Understanding IBM Spectrum Scale for Linux on z Systems (Express Edition)

IBM Redbooks Solution Guide

IBM Spectrum Scale for Linux on IBM® z Systems® is an extremely powerful file system. It is based on the IBM General Parallel File System (GPFS™) technology, which is a proven, scalable, high-performance data and file management solution, and also enabled for technical computing, Big Data & Analytics, and Cloud. IBM Spectrum Scale is being used extensively across multiple industries worldwide. This IBM Redbooks® Solution Guide describes the benefits of IBM Spectrum Scale.

Did you know?

IBM Spectrum Scale for Linux on z Systems supports extended count key data (IBM ECKD™) direct access storage device (DASD) disks and Fibre Channel Protocol (FCP) attached Small Computer System Interface (SCSI) disks.

IBM Spectrum Scale for Linux on z Systems nodes can communicate with each other through HiperSockets devices within one z Systems server, which provides high-speed IP network communication. Therefore, it can have better file system performance, especially in Network Shared Node (NSD) mode. The HiperSockets devices in two z Systems servers can be connected through a HiperSockets Bridge.

Each clustered file system has metadata. Some clustered file systems require a centralized metadata server, which can become a performance bottleneck for metadata-intensive operations and can represent a single point of failure. IBM Spectrum Scale solves this problem by managing metadata at the node that is using the file or, in the case of concurrent access to the file, at a dynamically selected node that is using the file.

Business value

Today’s data growth is challenging traditional storage and data management solutions. Limited data access, good performance, and reliability are required for IT environments. Also, application performance is affected by data access bottlenecks that delay schedules and waste expensive resources. Workloads are scaled up to large numbers of application nodes and disks, and because not all components are working correctly at all times, IT environments are required to handle component failures and continue the operation.

IBM Spectrum Scale for Linux on z Systems will enable enterprise clients to use a highly available clustered file system with Linux in a logical partition (LPAR) or as a Linux guest on IBM z/VM®.

IBM and independent software vendor (ISV) solutions will provide higher value for Linux on z Systems clients by exploiting IBM Spectrum Scale functionality:

- A highly available cluster architecture: IBM Spectrum Scale improves data availability through data access even when the cluster experiences storage or node malfunctions.
- Capabilities for high-performance parallel workloads: Concurrent high-speed, reliable file access from multiple nodes in the cluster environment.
- Smooth, nondisruptive capacity expansion and reduction are possible.
- Services are available to effectively manage large and growing quantities of data.

IBM Spectrum is designed to provide high availability through advanced clustering technologies, dynamic file system management, and data replication. IBM Spectrum can continue to provide data access even when the cluster experiences storage or node malfunctions. IBM Spectrum Scale scalability and performance are designed to meet the needs of data-intensive applications.

Solution overview

The first version of IBM Spectrum Scale for Linux on z Systems is based on IBM Spectrum Scale 4.1 Express Edition, which includes most base level features. IBM intends to offer additional functionality that is in the Standard and Advanced Editions in future versions of IBM Spectrum Scale for Linux on z Systems. The functions in the Express Edition include, but are not limited to the following functions:

- Snapshots
- NSD client/server capability
- Server failover
- Online or nondisruptive file system management, for example, adding and removing nodes and disks
- Logging

The Linux instances or nodes can be either Red Hat Enterprise Linux or SUSE Linux Enterprise Server, and they can run in LPARs or under z/VM as guest machines. The nodes also can be running on the same or different z Systems servers.

The IBM Spectrum Scale for Linux on z Systems is a clustered file system defined over one or more nodes. On each node in the cluster, it contains three basic components:

- GPFS administration commands
- Portability layer (kernel modules)
- Multithreaded daemon

IBM Spectrum Scale for Linux on z Systems uses a portability layer (kernel modules) that enables the GPFS daemon to interact with the Linux kernel. During the installation, you build the portability layer on your Linux instance, which fits in a wide variety of Linux Kernel versions and configurations. Figure 1 shows the basic IBM Spectrum Scale structure.

