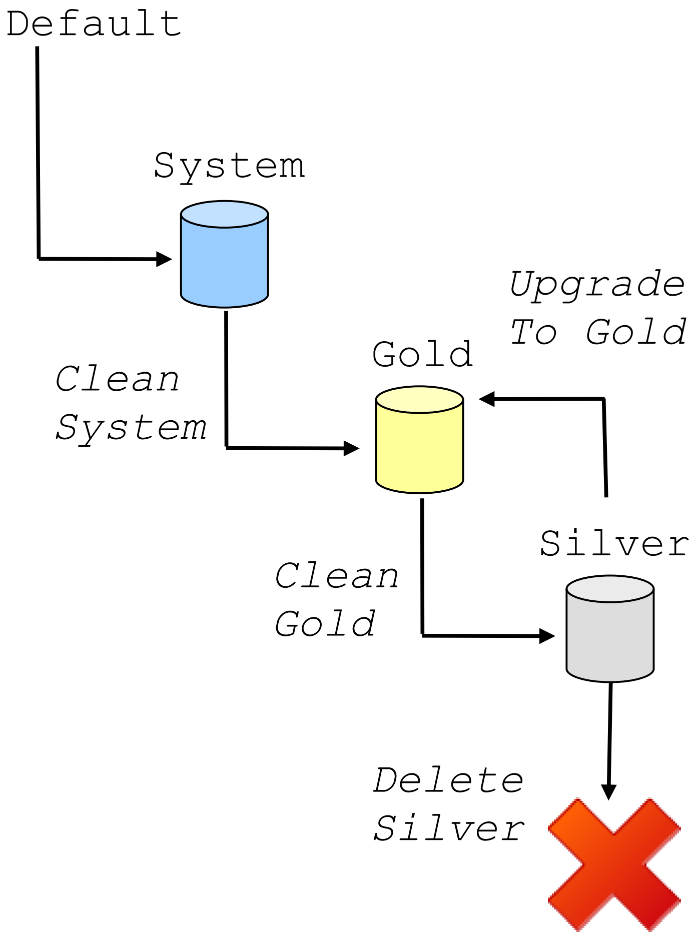


Policies and rules

- Policy: A set of user-specified rules that match data to the appropriate pool
  - SQL-like syntax for selecting files based on file attributes

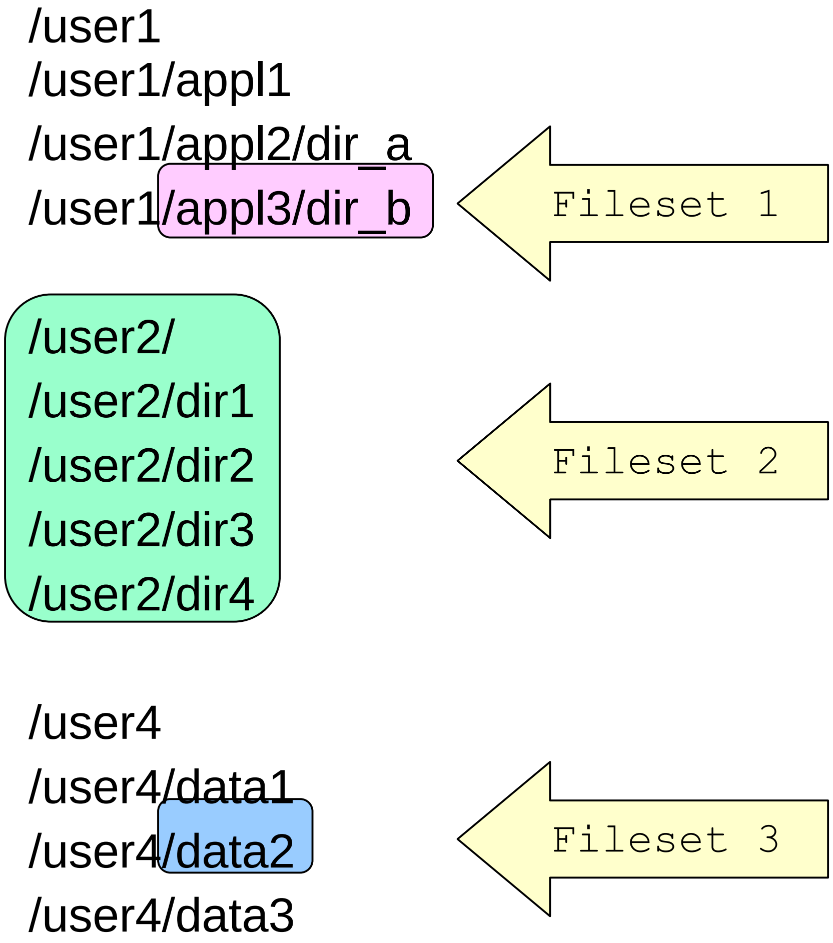
```
RULE 'Clean System' MIGRATE FROM
   POOL 'System' THRESHOLD(85,40)
   WEIGHT(KB_ALLOCATED)
   TO POOL 'Gold'
```

