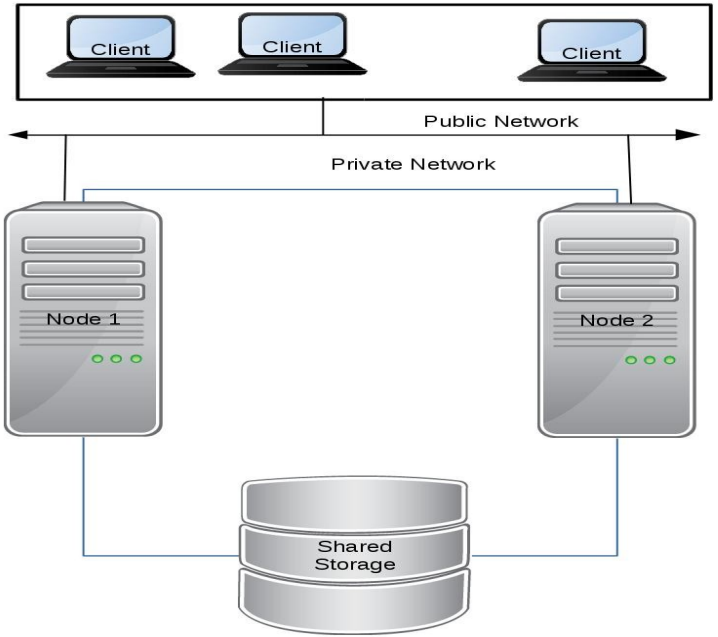
redhat.

MS Cluster on KVM

Vadim Rozenfeld
vrozenfe@redhat.com
25 Aug, 2016

Cluster: Servers Combined to Improve Availability and Scalability.

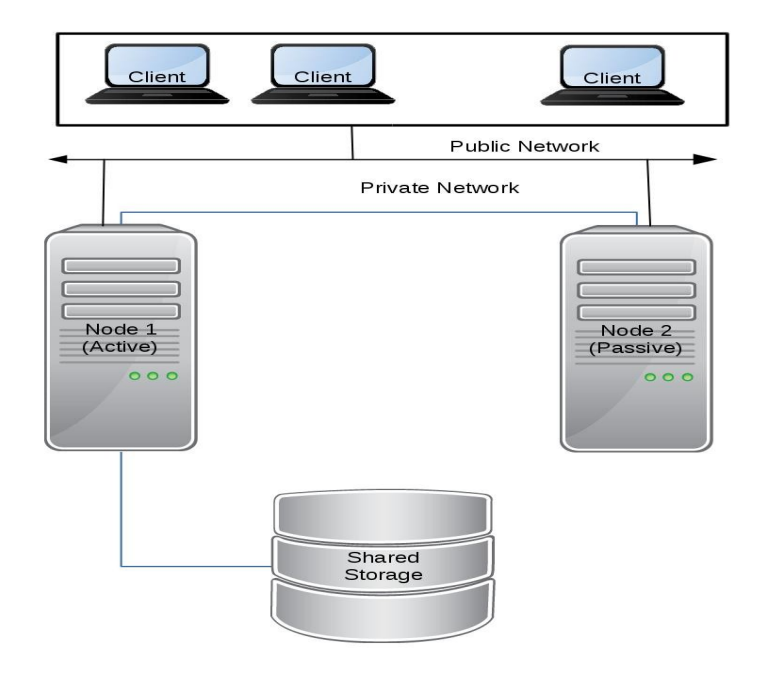


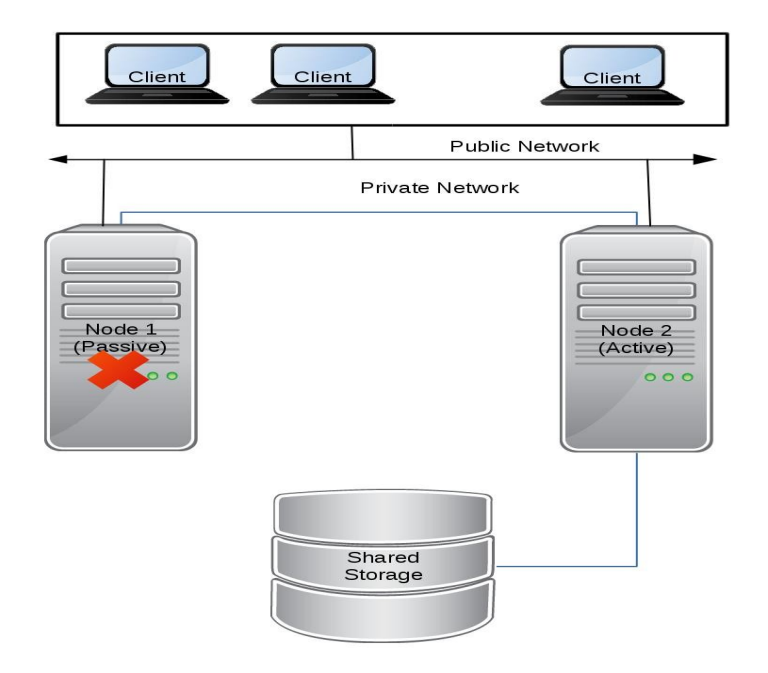
- Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.

- Node: A server in a cluster.

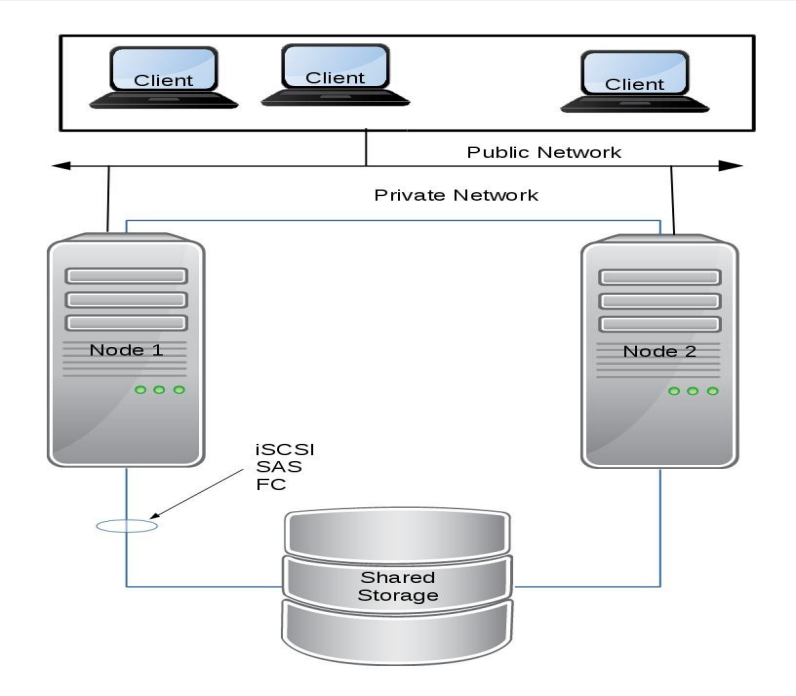
- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