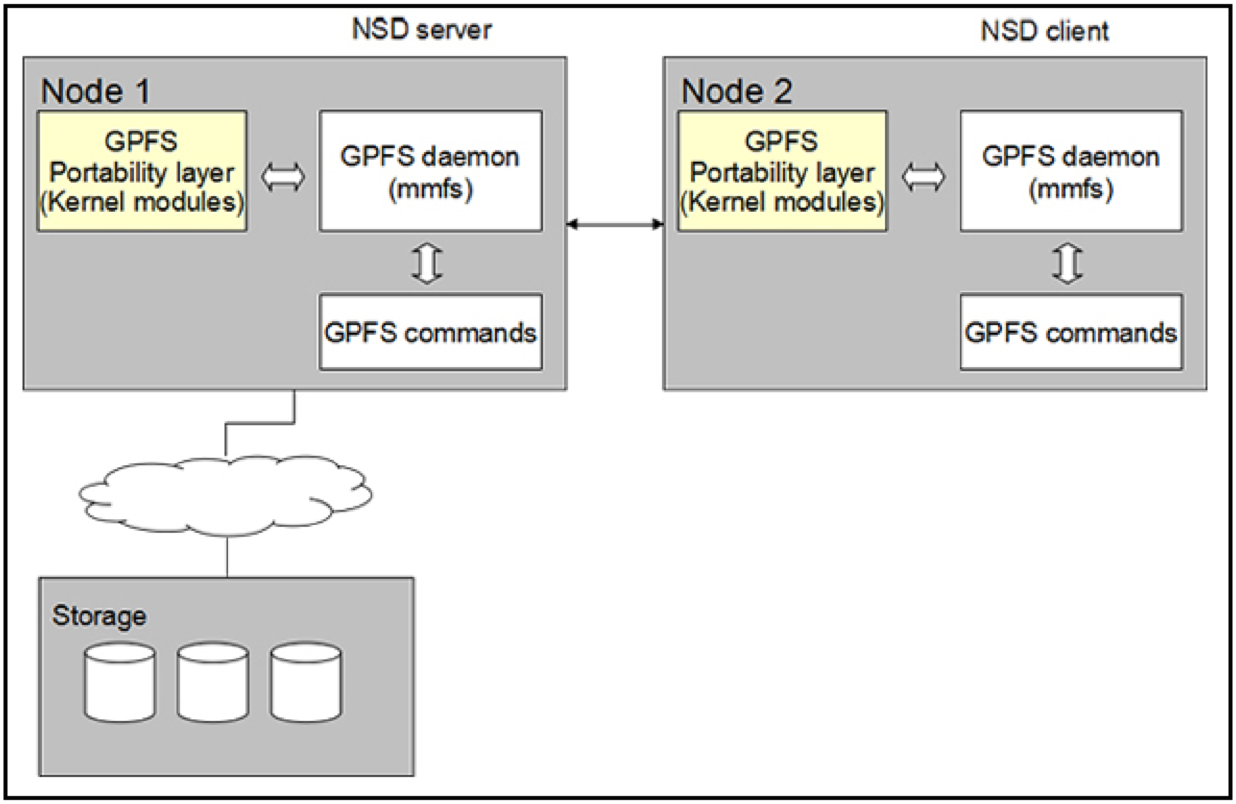

Figure 1. IBM Spectrum Scale basic structure

IBM Spectrum Scale depends on the correct operation of an IP network to communicate with other nodes. In z Systems, if the nodes are in the same z Systems server, the communication can use HiperSockets devices, which can provide higher network speed and more secure and better connectivity performance. It is particularly suitable for Network Shared Disk (NSD) model (NSD server/client structure) because it generates large amounts of data traffic between nodes.

The current version of IBM Spectrum Scale for Linux on z Systems can support up to 32 nodes and a heterogeneous cluster in Network Shared Disk (NSD) mode. In a heterogeneous cluster, the NSD server must be Linux on z Systems and the NSD clients (without direct storage access) can be also based on AIX®, Red Hat and SUSE Linux distributions on IBM Power® Systems and x86 Linux.

Solution architecture

IBM Spectrum Scale for Linux on z Systems can work in two modes:

- Shared Disk (SAN) model
- Network Shared Disk (NSD) model

Shared Disk (SAN) model

In this type of configuration, all of the nodes in the IBM Spectrum Scale cluster are connected to a common set of disks, as shown in Figure 2. In this model, the disk I/O can perform better because all the nodes connect to the storage servers directly. This configuration does not support heterogeneous platforms.

