Cluster: Servers Combined to Improve Availability and Scalability.

- Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”. 
Failover Cluster
Cluster storage

Hardware requirements:

- iSCSI
- SAS
- Fiber Channel
- Fibre Channel over Ethernet (FcoE)
iSCSI
iSCSI (cont)
iSCSI vs. virtio-scsi performance test
iSCSI vs. virtio-scsi performance test (cont.)

Graph 1: 256K Reads (IOPS)

Graph 2: 256K Writes (IOPS)
MS Exchange Jetstress
Jetstress latency results
Failover Cluster Manager
Failover Cluster Manager (cont.)

Inventory virtio-scsi

---

Microsoft

Failover Cluster Validation Report

Node: WIN-FC0.corp.vrozenfe.com
Checked: 9/6/2015 7:00:45 AM
Completed: 9/4/2015 7:00:49 AM

Inventory

<table>
<thead>
<tr>
<th>Name</th>
<th>Result</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>List SAS Host Bus Adapters</td>
<td>✔️</td>
<td>Success</td>
</tr>
</tbody>
</table>

Overall Result

Testing has completed for the tests you selected. To confirm that your cluster solution is supported, you must run all tests. A cluster solution is supported by Microsoft only if it passes all tests in the wizard.

List SAS Host Bus Adapters

List Serial Attached SCSI (SAS) host bus adapters on each node.

WIN-FC0.corp.vrozenfe.com

Gathering SAS Host Bus Adapter information for WIN-FC0.corp.vrozenfe.com

None found ...
Failover Cluster Manager (cont.)

Inventory Isi_sas (VMWare Fusion)

List SAS Host Bus Adapters

List Serial Attached SCSI (SAS) host bus adapters on each node.

win-fc0.corp.fusion.com

Gathering SAS Host Bus Adapter information for win-fc0.corp.fusion.com

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Model</th>
<th>Driver Name</th>
<th>Number of Ports</th>
<th>Driver Version</th>
<th>Firmware Version</th>
<th>Serial Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI Corporation</td>
<td>SAS2444</td>
<td>Isi_sas</td>
<td>2</td>
<td>1.20.02.52</td>
<td>01.03.41.32</td>
<td></td>
</tr>
</tbody>
</table>
Windows Management Instrumentation
WMI discovering GUID List
WMI discovering GUID List (cont)

scsiwmi.h
Abstract:
   This module contains the internal structure definitions and APIs used by the SCSI WMILIB helper functions
   //
   // This structure supplies context information for SCsSIWMILIB to process the WMI srbs.

typedef struct _SCSIIWMILIB_CONTEXT
{
   // WMI data block guid registration info
   ULONG GuidCount;
   PSCSIWMIGUIDREGINFO GuidList;
   // WMI functionality callbacks
   PSCSIWMI_QUERY_REGINFO QueryWmiRegInfo;
   .......
} SCSI_WMILIB_CONTEXT, *PSCSI_WMILIB_CONTEXT;

typedef struct
{
   LPCGUID Guid;       // Guid representing data block
   ULONG InstanceCount; // Count of Instances of Datablock. If this count is 0xffffffff then the guid is assumed to be dynamic instance names
   ULONG Flags;        // Additional flags (see WMIREGINFO in wmistr.h)
} SCSIWMIGUIDREGINFO, *PSCSIWMIGUIDREGINFO;
WMI discovering GUID List
WMI discovering GUID List

//****************************************************************************
//
//  hbapiwmi.h
//
//  Module: WDM classes to expose HBA api data from drivers
//
//  Purpose: Contains WDM classes that specify the HBA data to be exposed
//           via the HBA api set.
//
//  NOTE: This file contains information that is based upon:
//        SM-HBA Version 1.0 and FC-HBA 2.18 specification.
//
#define MS_SM_AdapterInformationQueryGuid \
   { 0xbdc67efa,0xe5e7,0x4777, { 0xb1,0x3c,0x62,0x14,0x59,0x65,0x70,0x99 } }

#define MS_SM_PortInformationMethodsGuid \
   { 0x5b6a8b86,0x708d,0x4ec6, { 0x82,0xa6,0x39,0xad,0xcf,0x6f,0x64,0x33 } }

Failover Cluster Manager (cont.)