- 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



Filesets categorization

- A fileset is a sub-tree of a file system
  - That provides a means of partitioning the file system
  - That allows you 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
/user2/dir3
/user2/dir4

Fileset 2

/user4
/user4/data1
/user4/data2
/user4/data3

Fileset 3

Snapshots

Snapshots are a logical, read-only copy of the file system; changes can only be made to the active files and directories .

- Snapshots can be created at file system, fileset and file level
  - Each file system can have multiple snapshots of any of the types at the same time.
- Snapshots are very fast and space efficient
  - A file in a snapshot does not occupy disk space until the file is modified or deleted.
- Snapshots typically used
  - To run a file system backup on a consistent state of the file system
  - On-line access to previous file system state
  - Protect data from user errors

Active File Management (AFM)

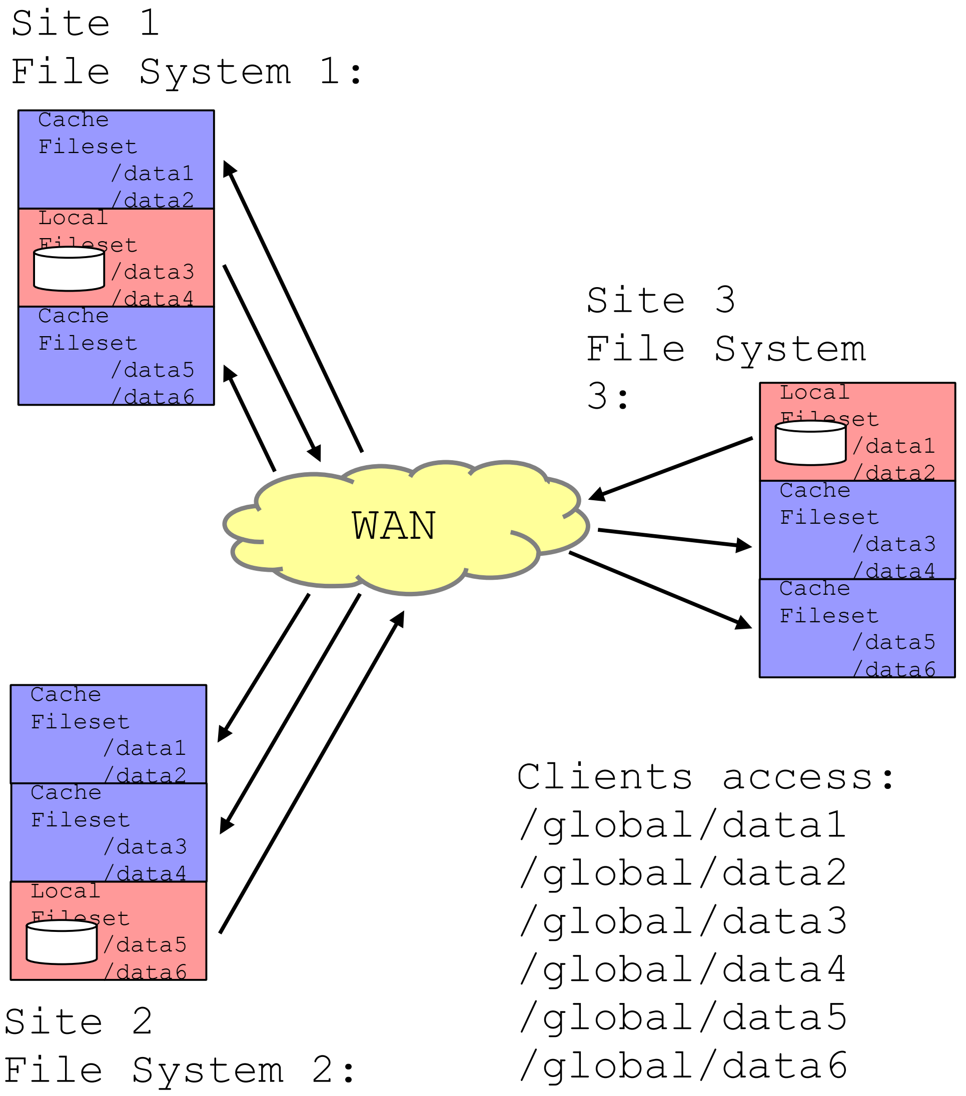
Spectrum Scale Standard Edition V4.1.1

Introduction to Active File Management (AFM)