Failover Cluster





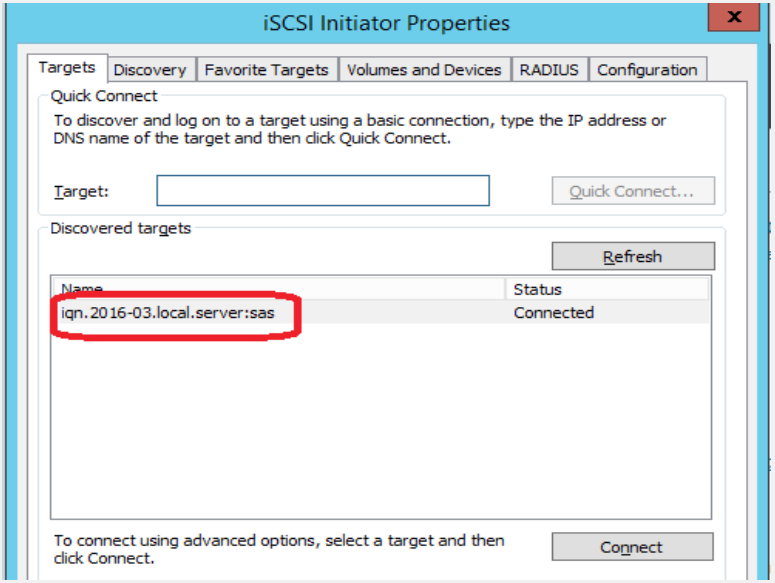
Cluster storage

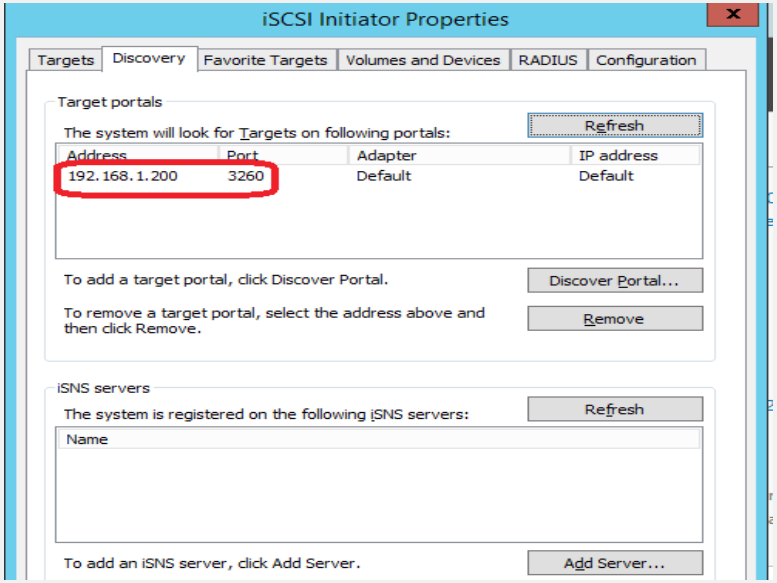


Hardware requirements :

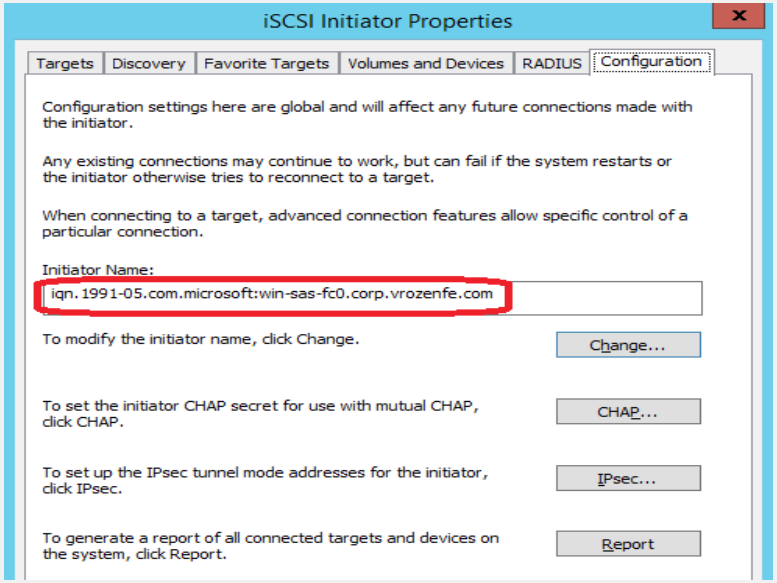
- iSCSI
- SAS
- Fiber Channel
- Fibre Channel over Ethernet (FcoE)

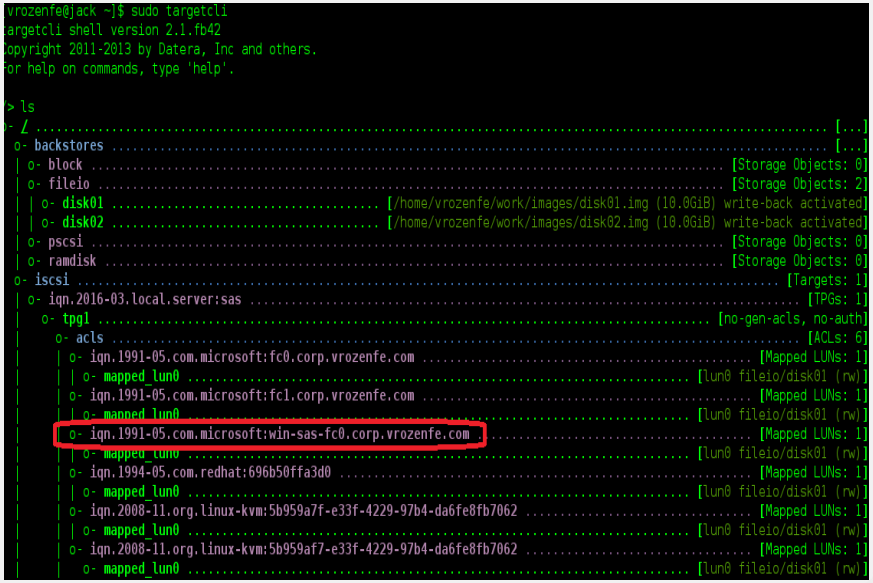
iSCSI



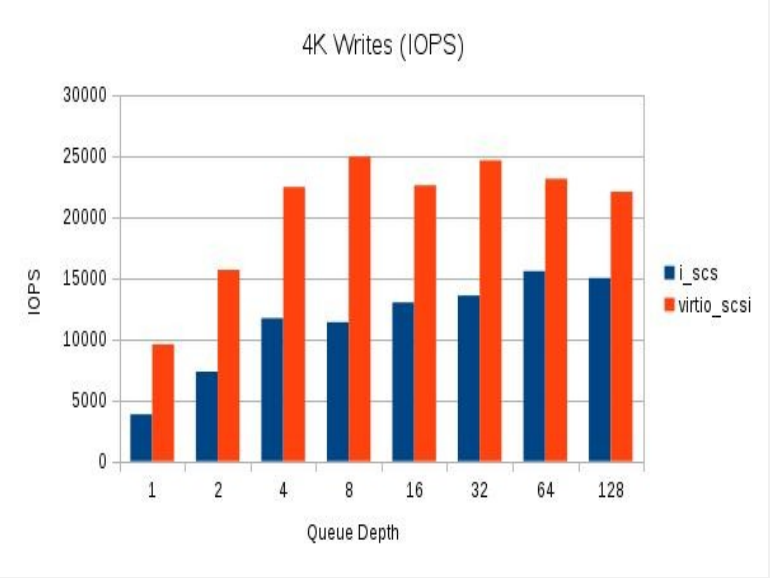


iSCSI (cont)