List All Disks

<table>
<thead>
<tr>
<th>Disk Number</th>
<th>Disk Identifier</th>
<th>Disk Bus Type</th>
<th>Disk Slot Type</th>
<th>Disk Address (scsiPath/shell)</th>
<th>Adapter Description</th>
<th>Eligible for Validation</th>
<th>Disk Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhysicalDrive0</td>
<td>2x24000D</td>
<td>5G</td>
<td>5G</td>
<td>0.0.0.0</td>
<td>IBM Channel</td>
<td>False</td>
<td>Disk is a boot volume. Do not use for RAID. Disk type does not support clustering. Disk bus type does not support clustering. Use the system bus. Disk partition type is HBD. Disk partition type is HBD. Port driver of the disk does not support clustering. Do not use for RAID. Disk partition type is HBD.</td>
</tr>
<tr>
<td>PhysicalDrive1</td>
<td>3001a74</td>
<td>5G</td>
<td>5G</td>
<td>0.0.0.0</td>
<td>Red Hat VirtIO SC5 controller</td>
<td>False</td>
<td>Disk is a boot volume. Do not use for RAID. Disk type does not support clustering. Disk bus type does not support clustering. Use the system bus. Disk partition type is HBD. Disk partition type is HBD. Port driver of the disk does not support clustering. Do not use for RAID. Disk partition type is HBD.</td>
</tr>
</tbody>
</table>
Failover Cluster Manager (cont.)

List All Disks log file

```
Z:\EC\Reports\ValidateStorage.log
000009bc.00000288::19:03:56.089 DoIoctlAndAlloc: ControlCode 0x7f050, retCode 1, status 122
000009bc.00000288::19:03:56.089 DoIoctlAndAlloc: ControlCode 0x7f050, retCode 1, status 122
000009bc.00000288::19:03:56.089 IsDynamicDisk: Exit IsDynamicDisk: DynamicDisk 0, status 0
000009bc.00000288::19:03:56.089 CprepDiskGetProps: Exit CprepDiskGetProps: hr 0x0, DiskProps->Flags 0x9317
000009bc.00000288::19:03:57.005 CprepDiskGetProps: Enter CprepDiskGetProps: DiskIdType 4000 DiskSignature 1
000009bc.00000288::19:03:57.005 DoIoctlAndAlloc: ControlCode 0x74208, retCode 1, status 0
000009bc.00000288::19:03:57.021 CreateNtFile: Path \Device\ScsiPort3, status 0
000009bc.00000288::19:03:57.021 IsClusterSupported: Port driver does not support clustering
000009bc.00000288::19:03:57.021 IsClusterSupported: Exit IsClusterSupported: \Device\ScsiPort3, ClusterSupported 0, status 0
000009bc.00000288::19:03:57.036 CprepDiskGetProps: Port driver does not support clustering
000009bc.00000288::19:03:57.036 GetAdapterBusType: Exit GetAdapterBusType: BusType 10, status 0
000009bc.00000288::19:03:57.036 EnumerateDevices: Enter EnumerateDevices: EnumDevice 0
000009bc.00000288::19:03:57.052 EnumerateDevices: opened file \\ide\diskqemu_harddisk_2.3.50_5817595
```
Failover Cluster Manager (cont.)

Clusters.dll
Failover Cluster Manager (cont.)

