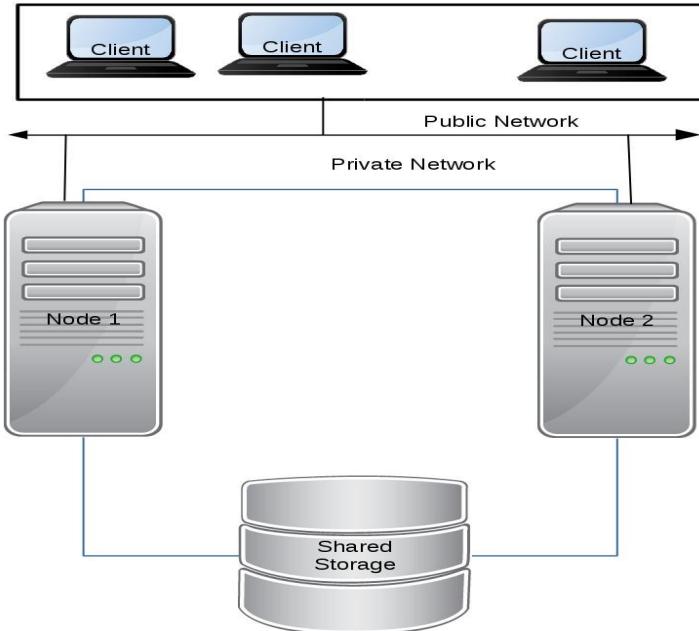




# MS Cluster on KVM

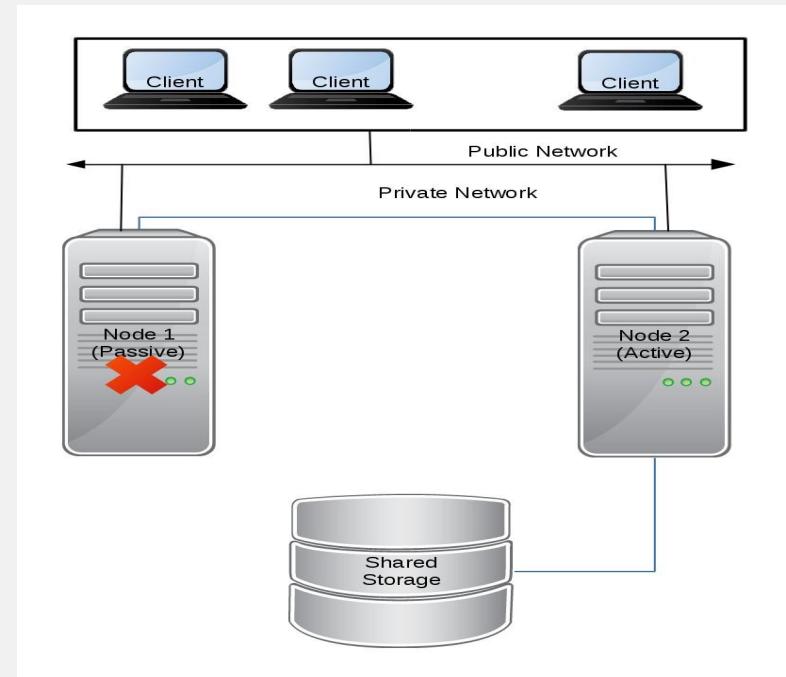
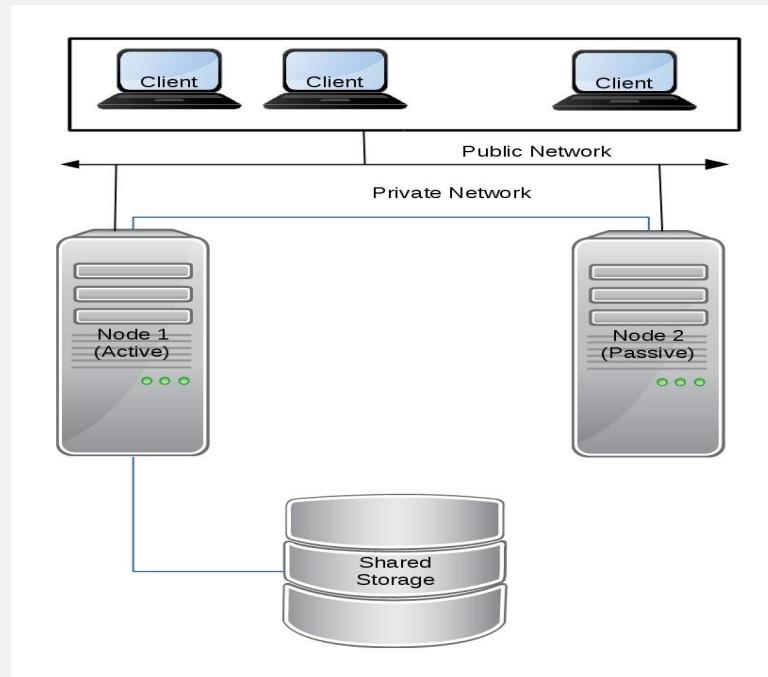
Vadim Rozenfeld  
[vrozenfe@redhat.com](mailto:vrozenfe@redhat.com)  
25 Aug, 2016