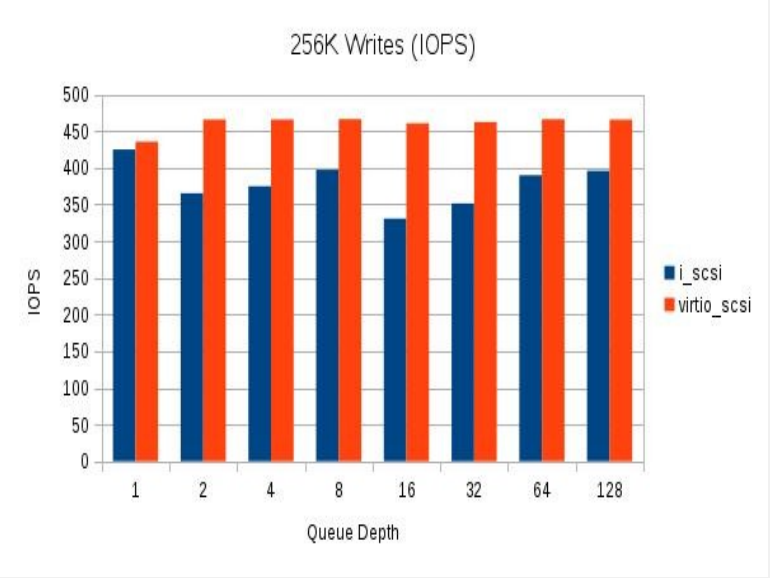




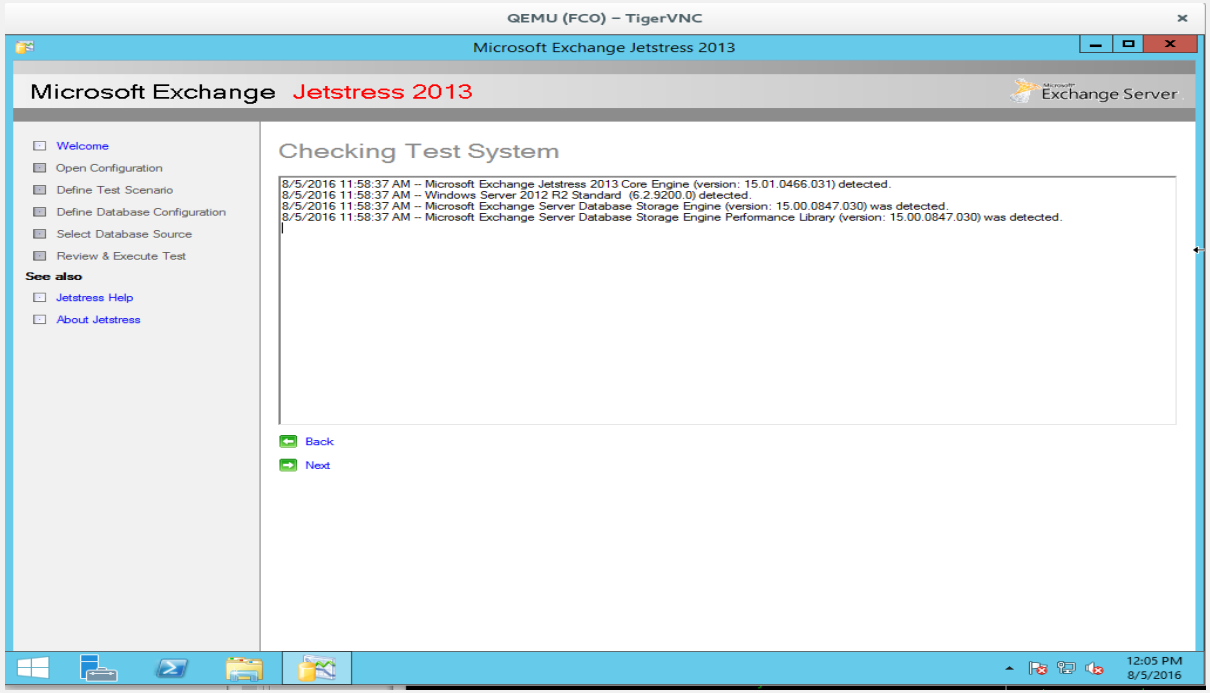
iSCSI vs. virtio-scsi performance test



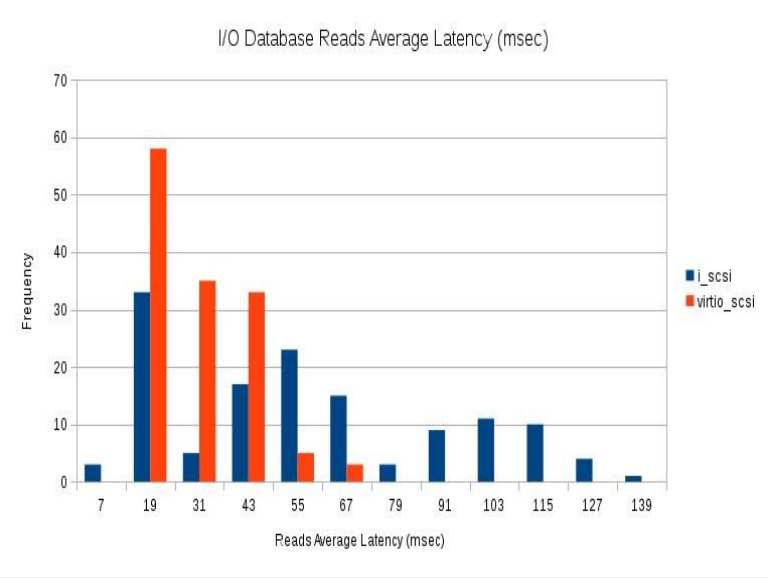
iSCSI vs. virtio-scsi performance test (cont.)