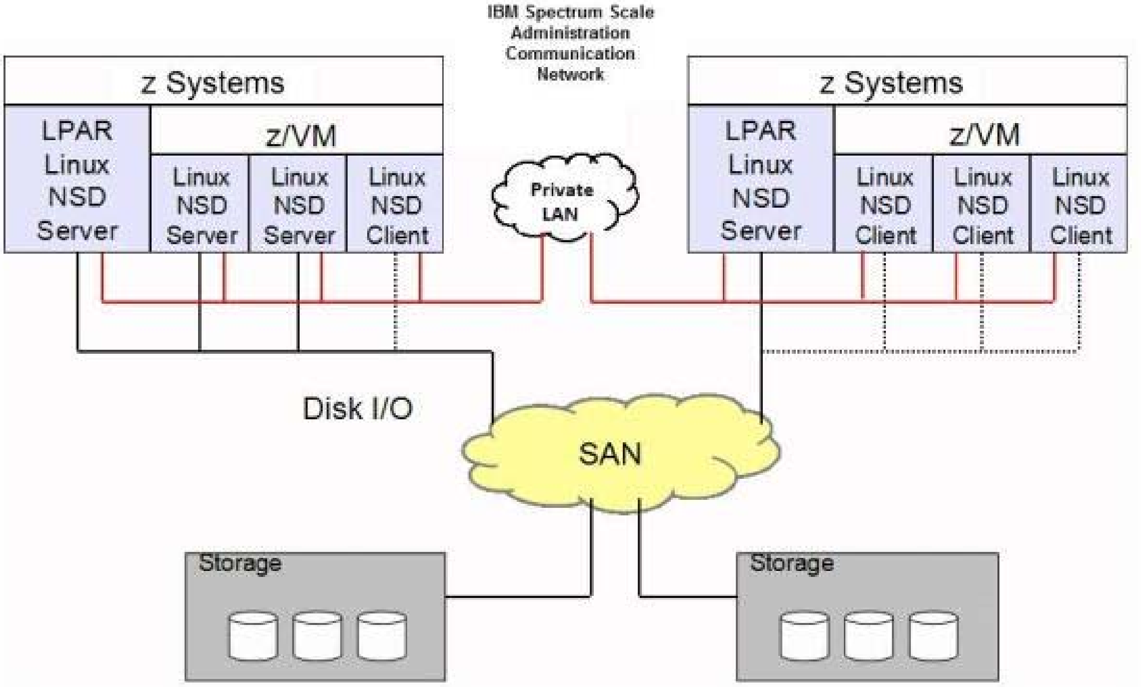

Figure 2. IBM Spectrum Scale Shared Disk (SAN) mode

Network Shared Disk (NSD) model

You can configure an an IBM Spectrum Scale cluster in which some nodes attach directly to the disks and other nodes access the disks through the an IBM Spectrum Scale server nodes. This configuration is often used in large clusters or to provide a cost-effective, potential high-performance solution.

When an an IBM Spectrum Scale node provides access to a disk for anotheran IBM Spectrum Scale node, it is called an NSD server. Thean IBM Spectrum Scale node accessing the data through an NSD server is called an NSD client. In Figure 3, the NSD servers connect to storage servers directly, and the NSD client accesses the file system through a high-speed network connecting to NSD servers.