- The goal of AFM is to share data across unreliable and high latency networks and across different location.
  - For example, it can also be used for duplicating data to a remote location for disaster recovery purposes without suffering from wide area network (WAN) latencies.
- AFM allows you to create associations from a local IBM Spectrum Scale cluster to a remote cluster
- With AFM you can define the location and flow of file data to automate the management of the data.
- This allows you to implement a single namespace view across sites around the world.

Global Namespace and AFM cache

- Relationships between clusters using AFM are defined at the fileset level.
- A fileset in a file system can be created as a “cache” that provides a view to a file system in another Spectrum Scale cluster called the “home”.
- File data is moved into a cache fileset on demand.
- A file system can contain multiple homes, caches and non-cached data
- Home and cache entities can be combined to create a global namespace
- See all data from any cluster


Site 1
File System 1:
Cache Fileset /data1 /data2
Local Fileset /data3 /data4
Cache Fileset /data5 /data6
WAN
Site 3
File System 3:
Local Fileset /data1 /data2
Cache Fileset /data3 /data4
Cache Fileset /data5 /data6
Cache Fileset /data1 /data2
Cache Fileset /data3 /data4
Local Fileset /data5 /data6
Site 2
File System 2:

Clients access:
/global/data1
/global/data2
/global/data3
/global/data4
/global/data5
/global/data6

AFM concepts

- Communication is done using NFSv3
- NFS server has to run at the 'home' cluster and the export configuration (/etc/exports) has to be updated to export the path of the fileset

- Whole file is fetched over NFS and stored on disk in a Spectrum Scale file system
- More than one copy of a file is available → home site and every cache site
- Data read is done in parallel across multiple nodes
- Application can continue after required data is in cache while the remaining file is being fetched

- Data written to the cache is copied back to home as quickly as possible → asynchronous writeback
- Writeback coalesce updates and accommodate out-of-order and parallel writes

AFM concepts

- In case of a disconnection between 'home' and 'cache' cluster,
  - AFM filesets on the 'cache' site continue to function independent of the 'home' site
  - Data access to cached data will fetch local data provided that the file is already cached
  - Writes can continue when the WAN is unavailable
- As soon as the connection is available again the data is written back to the home site
  - Conflict resolution: 'the last writer wins'

- Data is managed like a cache but stored on disk in a GPFS file system.
- Once retrieved, data can be accessed at local cluster speeds
- Duration of data in a cache (cache eviction) is dependent on the configuration and size of the cache (vs home)