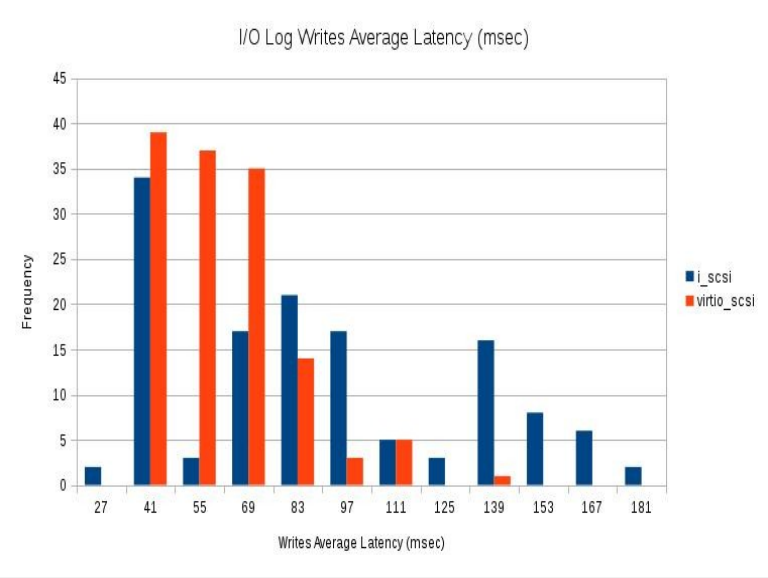


MS Exchange Jetstress

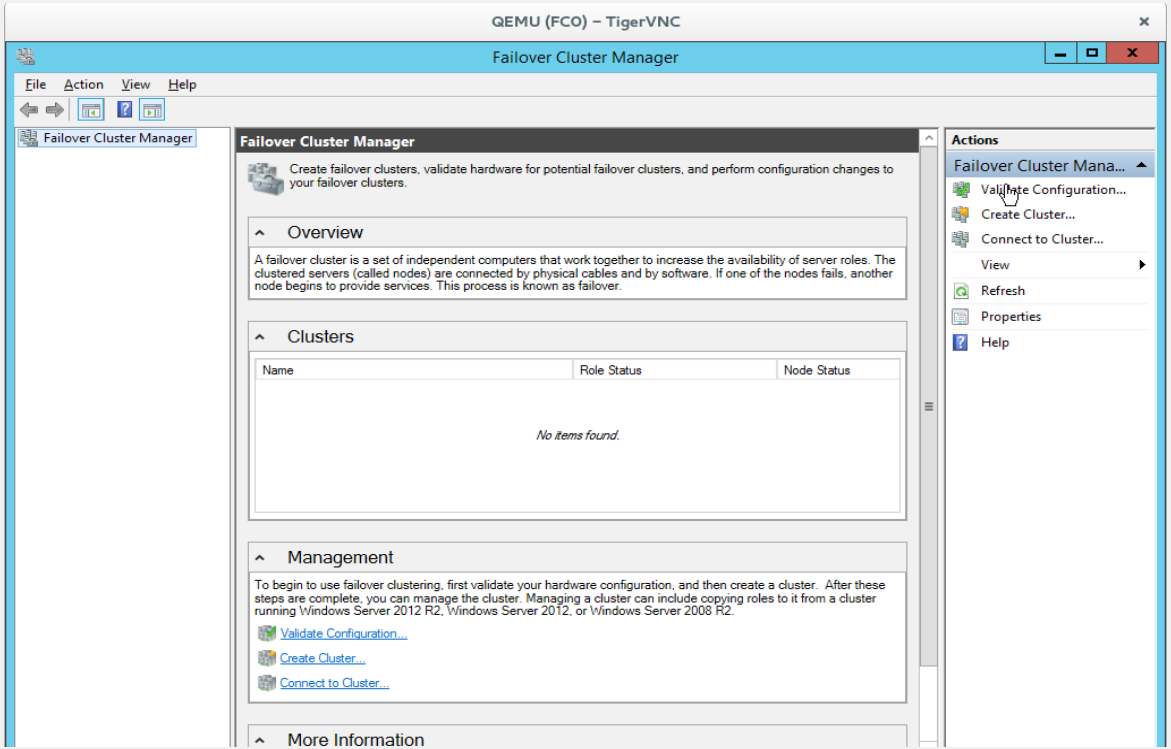


Jetstress latency results





Failover Cluster Manager



Failover Cluster Manager (cont.)

Inventory virtio-scsi

Microsoft

Failover Cluster Validation Report

Node: WIN-FC0.corp.vrozenfe.com
Started 9/4/2015 7:09:49 AM
Completed 9/4/2015 7:09:49 AM

Inventory

Name	Result	Description
List SAS Host Bus Adapters		Success

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

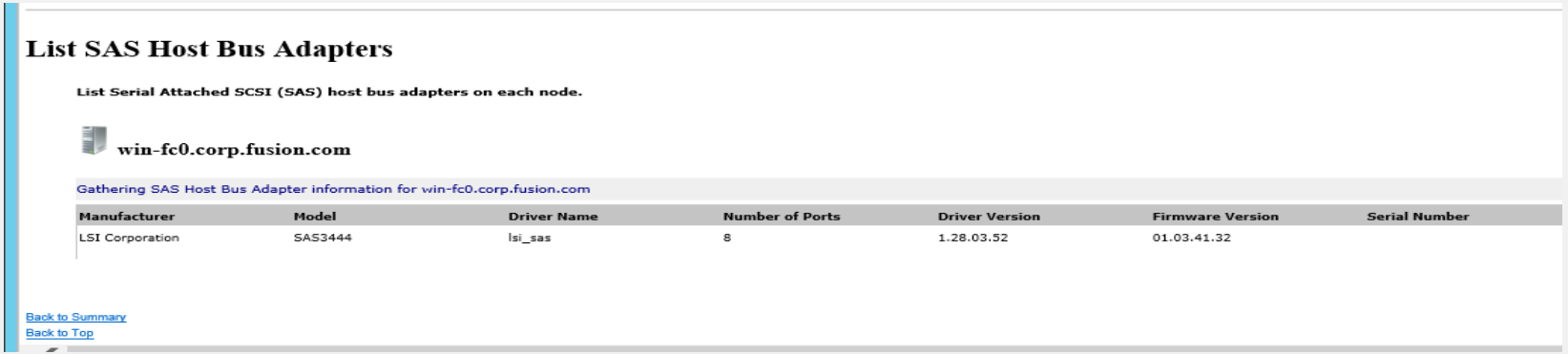
Gathering SAS Host Bus Adapter information for WIN-FC0.corp.vrozenfe.com

None found...

Back to Summary
Back to Top

Failover Cluster Manager (cont.)

Inventory lsi_sas (VMWare Fusion)