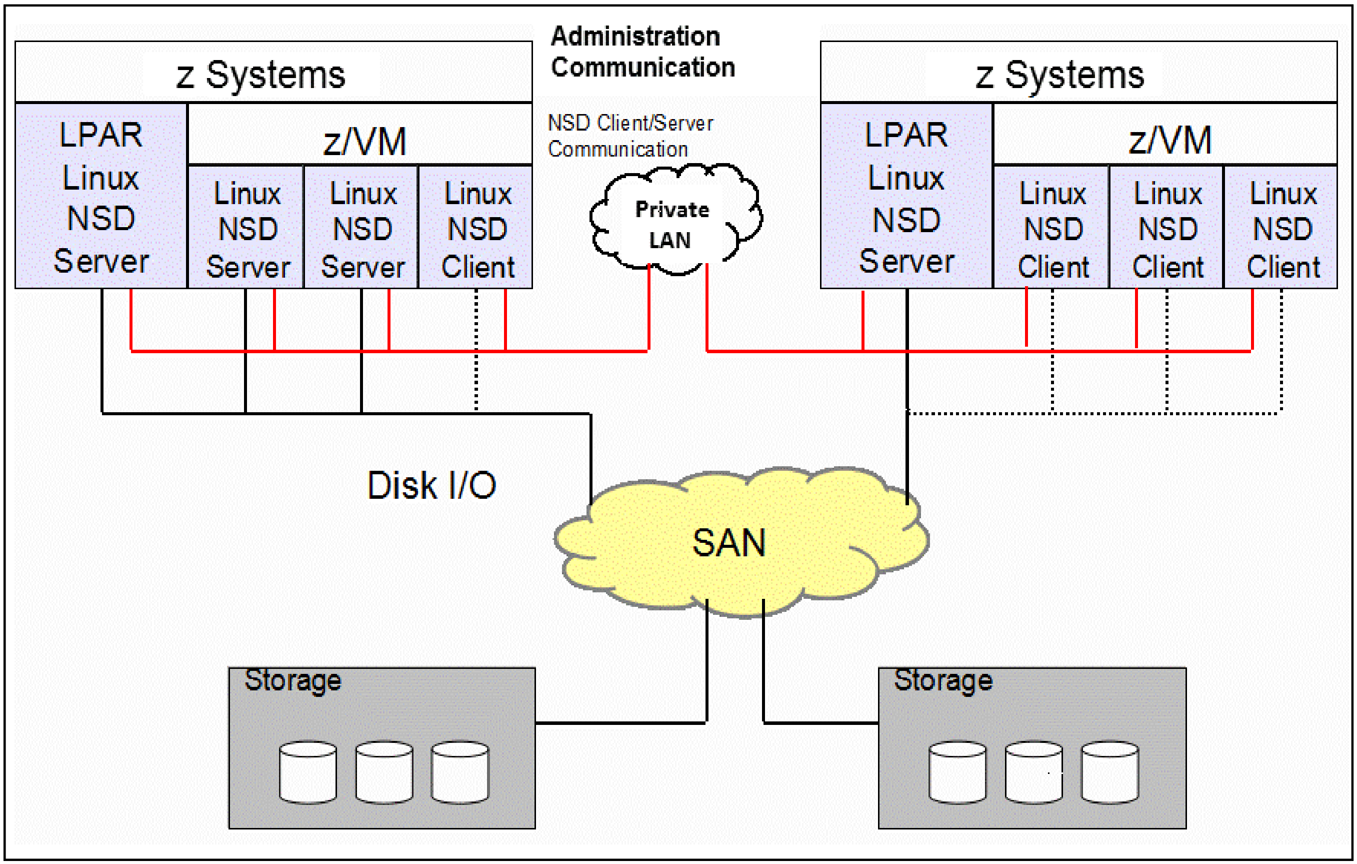

Figure 3. Network Shared Disk (NSD) model

Note: For nodes without direct attachment to the shared storage, NDS (block/disk data traffic) access is done over the network.

Usage scenarios

This guide describes IBM Spectrum Scale use cases that you can use on the IBM z Systems platform to help you achieve better reliability and performance. IBM Spectrum Scale for Linux on z Systems is supported in many more scenarios than the scenarios described here.

High availability with IBM WebSphere MQ Multi -Instance Queue Manager (MIQM)

For business continuity, high availability solutions need to be employed. There are multiple high availability solutions for IBM WebSphere® MQ from both hardware and middleware perspectives. WebSphere MQ Multi-Instance Queue Manager (MIQM) is one of the predominant high availability solutions. MIQM is a software-based high availability (active-standby) solution. It defines an active instance of the queue manager on one server and a standby instance on another server. The active instance processes messages and accepts connections from applications and from other queue managers. It holds a lock on the queue manager data to ensure that there is only one active instance of the queue manager at a specific time accessing the data. Message input queues and logs for Multi-Instance Queue Managers are held on network storage, such as NFS (Figure 4) and the and IBM Spectrum Scale clustered file system (Figure 5) and are shared by the two servers.