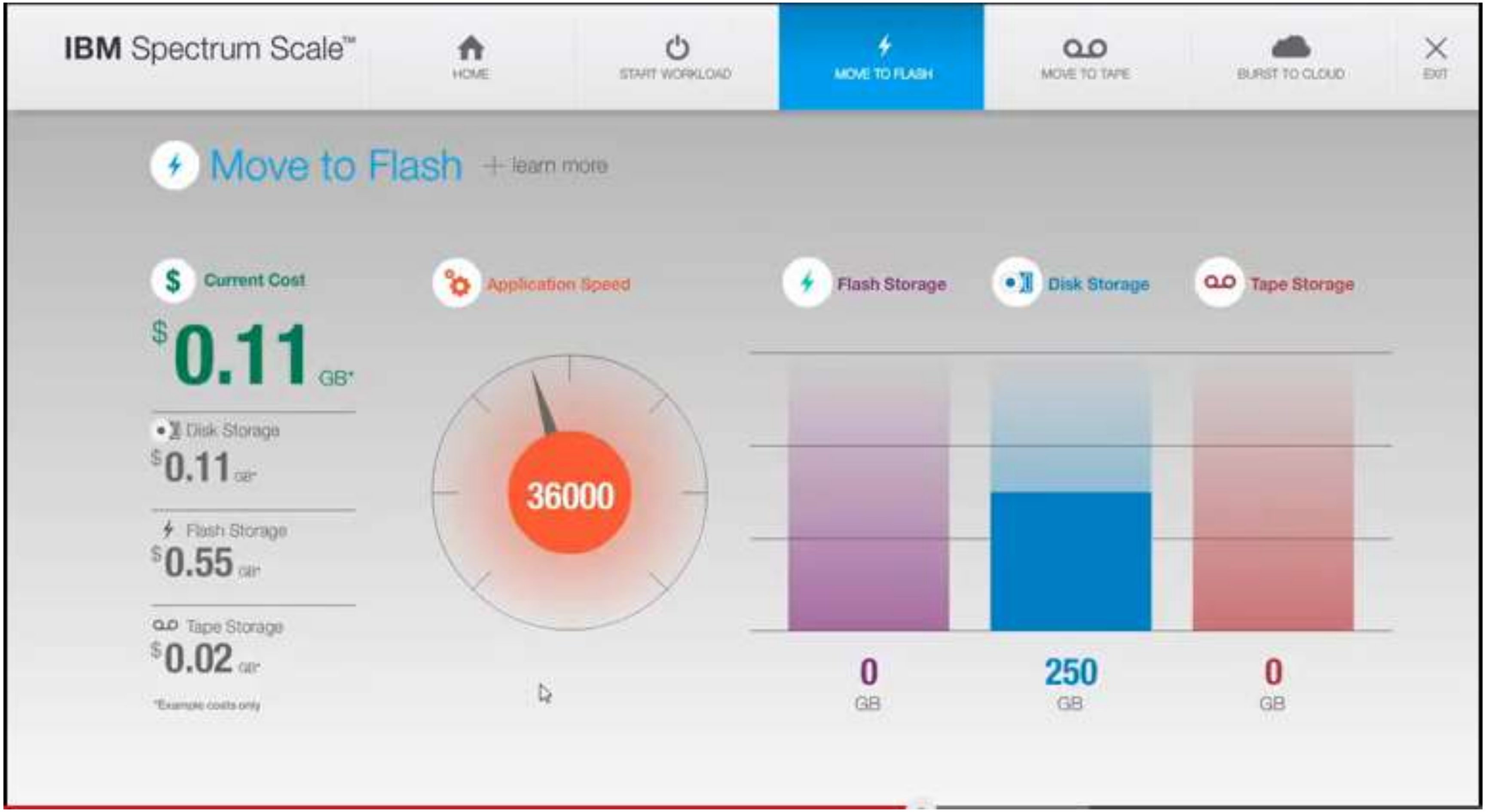
IBM Spectrum Scale Standard Edition v4.1.1 for Linux on z Systems Specifics disk support

- Support for ECKD™-based and FCP-based storage

- Supported storage systems
  - IBM System Storage® DS8000 Series
  - IBM Storwize® V7000 Disk Systems
  - IBM XIV® Storage Systems
  - IBM FlashSystem™ Systems
  - IBM System Storage SAN Volume Controller (SVC)
  - Competitive storage systems

- Minimum supported Linux distributions
  - Red Hat Enterprise Linux (RHEL) 7.0 and 6.5 (with specific errata)
  - SUSE Linux Enterprise Server (SLES) 12 and 11 SP3 (with specific maintweb) or later SP

IBM Spectrum Scale™ for Big Data & Analytics



Current Cost
$0.11 GB*

Disk Storage $0.11 GB

Flash Storage $0.55 GB

Tape Storage $0.02 GB

*Example costs only

Flash Storage 0 GB

Disk Storage 250 GB

Tape Storage 0 GB

https://www.youtube.com/watch?v=SmRZkynQSyc&feature=youtu.be

Use Cases

IBM WebSphere MQ and IBM WebSphere Application Server *Multi-Instance Queue Manager (MIQM) and HA Cluster*