List SAS Host Bus Adapters

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

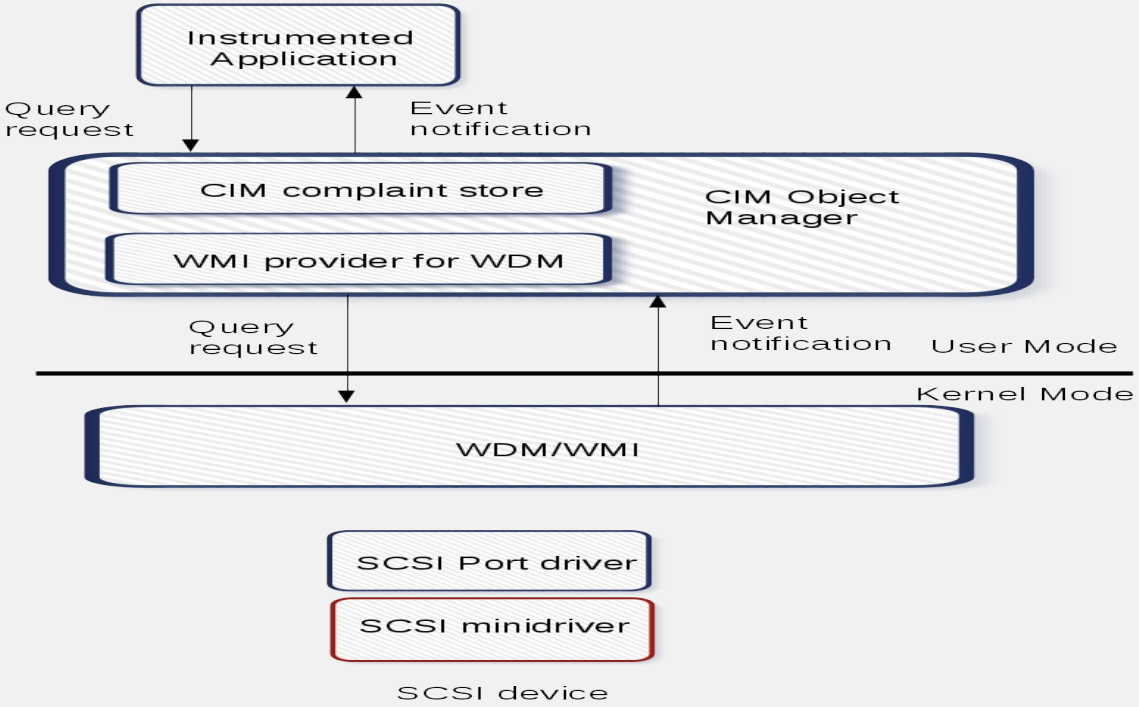
win-fc0.corp.fusion.com

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

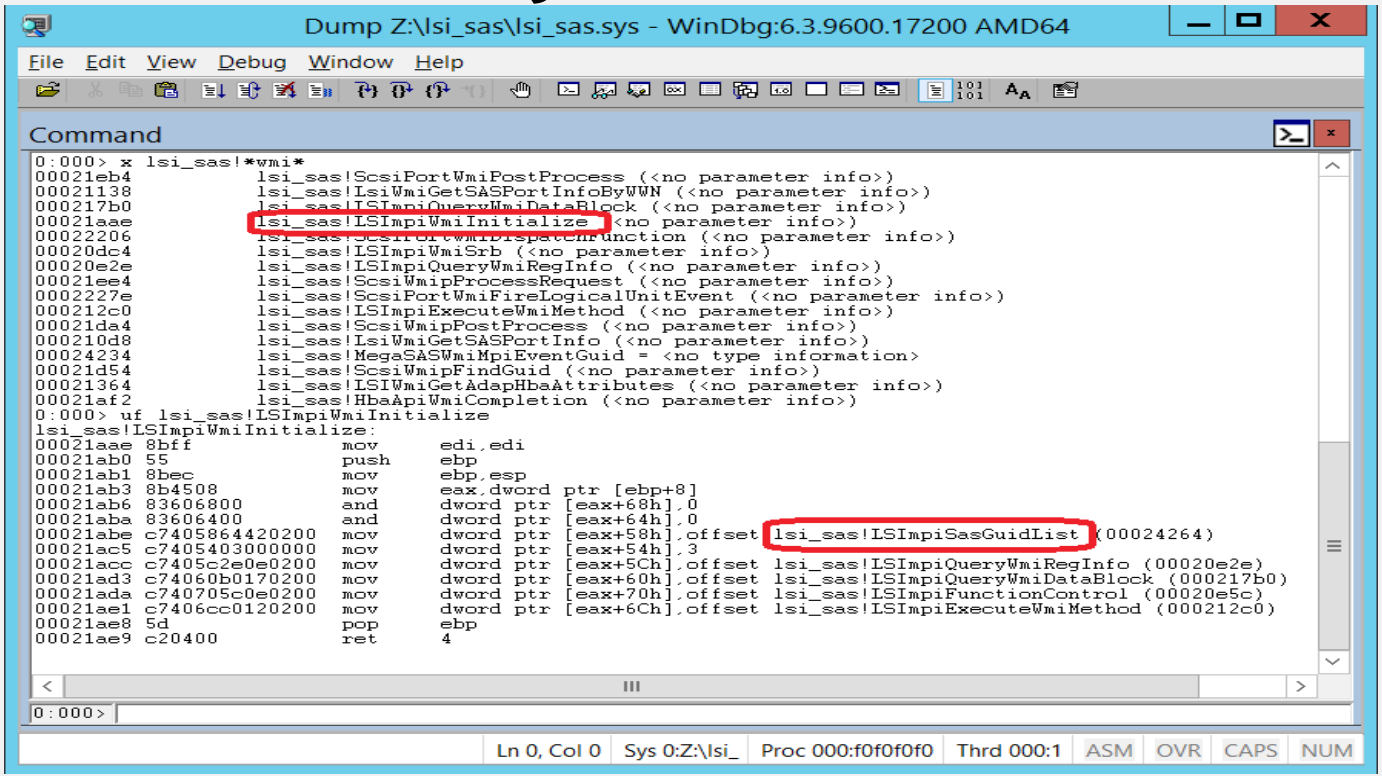
Manufacturer	Model	Driver Name	Number of Ports	Driver Version	Firmware Version	Serial Number
LSI Corporation	SAS3444	lsi_sas	8	1.28.03.52	01.03.41.32	

Back to Summary
Back to Top

Windows Management Instrumentation



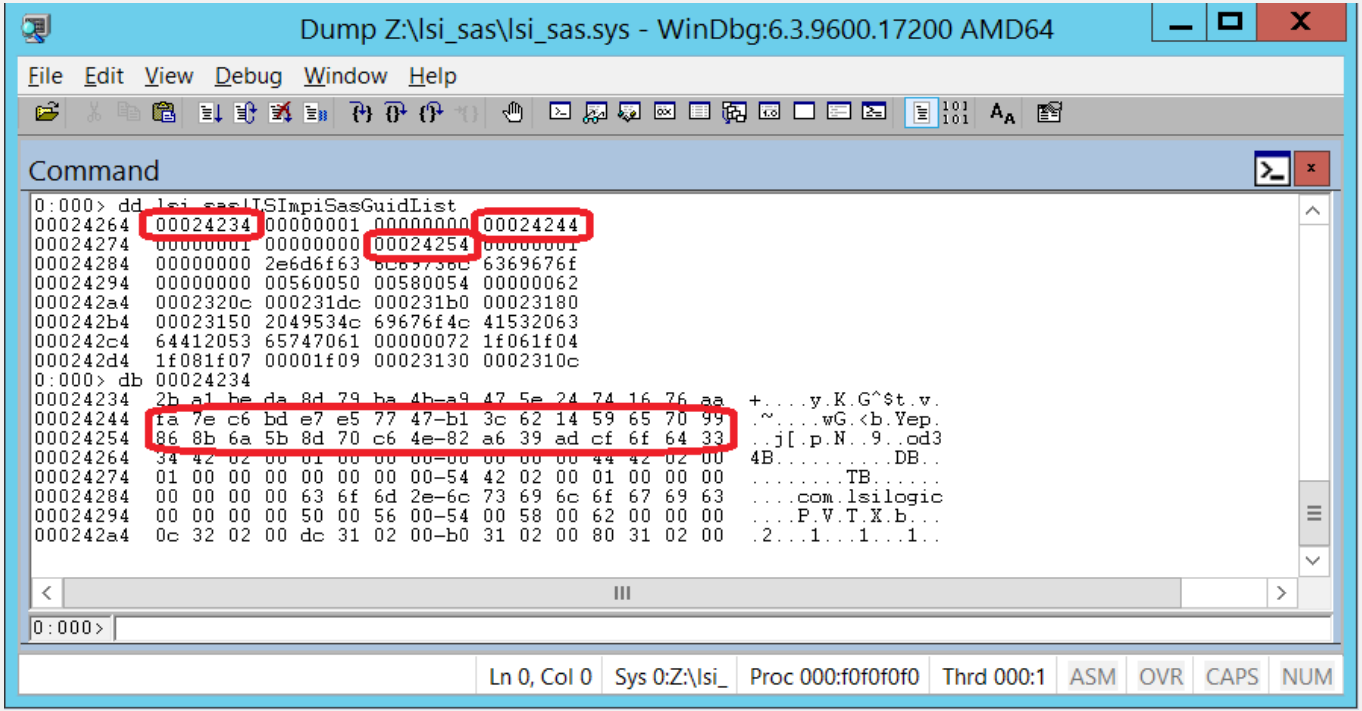
WMI discovering GUID List



WMI discovering GUID List (cont)