# Cluster: Servers Combined to Improve Availability and Scalability.

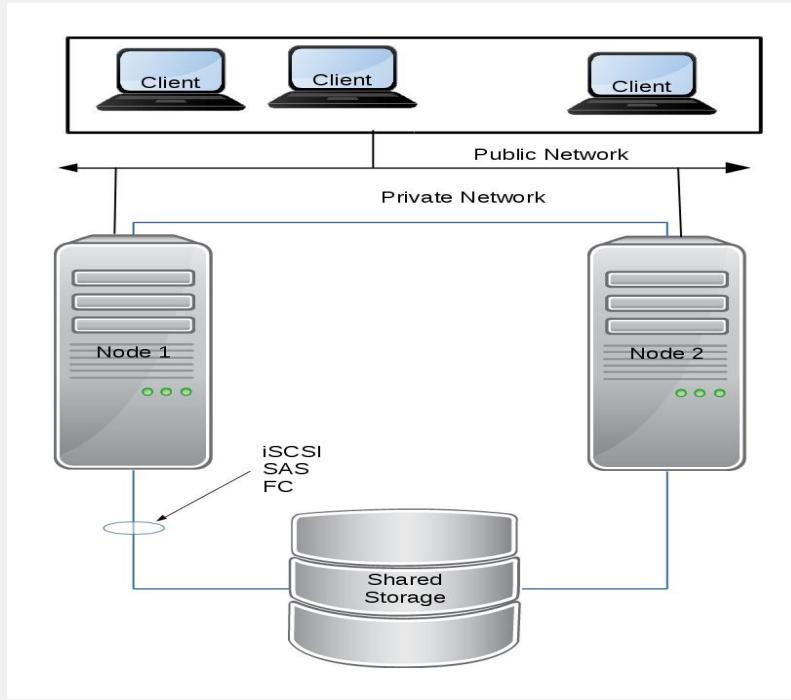


- Cluster: A group of independent systems working together as a single system. Clients see scalable and fault tolerance service.
- Node: A server in a cluster.
- Interconnect: Communication link used for intra-cluster status info such as “heartbeats”.