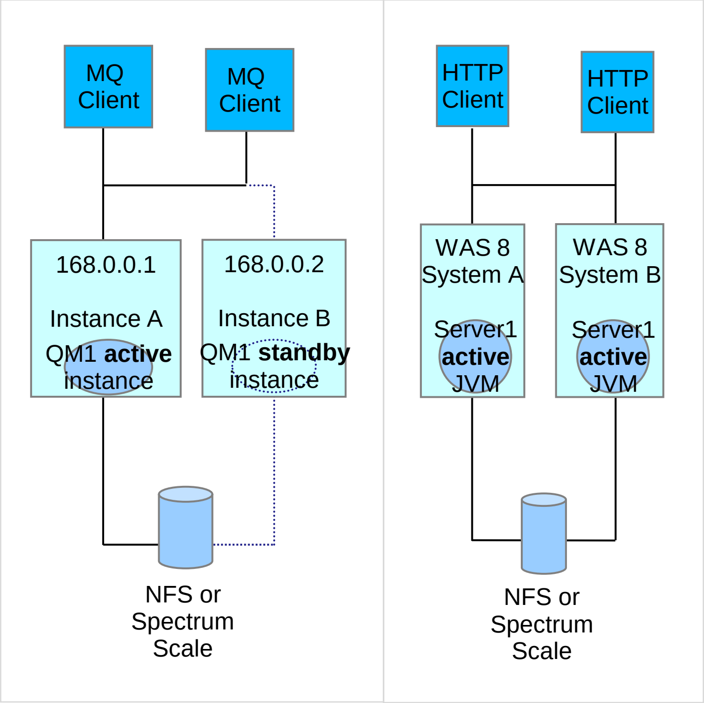
Need

High availability configuration of WebSphere MQ with two instances of the queue manager running on different servers, and either instance can be active.

High availability configuration of WebSphere Application Server (WAS) with two instances of the application running on different servers, and both instances are active.

Solution

A shared file system is required on networked storage, such as a NFS, or a clustered file system such as ***Spectrum Scale***



Advantages of ***Spectrum Scale*** versus NFS

- No single-server bottleneck
- No protocol overhead for data (network) transfer
- Interacts with applications like a local file system, while delivering high performance, scalability and fault tolerance by allowing data access from multiple systems directly and in parallel
- Maintaining file-data integrity while allowing multiple applications / users to share access to a single file simultaneously

Summary: IBM Spectrum Scale for Linux on z Systems

- **Quick access to enterprise file data**
  - No single-server bottleneck or protocol overhead for data transfer
- **Designed to deliver high performance, scalability and fault tolerance**
  - Allowing access to the data from multiple systems directly and in parallel
- **Shared access to a single file simultaneously while maintaining file-data integrity**
- **Takes file management beyond a single system**
  - Provides scalable access from multiple systems
- **Effective management of growing quantities of unstructured data**
- **Optimal use of enterprise available storage to deliver high performance**
  - Automatically spread across multiple storage devices
- **Reduces the amount of storage and management overhead**

Detailed Description: Getting started with IBM Spectrum Scale for Linux on z Systems

The IBM technical paper: Getting started with IBM Spectrum Scale for Linux on z Systems provides detailed information on:

- How to set up Linux on z Systems for Spectrum Scale (formerly IBM GPFS)
- How to install Spectrum Scale, how to set up and configure a Spectrum Scale cluster, and how to create a file system
- Best practices, hints, and tips

Download available at: ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGE_ZS_ZS_USEN&htmlfid=ZSW03272USEN&attachment=ZSW03272USEN.PDF

June 2015

IBM

Getting started with IBM Spectrum Scale for Linux on z Systems

Formerly IBM GPFS

Susanne Wintenberger
swinten@de.ibm.com
IBM R&D Boeblingen

Hendrik Brueckner
brueckner@de.ibm.com
IBM R&D Boeblingen

Resources

- ibm.com:

ibm.com/systems/platformcomputing/products/gpfs/

- Public Wiki:

ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/General Parallel File System (GPFS)

- IBM Knowledge Center:

ibm.com/support/knowledgecenter/SSFKCN/gpfs_welcome.html?lang=en

- Data sheet: IBM General Parallel File System (GPFS) Version 4.1

http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=SP&infotype=PM&appname=STGE_DC_ZQ_USEN&htmlfid=DCD12374USEN&attachment=DCD12374USEN.PDF

Questions?

IBM

Wilhelm Mild

IBM Executive IT Architect

The Open Group Open Certified
Distinguished Architect
IT Architecture
Chief/Lead IT Architect

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

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

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*
DB2*
DS8000*
ECKD

FlashSystem
IBM*
IBM (logo)*
MQSeries*

Storwize*
Spectrum Scale*
System p*
System x*
System z*

Tivoli*
WebSphere*
XIV*
z/VM*
Z Systems*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.

TEALEAF is a registered trademark of Tealeaf, an IBM Company.

Windows Server and the Windows logo are trademarks of the Microsoft group of countries.

Worklight is a trademark or registered trademark of Worklight, an IBM Company.

UNIX is a registered trademark of The Open Group in the United States and other countries.

* Other product and service names might be trademarks of IBM or other companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon 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 throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the 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.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.