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

typedef struct _SCSIWMILIB_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

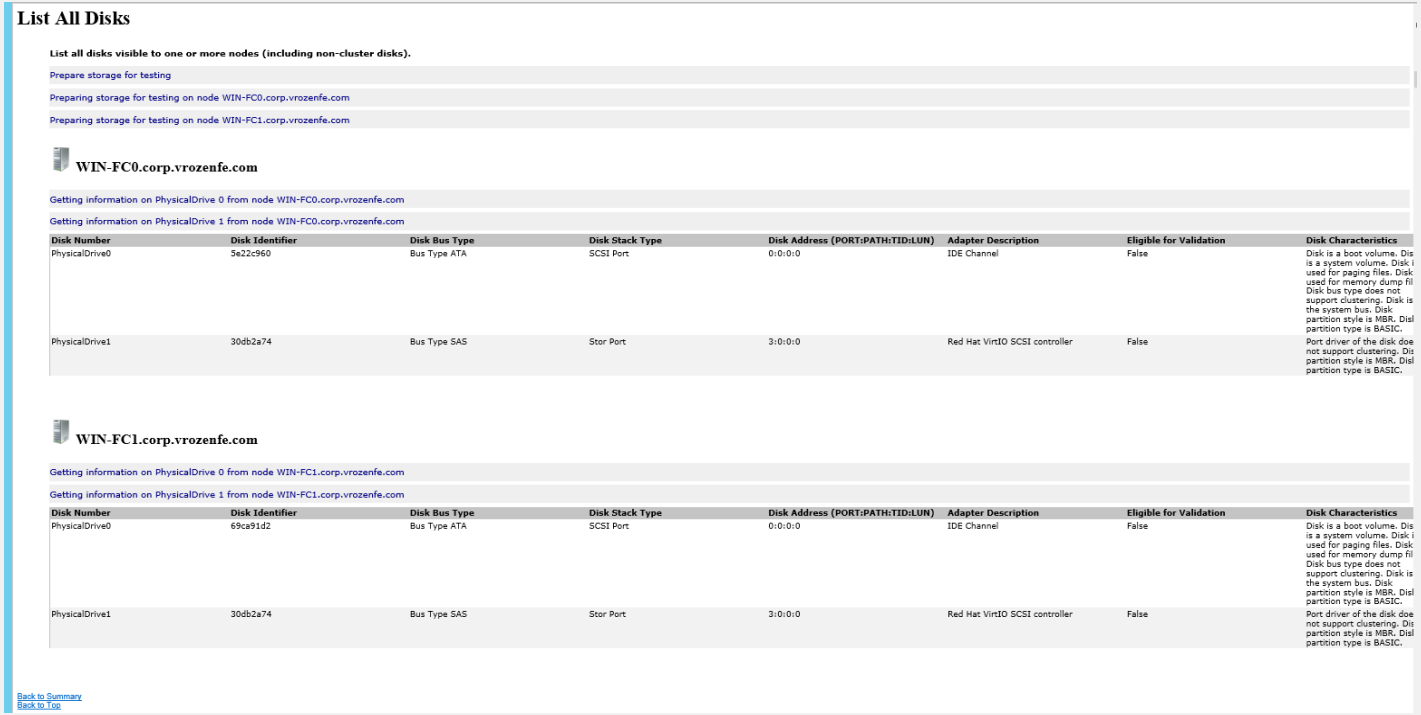


WMI discovering GUID List

```
//***************************************************************************
//
//  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



List All Disks

List all disks visible to one or more nodes (including non-cluster disks).

Prepare storage for testing

Preparing storage for testing on node WIN-FC0.corp.vrozenfe.com

Preparing storage for testing on node WIN-FC1.corp.vrozenfe.com

WIN-FC0.corp.vrozenfe.com

Getting information on PhysicalDrive 0 from node WIN-FC0.corp.vrozenfe.com

Getting information on PhysicalDrive 1 from node WIN-FC0.corp.vrozenfe.com

Disk Number	Disk Identifier	Disk Bus Type	Disk Stack Type	Disk Address (PORT:PATH:TID:LUN)	Adapter Description	Eligible for Validation	Disk Characteristics
PhysicalDrive0	5e22c960	Bus Type ATA	SCSI Port	0:0:0:0	IDE Channel	False	Disk is a boot volume. Dis is a system volume. Disk used for paging files. Disk used for memory dump fil Disk bus type does not support clustering. Disk is the system bus. Disk partition style is MBR. Disl partition type is BASIC.
PhysicalDrive1	30db2a74	Bus Type SAS	Stor Port	3:0:0:0	Red Hat VirtIO SCSI controller	False	Port driver of the disk doe not support clustering. Dis partition style is MBR. Disl partition type is BASIC.