# Failover Cluster



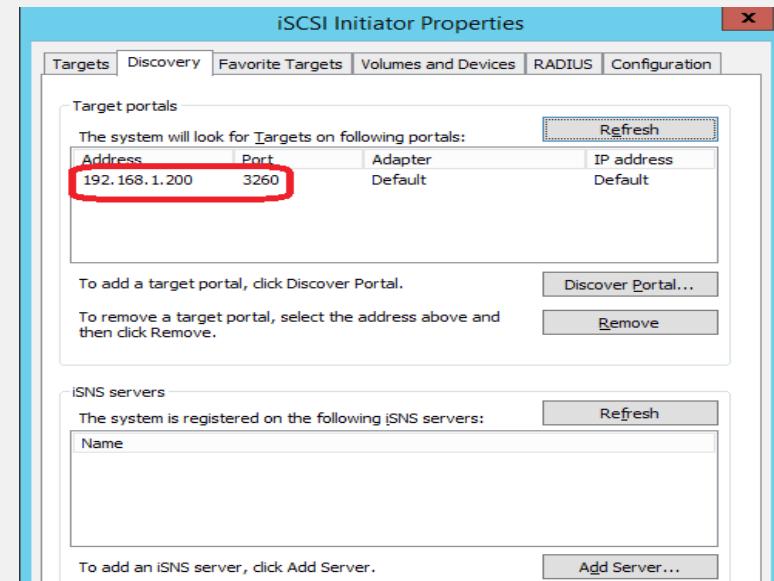
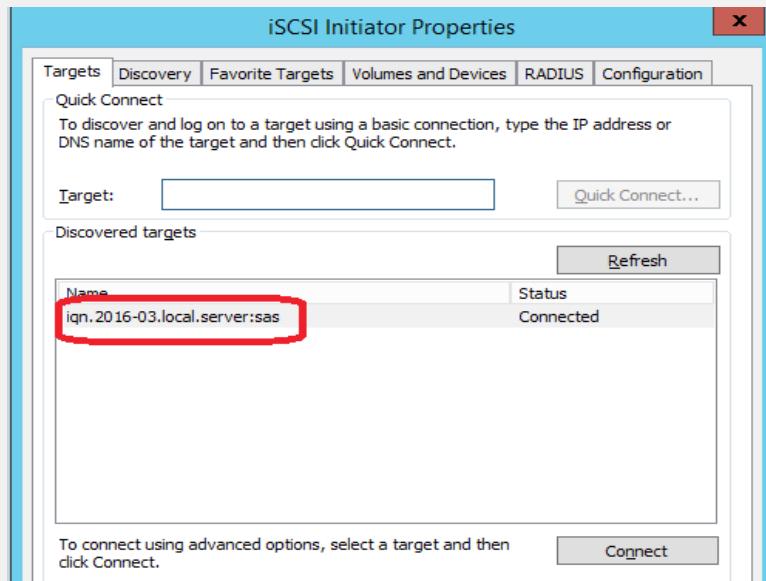
# Cluster storage



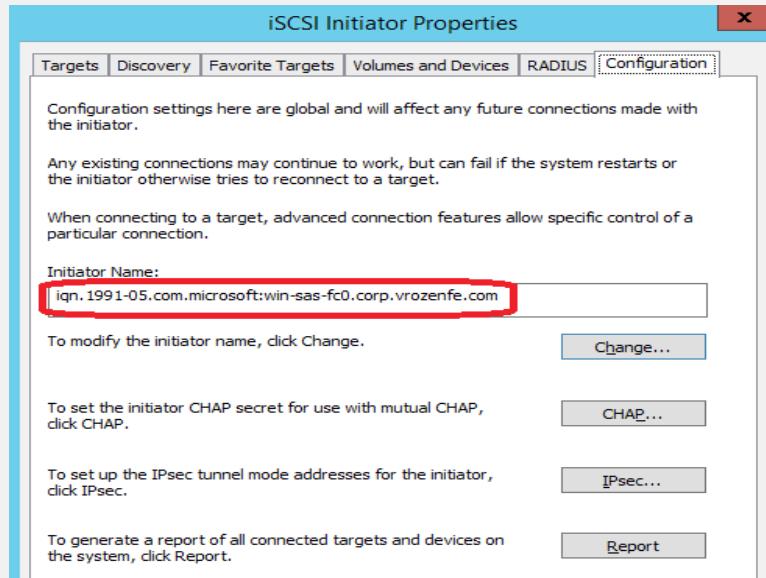
## Hardware requirements :