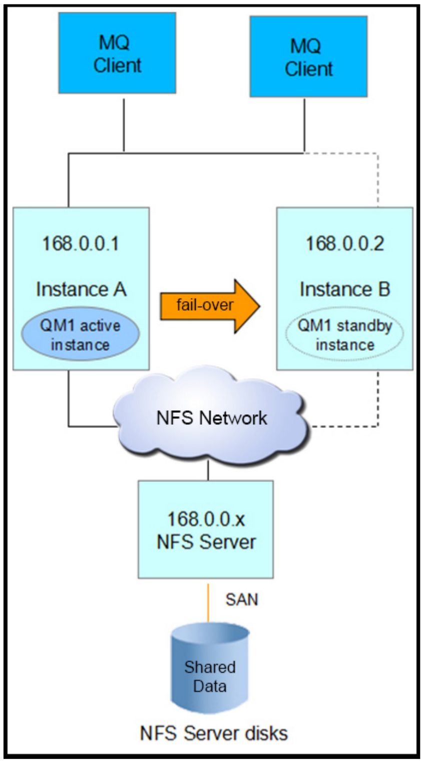

Figure 4. MIQM HA solution with NFS

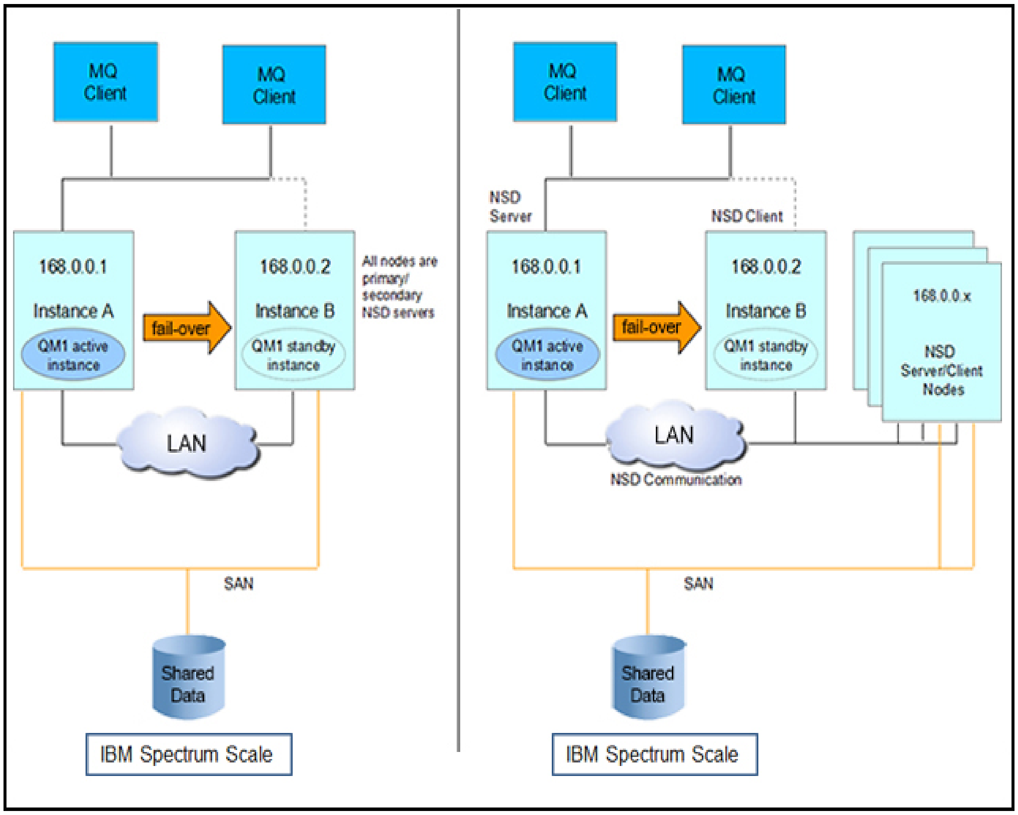

Figure 5. MIQM HA Solution with IBM Spectrum Scale

As seen in Figure 4 and Figure 5, if the high availability solution uses NFS as its shared network storage, it requires an NFS server to hold the shared data on its disks. If we adopt IBM Spectrum Scale in this solution, it does not need a separate server to hold the shared data. The data is across all the IBM Spectrum Scale nodes. According to the configurations of theIBM Spectrum Scale file system, the MIQM High Availability (HA) solution can be defined in two different modes - Shared Disk (SAN) model or Network Shared Disk (NSD) model.

Shared Disk (SAN) model

On the left side of Figure 5, both servers attach to the shared data disks physically. In this model, both nodes act as primary or secondary Network Shared Disk (NSD) servers. All the nodes in this configuration have good disk I/O performance. The network communications between the two nodes are through a private network, for example, it can be configured with HiperSockets devices or Shared Open Systems Adapter (OSA).

Network Shared Disk (NSD) model

On the right side of Figure 5, we defined the active Queue Manager server as an NSD server, where the shared disks are attached, and we defined the standby instance as the NSD client. The NSD nodes in the same z Systems server can use the high-speed HiperSockets devices for better performance. There are other nodes acting as NSD servers in a customer's installation to help avoid a single NSD server failure. This model is suitable when IBM Spectrum Scale nodes already exist and you need to add another node to access the shared disks (for example, creating a MIQM cluster). In that case, a user can add a standby Queue Manager instance without any configuration changes related to the shared disks' access.