WIN-FC1.corp.vrozenfe.com

Getting information on PhysicalDrive 0 from node WIN-FC1.corp.vrozenfe.com

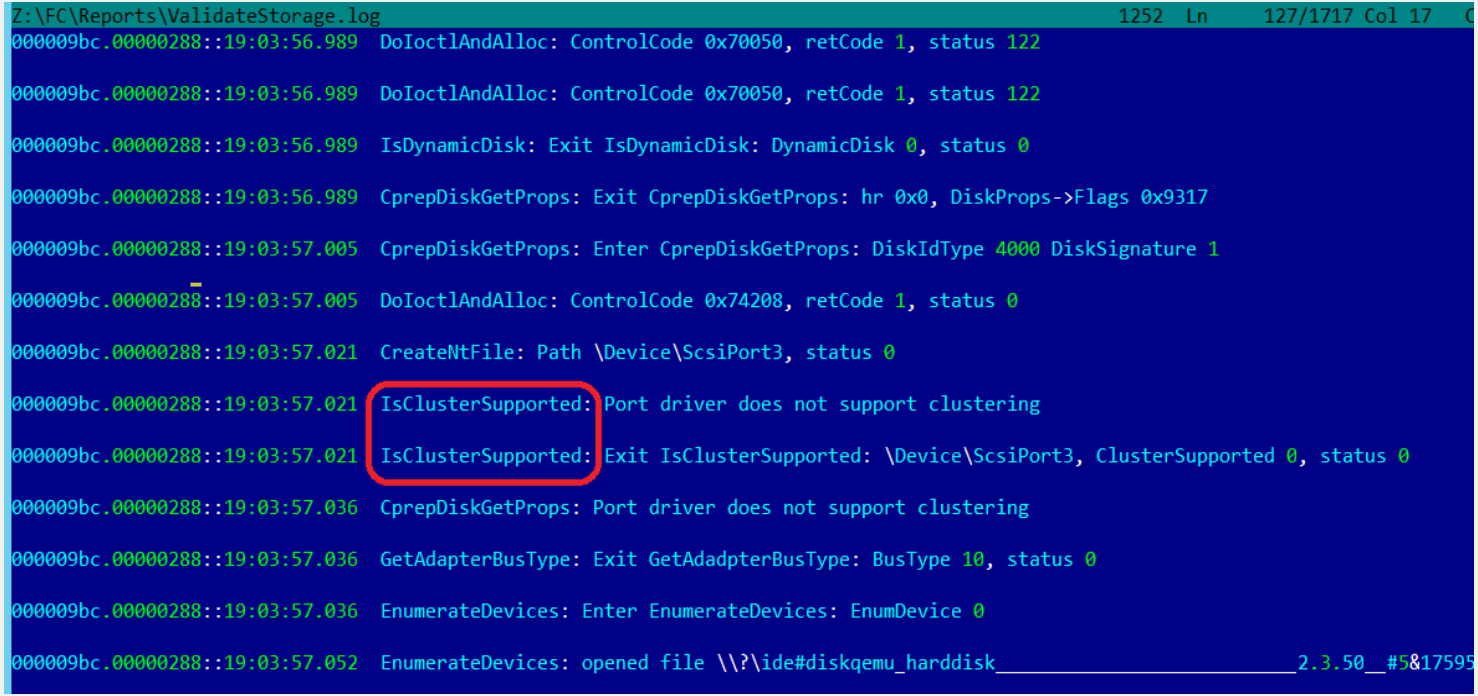
Getting information on PhysicalDrive 1 from node WIN-FC1.corp.vrozenfe.com

Disk Number	Disk Identifier	Disk Bus Type	Disk Stack Type	Disk Address (PORT:PATH:TID:LUN)	Adapter Description	Eligible for Validation	Disk Characteristics
PhysicalDrive0	69ca91d2	Bus Type ATA	SCSI Port	0:0:0:0	IDE Channel	False	Disk is a boot volume. Dis is a system volume. Disk used for paging files. Disk used for memory dump fil Disk bus type does not support clustering. Disk is the system bus. Disk partition style is MBR. Disl partition type is BASIC.
PhysicalDrive1	30db2a74	Bus Type SAS	Stor Port	3:0:0:0	Red Hat VirtIO SCSI controller	False	Port driver of the disk doe not support clustering. Dis partition style is MBR. Disl partition type is BASIC.

Back to Summary
Back to Top

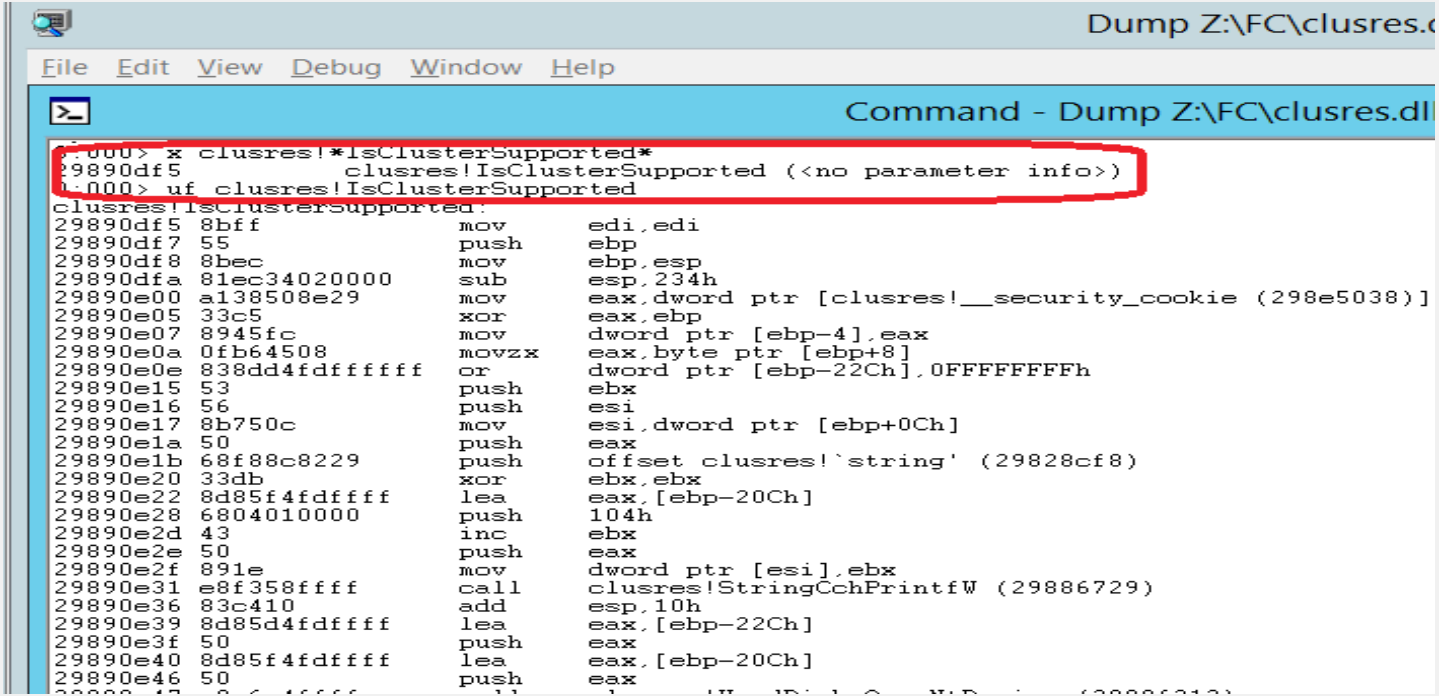
Failover Cluster Manager (cont.)

List All Disks log file



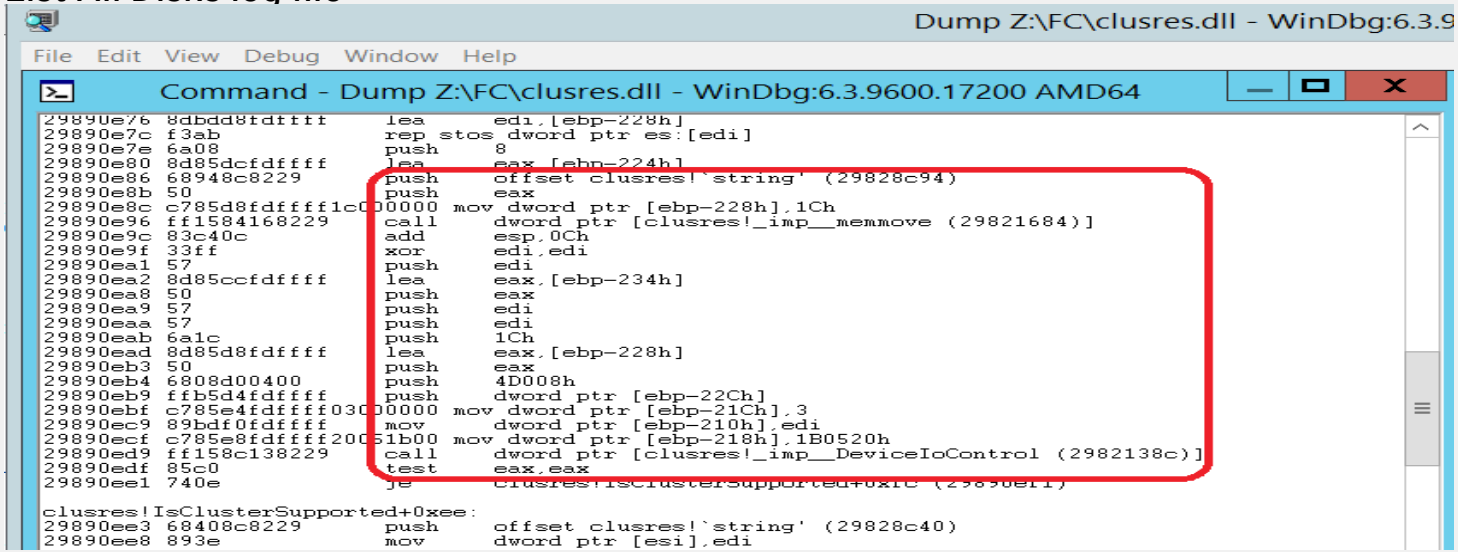
Failover Cluster Manager (cont.)

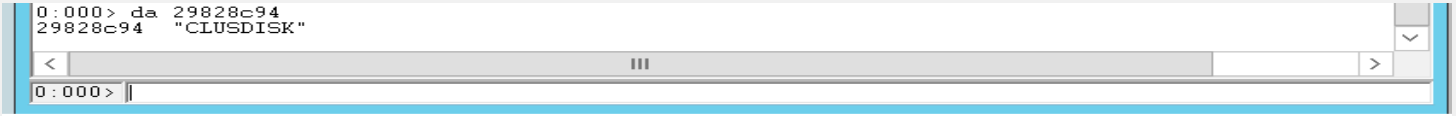
Clusters.dll



Failover Cluster Manager (cont.)

List All Disks log file





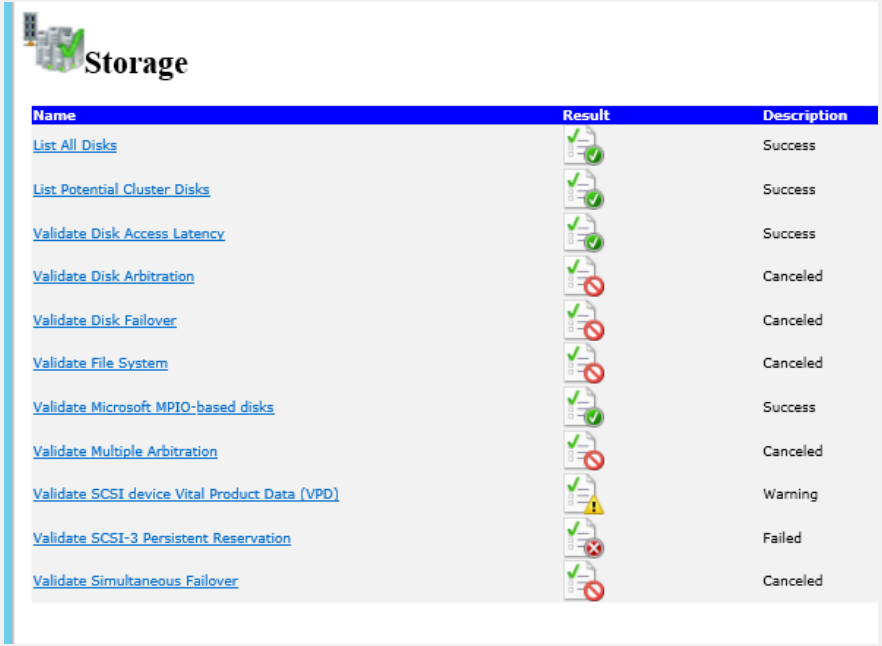
IOCTL_SCSI_MINIPORT