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

# iSCSI



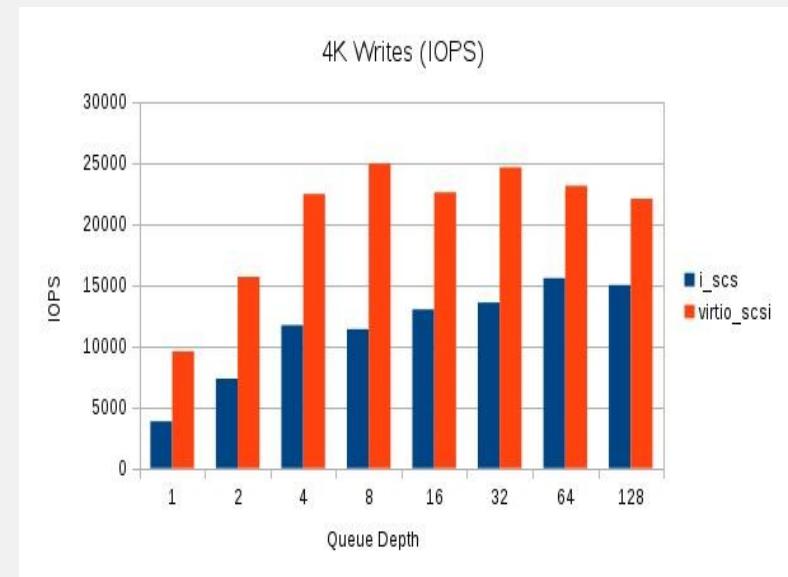
# iSCSI (cont)



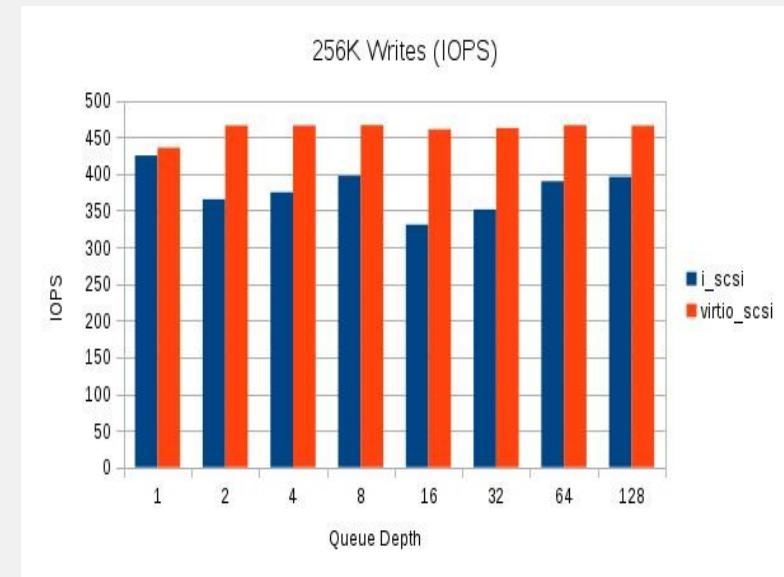
```
yrozenfe@jack ~]$ sudo targetcli
targetcli shell version 2.1.fb42
Copyright 2011-2013 by Datera, Inc and others.
For help on commands, type 'help'.

> ls
+ / .....
o backstores
| o block .....
| o fileio .....
| | o disk01 ..... [/home/vrozenfe/work/Images/disk01.img (10.0GiB) write-back activated]
| | o disk02 ..... [/home/vrozenfe/work/Images/disk02.img (10.0GiB) write-back activated]
o pscsi .....
o ramdisk .....
o iscsi .....
o iqn.2016-03.local.server:sas .....
o tgtl .....
o acls .....
| o iqn.1991-05.com.microsoft:fc0.corp.vrozenfe.com .....
| | o mapped_lun0 .....
| | o iqn.1991-05.com.microsoft:fcl.corp.vrozenfe.com .....
| | o mapped_lun0 .....
| | o iqn.1991-05.com.microsoft:win-sas-fc0.corp.vrozenfe.com .....
| | | o mapped_lun0 .....
| | | o iqn.1994-05.com.redhat:696b50ffa3d0 .....
| | | o mapped_lun0 .....
| | | o iqn.2008-11.org.Linux-kvm:5b959af7-e33f-4229-97b4-da6fe8fb7062 .....
| | | o mapped_lun0 .....
| | | o iqn.2008-11.org.Linux-kvm:5b959af7-e33f-4229-97b4-da6fe8fb7062 .....
| | | o mapped_lun0 .....
```