High Availability with a WebSphere Application Server cluster

In certain circumstances, you might require the WebSphere Application Server High Availability solution to provide workload management and failover for applications that reside on the application server cluster. IBM WebSphere Application Server offers a built-in application server clustering function and the HAManager for protecting WebSphere singleton services. The HAManager enhances the high availability of WebSphere singleton services, such as transaction or messaging services. It provides a peer recovery mechanism for in-flight transaction logs or messages among clustered WebSphere application servers. The WebSphere Application Server Transaction Manager writes to its transaction recovery logs when it handles global transactions that involve two or more resources. Transaction recovery logs are stored on disks and are used for recovering in-flight transactions from system crashes or process failures. To enable WebSphere Application Server transaction peer recovery, it is necessary to place the recovery logs on a highly available file system, such as an IBM SAN file system or network-attached storage (NAS), for all the application servers within the same cluster to access. All application servers must be able to read from and write to the logs. Before IBM Spectrum Scale for Linux on z Systems became available, customers used a solution with a Network File System (NFS), for example, such as the solution in Figure 6.

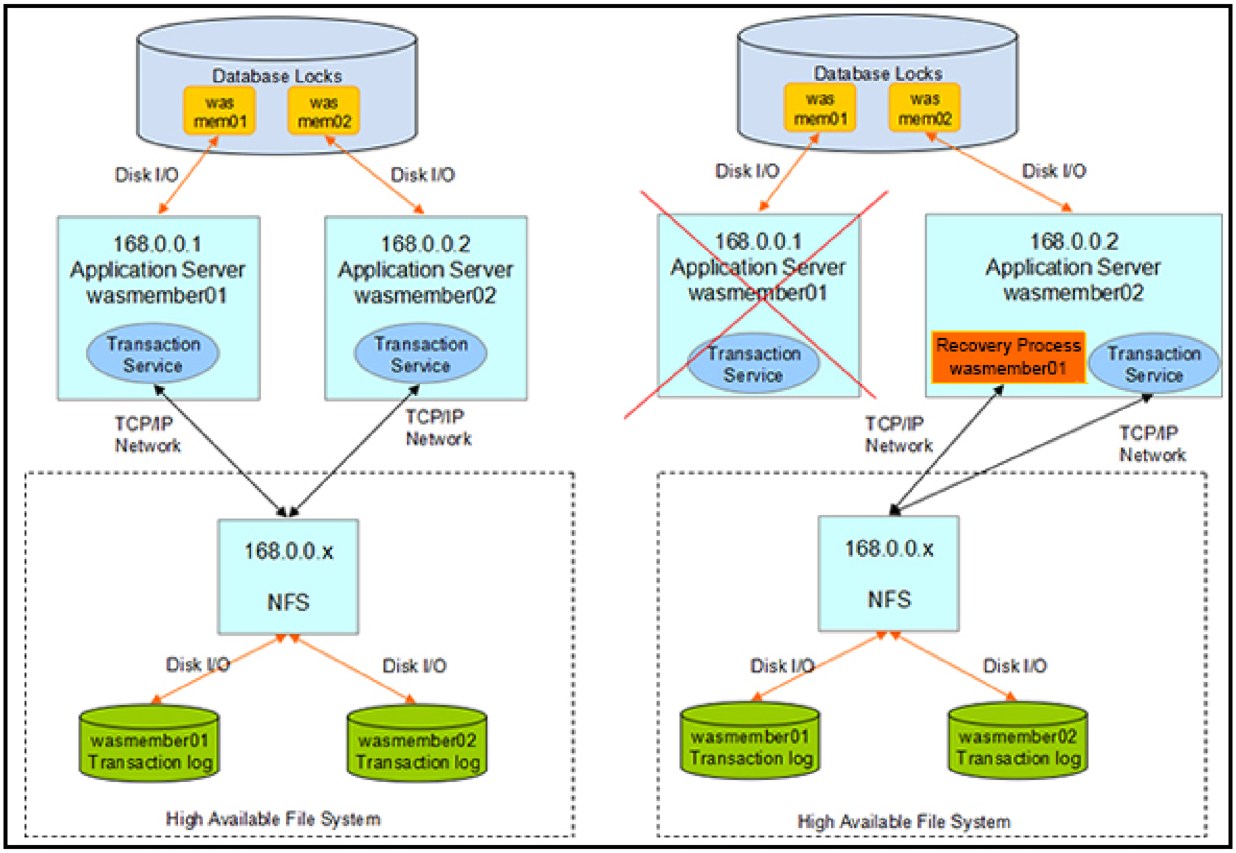
Database Locks
was mem01
was mem02
Disk I/O
Disk I/O
168.0.0.1 Application Server wasmember01
Transaction Service
168.0.0.2 Application Server wasmember02
Transaction Service
TCP/IP Network
TCP/IP Network
168.0.0.x
NFS
Disk I/O
Disk I/O
wasmember01 Transaction log
wasmember02 Transaction log
High Available File System
Database Locks
was mem01
was mem02
Disk I/O
Disk I/O
168.0.0.1 Application Server wasmember01
Transaction Service
168.0.0.2 Application Server wasmember02
Recovery Process wasmember01
Transaction Service
TCP/IP Network
TCP/IP Network
168.0.0.x
NFS
Disk I/O
Disk I/O
wasmember01 Transaction log
wasmember02 Transaction log
High Available File System