```
inc\api\ntddscsi.h
#define IOCTL_SCSI_MINIPORT           CTL_CODE(IOCTL_SCSI_BASE, 0x0402, METHOD_BUFFERED, FILE_READ_ACCESS | FILE_WRITE_ACCESS)


inc\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;
```

IOCTL_SCSI_MINIPORT

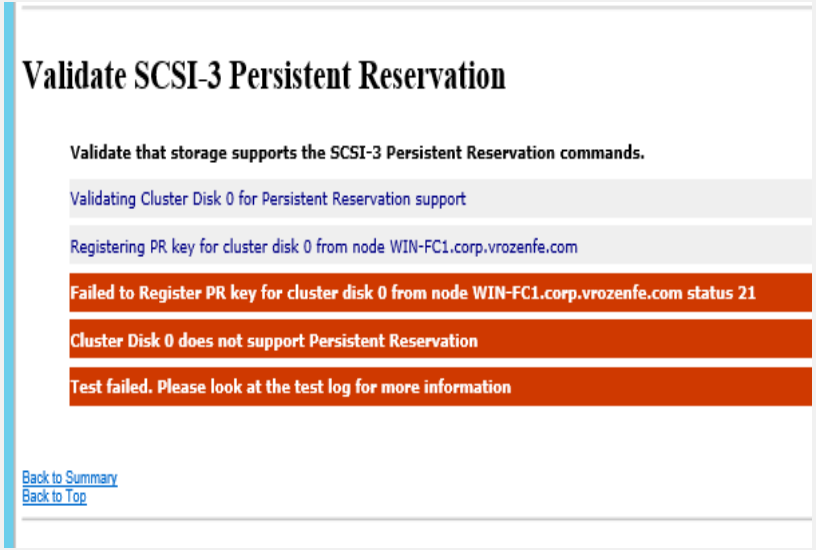
```
unsigned size = sizeof(SRB_IO_CONTROL);
SRB_IO_CONTROL srbc;
DWORD num_out;

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

Name	Result	Description
List All Disks		Success
List Potential Cluster Disks		Success
Validate Disk Access Latency		Success
Validate Disk Arbitration		Canceled
Validate Disk Failover		Canceled
Validate File System		Canceled
Validate Microsoft MPIO-based disks		Success
Validate Multiple Arbitration		Canceled
Validate SCSI device Vital Product Data (VPD)		Warning
Validate SCSI-3 Persistent Reservation		Failed
Validate Simultaneous Failover		Canceled



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

Back to Summary
Back to Top

QEMU – always use SG_IO

commit 8fdc7839e40f43a426bc7e858cf1dbfe315a3804
Author: Paolo Bonzini <pbonzini@redhat.com>
Date: Tue May 10 10:50:44 2016 +0200

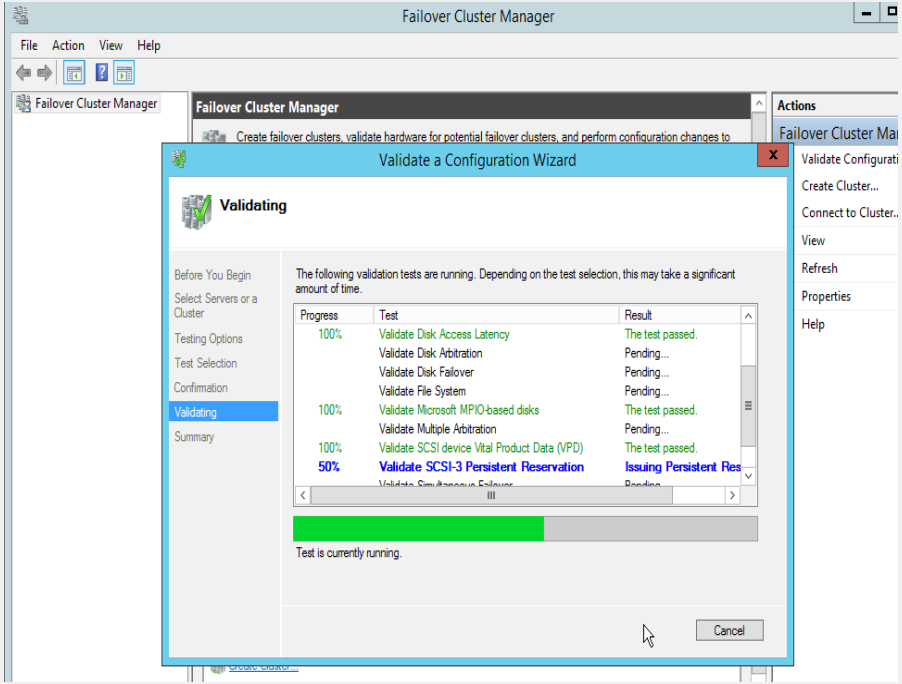
    scsi-block: always use SG_IO

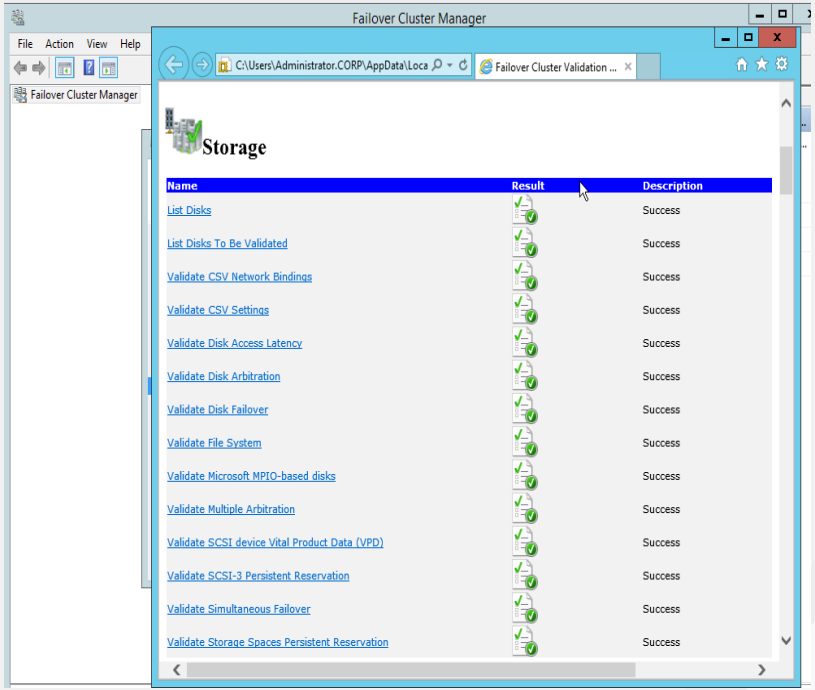
    Using pread/pwrite or io_submit has the advantage of eliminating the
    bounce buffer, but drops the SCSI status. This keeps the guest from
    seeing unit attention codes, as well as statuses such as RESERVATION
    CONFLICT. Because we know scsi-block operates on an SBC device we can
    still use 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





redhat.

THANK YOU

plus.google.com/+RedHat

facebook.com/redhatinc

linkedin.com/company/red-hat

twitter.com/RedHatNews

youtube.com/user/RedHatVideos