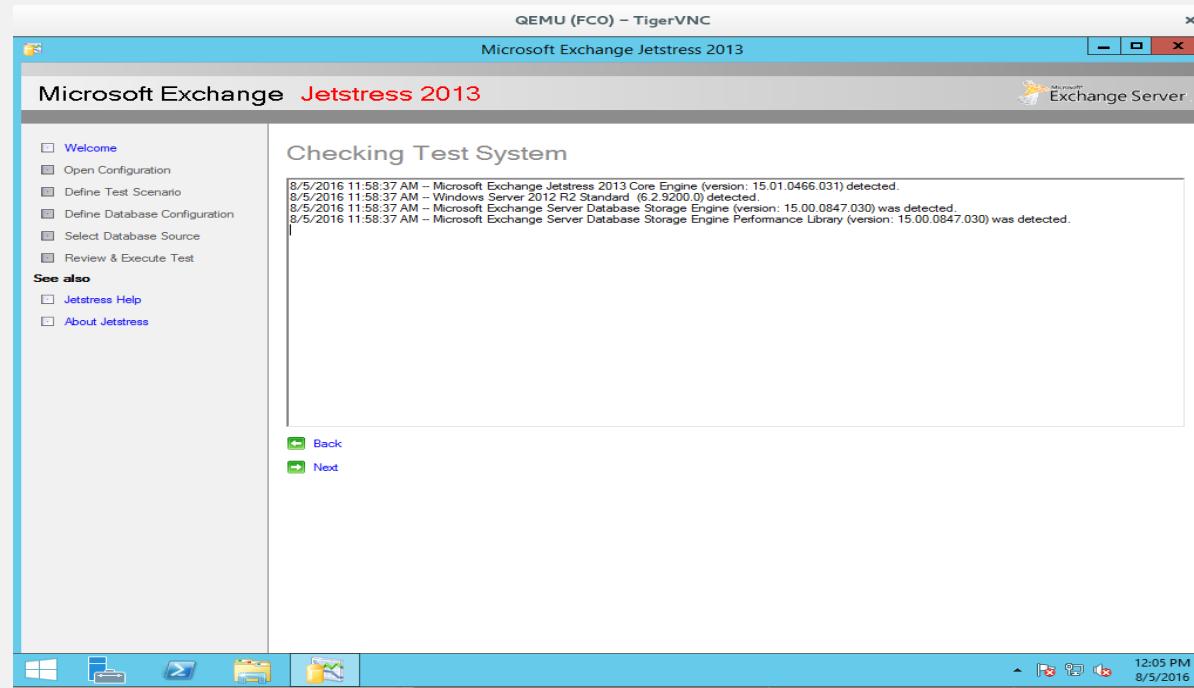
# iSCSI vs. virtio-scsi performance test