List All Disks log file
# IOCTL_SCSI_MINIPORT

```c
#include "ntddscsi.h"
#define IOCTL_SCSI_MINIPORT CST_CODE(IOCTL_SCSI_BASE, 0x0402, METHOD_BUFFERED, FILE_READ_ACCESS | FILE_WRITE_ACCESS)

#include "ddk/scsi.h"
#define IOCTL_SCSI_MINIPORT_NOT_QUORUM_CAPABLE ((FILE_DEVICE_SCSI << 16) + 0x0520)

typedef struct _SRB_IO_CONTROL {
    ULONG HeaderLength;
    UCHAR Signature[8];
    ULONG Timeout;
    ULONG ControlCode;
    ULONG ReturnCode;
    ULONG Length;
} SRB_IO_CONTROL, *PSRB_IO_CONTROL;
```
unsigned size = sizeof(SRB_IO_CONTROL);
SRB_IO_CONTROL srbc;
DWORD num_out;

srbc.HeaderLength = size;
memcpy(srbc.Signature, "CLUSDISK", 8);
srbc.Timeout = 3;
srbc.ControlCode = IOCTL_SCSI_MINIPORT_NOT_QUORUM_CAPABLE;

if (!DeviceIoControl(hdevice, IOCTL_SCSI_MINIPORT, 
    &srbc, size, NULL, 0, &num_out, NULL)) {

### Storage Test

#### Storage

<table>
<thead>
<tr>
<th>Name</th>
<th>Result</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>List All Disks</td>
<td>✔️</td>
<td>Success</td>
</tr>
<tr>
<td>List Potential Cluster Disks</td>
<td>✔️</td>
<td>Success</td>
</tr>
<tr>
<td>Validate Disk Access Latency</td>
<td>✔️</td>
<td>Success</td>
</tr>
<tr>
<td>Validate Disk Arbitration</td>
<td>✔️</td>
<td>Success</td>
</tr>
<tr>
<td>Validate Disk Failover</td>
<td>❌</td>
<td>Cancelled</td>
</tr>
<tr>
<td>Validate File System</td>
<td>❌</td>
<td>Cancelled</td>
</tr>
<tr>
<td>Validate Microsoft MP10-based disks</td>
<td>✔️</td>
<td>Success</td>
</tr>
<tr>
<td>Validate Multiple Arbitration</td>
<td>✔️</td>
<td>Success</td>
</tr>
<tr>
<td>Validate SCSI device Vital Product Data (VPD)</td>
<td>❌</td>
<td>Warning</td>
</tr>
<tr>
<td>Validate SCSI-3 Persistent Reservation</td>
<td>❌</td>
<td>Failed</td>
</tr>
<tr>
<td>Validate Simultaneous Failover</td>
<td>❌</td>
<td>Cancelled</td>
</tr>
</tbody>
</table>

#### Validate SCSI-3 Persistent Reservation

Validate that storage supports the SCSI-3 Persistent Reservation commands.

- Validating Cluster Disk 0 for Persistent Reservation support
- Registering PR key for cluster disk 0 from node WIN-FC1.corp.vrozenfe.com
- Failed to Register PR key for cluster disk 0 from node WIN-FC1.corp.vrozenfe.com status 21
- Cluster Disk 0 does not support Persistent Reservation
- Test failed. Please look at the test log for more information
commit 8fdc7839e40f43a426bc7e858cf1dbfe315a3804
Author: Paolo Bonzini <pbonzini@redhat.com>
Date: Tue May 10 10:50:44 2016 +0200

scsi-block: always use SG_IO

Using pread/pwrite or io_submit has the advantage of eliminating the bounce buffer, but drops the SCSI status. This keeps the guest from seeing unit attention codes, as well as statuses such as RESERVATION CONFLICT. Because we know scsi-block operates on an SBC device we can still use the DMA helpers with SG_IO; just remember to patch the CDBs if the transfer is split into multiple segments.

This means that scsi-block will always use the thread-pool unfortunately, instead of respecting aio=native.

Signed-off-by: Paolo Bonzini <pbonzini@redhat.com>
Storage Test
THANK YOU

plus.google.com/+RedHat
linkedin.com/company/red-hat
youtube.com/user/RedHatVideos
facebook.com/redhatinc
twitter.com/RedHatNews