Figure 6. WebSphere Application Server Transaction Manager Failover Solution with NFS

With the IBM Spectrum Scale clustered file system, you do not need NFS involved, as shown in Figure 7. In Figure 7, we use IBM Spectrum Scale Storage Shared Disk (SAN) mode, for example.The NSD mode is similar, as shown on the right side of Figure 5.

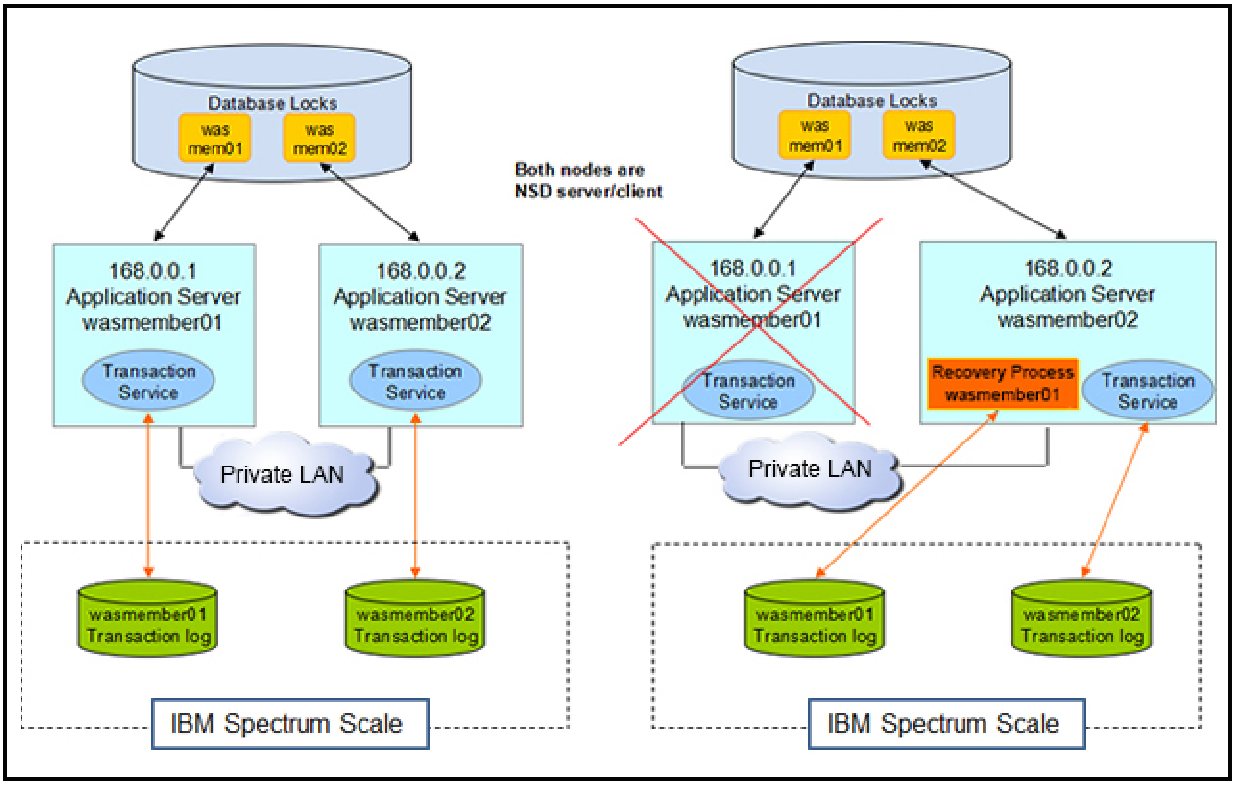

Figure 7. WebSphere Application Server Transaction Manager Failover Solution with IBM Spectrum Scale clustered file system

IBM Spectrum Scale file system compared to NFS

Compared to the high availability solutions with NFS, the IBM Spectrum Scale clustered file system solution offers the following advantages:

- No single point of failure in the shared file system: In a solution with NFS, the whole solution is jeopardized when the NFS server malfunctions, because both WebSphere MQ instances or WebSphere Application Server cannot access the shared data. Even if you can configure a clustered NFS with multiple NFS servers manually or with tools, it is still not reliable enough and adds additional resource overhead. However, IBM Spectrum Scale doesn’t have a single point of failure because both nodes can access the shared data concurrently.
- Nondisruptive file system scale-out: Whenever there are changes in NFS, for example, the file system size, all the clients need to remount the file system to refresh the changes, which means that the business is disrupted. However, with IBM Spectrum Scale for Linux on z Systems, the file system can be scaled out without stopping your business.
- No performance bottleneck in data and metadata: In NFS solutions, data and metadata performance are often the bottlenecks. The IBM Spectrum Scale file system is designed to support more files than NFS with high performance data and metadata access as part of the original design. The underlying NFS protocol does not support several of the features that are available in the IBM Spectrum Scale file system, for example, the capability to list enormous files. In terms of file open and creation performance, the IBM Spectrum Scale file system is superior to NFS. The NFS protocol was not designed for the type of performance that is required by large environments.
- No request I/O size restriction: In the NFS environment, you have to configure the rsize and wsize parameters from the performance perspective. However, with IBM Spectrum Scale, the I/O size requests made to the file system servers are generally the size of the I/O requests from the queue managers. So, the IBM Spectrum Scale clustered file system allows larger requests if the application can be configured or already makes larger requests than NFS supports.
- Synchronize write without data loss: NFS is always configured as asynchronized data writing to disk. It is part of the NFS design. Although you can configure the NFS in synchronized write, there is a performance cost. When NFS is working in async mode, it will experience data loss due to an NFS malfunction. Synchronization I/O is part of the design of the IBM Spectrum Scale file system, which means that it can avoid data loss. It also can reduce the CPU overhead to deal with the cache operations.

Integration

IBM Spectrum Scale for Linux on z Systems can be integrated with other IBM products and solutions as the base clustered file system, such as Business Analytics, Cloud, and storage HA solutions.

Supported platforms

The following platforms are supported by IBM Spectrum Scale for Linux on z Systems Express Edition (Version 4.1). Table 1 shows the supported Linux distributions.

Table 1. Supported Linux distributions

Distribution	Minimum level	Kernel
SUSE Enterprise Server 11	SUSE Linux Enterprise Server 11 SP3 + Maintweb Update or later maintenance update or Service Pack	3.0.101-0.15-default
Red Hat Enterprise Linux Server 6	Red Hat Enterprise Linux 6.5 + Errata Update RHSA-2014-0328, or later minor update	2.6.32-431.11.2.el6

Red Hat Enterprise Linux 7		3.10.0-123.6.3.el7.s390x

Table 2 shows the supported storage systems.

Table 2. Supported storage systems

Storage system	SCSI device	ECKD device
IBM DS8000® series	NSD or PR	NSD
IBM Storwize V7000	NSD or PR	N/A
IBM XIV®	NSD or PR	N/A
IBM FlashSystem™	NSD or PR	N/A
IBM SAN Volume Controller (SVC)	NSD or PR	N/A

Notes:
NSD: Network Shared Disk Leasing
PR: Persistent Reserve

Ordering information

The ordering information for IBM Spectrum Scale for Linux on z Systems is shown in the Table 3.

Table 3. Ordering part numbers and feature codes

Program name	PID number
IBM GPFS for Linux on System z	5725-S28

Related information

For more information, see the following documents:

- Solution brief: IBM Spectrum Scale for Linux on z Systems: http://www.ibm.com/common/ssi/cgi-bin/ssialias?subtype=SP&infotype=PM&appname=STGE_ZS_ZS_USEN&htmlfid=ZSS03118USEN
- Announcement letters ibm.com/common/ssi/SearchResult.wss?request_locale=en&dateval=index_customrange#ctype=ANCA&ctry=AMR|ASP|EUR|MDE&MPPEFSCH=GPFS for Linux on System z&MPPEFFDR=2014-10-06&MPPEFTDR=2014-10-06
- Home page: IBM Spectrum Scale ibm.com/systems/platformcomputing/products/gpfs
- Home page: Linux on IBM z Systems ibm.com/systems/z/os/linux

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service. IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you. This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you. 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 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. This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document was created or updated on February 16, 2015.

Send us your comments in one of the following ways:

- Use the online **Contact us** review form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips1211.html .

Trademarks