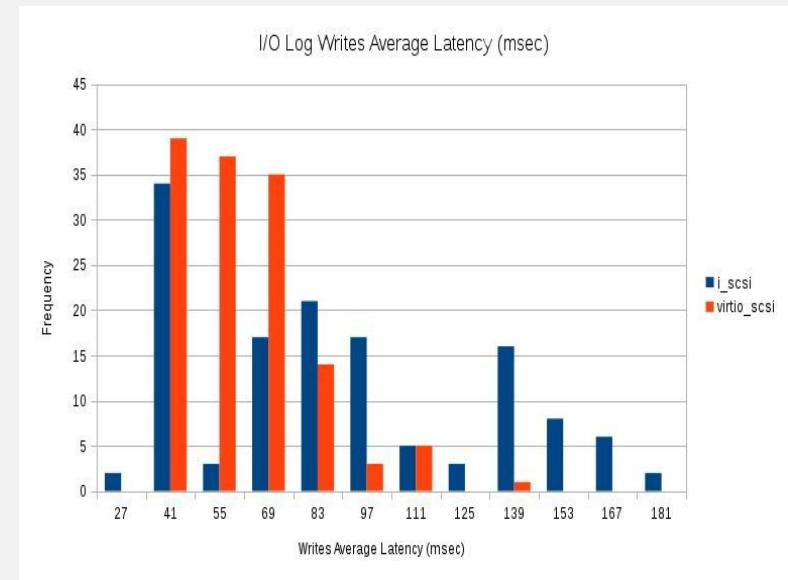
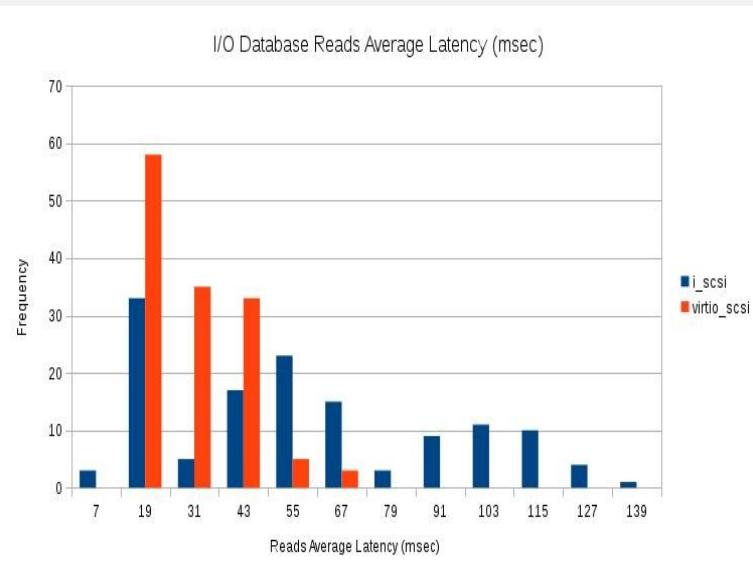
# iSCSI vs. virtio-scsi performance test (cont.)



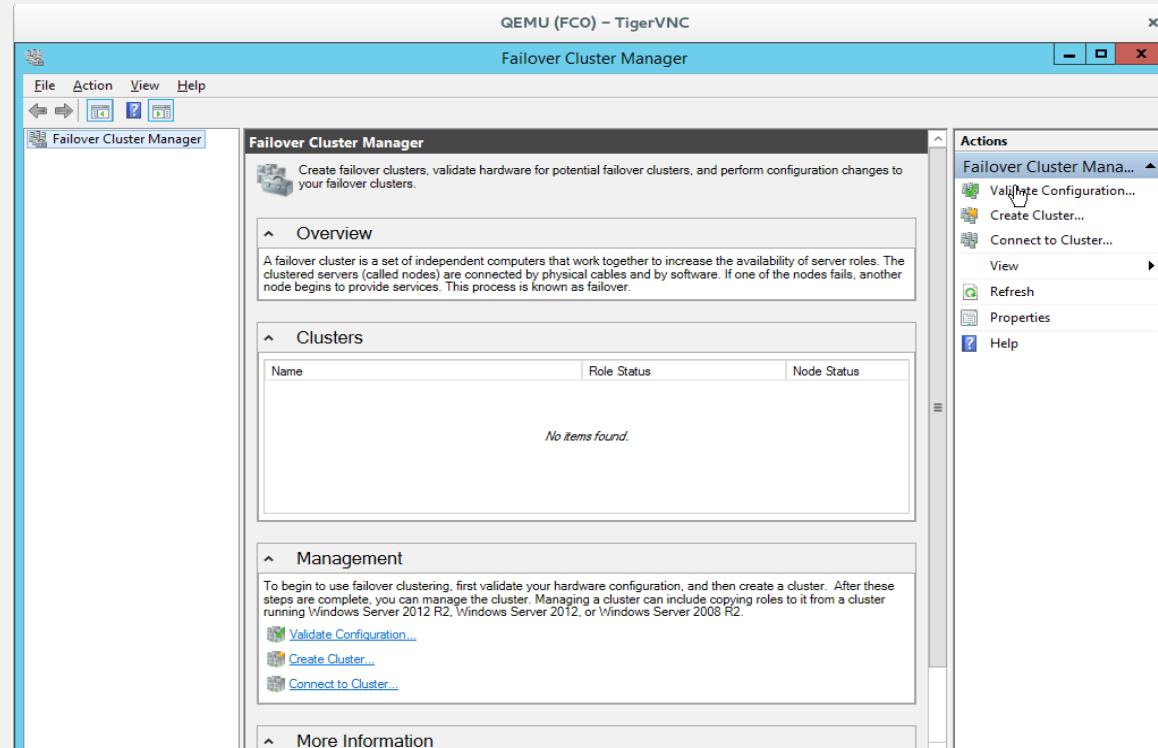
# MS Exchange Jetstress



# 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 Isi\_sas (VMWare Fusion)

**List SAS Host Bus Adapters**

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

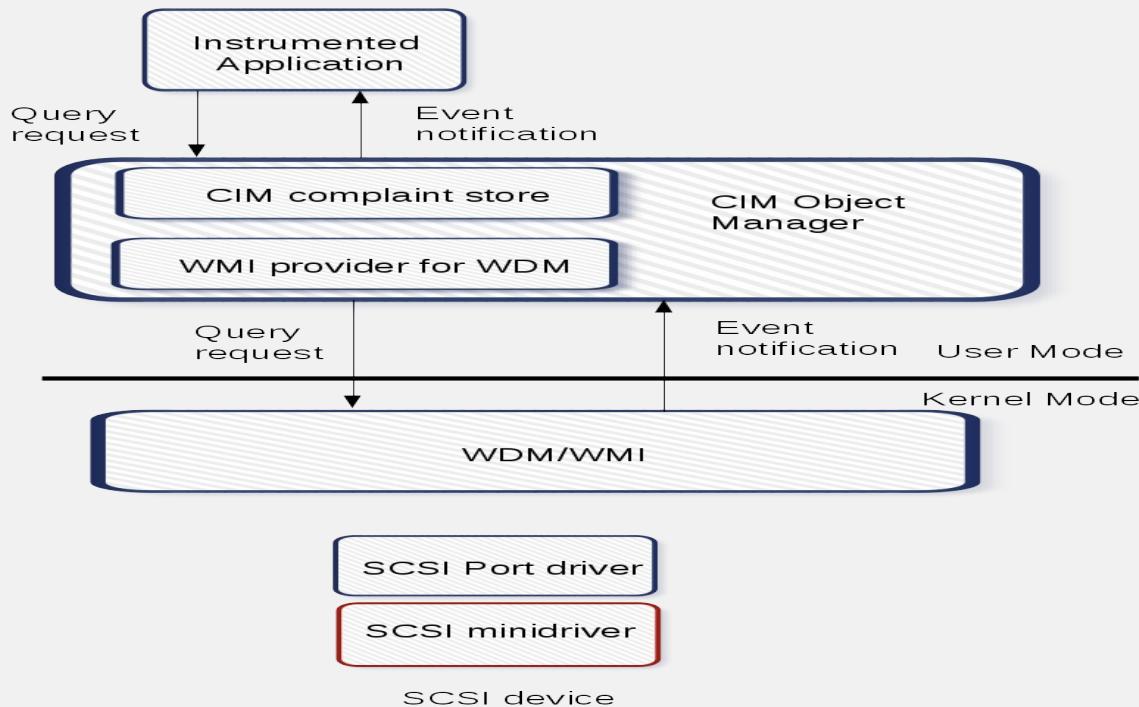
 win-fc0.corp.fusion.com

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

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

[Back to Summary](#)  
[Back to Top](#)

# Windows Management Instrumentation



# WMI discovering GUID List

Dump Z:\lsi\_sas\lsi\_sas.sys - WinDbg:6.3.9600.17200 AMD64

File Edit View Debug Window Help

Command

```
0:000> x lsi_sas!*wmi*
00021eb4 lsi_sas!ScsiPortWmiPostProcess (<no parameter info>)
00021138 lsi_sas!LsiWmiGetSASPortInfoByWWN (<no parameter info>)
000217b0 lsi_sas!LSIImpiQueryWmiDataBlock (<no parameter info>)
00021aae lsi_sas!LSIImpiWmiInitialize (<no parameter info>)
00022206 lsi_sas!ScsiWmiFunctionControlFunction (<no parameter info>)
00020dc4 lsi_sas!LSIImpiWmiSrb (<no parameter info>)
00020e2e lsi_sas!LSIImpiQueryWmiRegInfo (<no parameter info>)
00021ee4 lsi_sas!ScsiWmiPProcessRequest (<no parameter info>)
0002227e lsi_sas!ScsiPortWmiFireLogicalUnitEvent (<no parameter info>)
000212c0 lsi_sas!LSIImpiExecuteWmiMethod (<no parameter info>)
00021da4 lsi_sas!ScsiWmiPostProcess (<no parameter info>)
000210d8 lsi_sas!LsiWmiGetSASPortInfo (<no parameter info>)
00024234 lsi_sas!MegaSASWmiMpEventGuid = <no type information>
00021d54 lsi_sas!ScsiWmiFindGuid (<no parameter info>)
00021364 lsi_sas!LSIWmiGetAdapHbaAttributes (<no parameter info>)
00021af2 lsi_sas!HbaApiWmiCompletion (<no parameter info>)
0:000> uf lsi_sas!LSIImpiWmiInitialize
lsi_sas!LSIImpiWmiInitialize:
00021aae 8bff    mov     edi,edi
00021ab0 55      push    ebp
00021ab1 8bec   pop    ebp
00021ab3 8b4508  mov     eax,dword ptr [ebp+8]
00021ab5 83606800 and    dword ptr [eax+68h],0
00021ab7 83606400 and    dword ptr [eax+64h],0
00021abe c7405864420200 mov     dword ptr [eax+58h],offset lsi_sas!LSIImpiSasGuidList (00024264)
00021ac1 c7405403000000 mov     dword ptr [eax+54h],3
00021acc c7405c2e0e0200 mov     dword ptr [eax+5Ch],offset lsi_sas!LSIImpiQueryWmiRegInfo (00020e2e)
00021ad3 c74060b170200 mov     dword ptr [eax+60h],offset lsi_sas!LSIImpiDataBlock (000217b0)
00021ada c740705c0e0200 mov     dword ptr [eax+70h],offset lsi_sas!LSIImpiFunctionControl (00020e5c)
00021ael c7406cc0120200 mov     dword ptr [eax+6Ch],offset lsi_sas!LSIImpiExecuteWmiMethod (000212c0)
00021ae8 5d      pop    ebp
00021ae9 c20400  ret    4

< << >> >
```

Ln 0, Col 0 Sys 0:Z:\lsi\_ Proc 000:f0f0f0f0 Thrd 000:1 ASM OVR CAPS NUM

# 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 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

```
Dump Z:\lsi_sas\lsi_sas.sys - WinDbg:6.3.9600.17200 AMD64
File Edit View Debug Window Help
Command
0:000> dd lsi_sas!TSImpiSasGuidList
00024264 00024234 00000001 00000000 00024244
00024274 00000001 00000000 00024254 00000001
00024284 00000000 2e6d6f63 6c89738c 6369676f
00024294 00000000 00560050 00580054 00000062
000242a4 0002320c 000231dc 000231b0 00023180
000242b4 00023150 2049534c 69676f4c 41532063
000242c4 64412053 65747061 00000072 1f061f04
000242d4 1f081f07 00001f09 00023130 0002310c
0:000> db 00024234
00024234 2b a1 b6 da 8d 79 ba 4b-a9 47 5e 24 74 16 76 aa +....y.K.G^$t.v.
00024244 fa 7e c6 bd e7 e5 77 47-b1 3c 62 14 59 65 70 99 .~...wG.<b.Yep.
00024254 86 8b 6a 5b 70 c6 4e-82 a6 39 ad cf 6f 64 33 ..j[.p.N..9..od3
00024264 34 42 02 00 01 00 00 00-00 00 00 44 42 02 00 4B.....DB..
00024274 01 00 00 00 00 00 00 00-54 42 02 00 01 00 00 00 .....TB....
00024284 00 00 00 00 63 6f 6d 2e-6c 73 69 6c 6f 67 69 63 ....com.lsilogic
00024294 00 00 00 00 50 00 56 00-54 00 58 00 62 00 00 00 ...P.V.T.X.b...
000242a4 0c 32 02 00 dc 31 02 00-b0 31 02 00 80 31 02 00 2...1...1...1...
```

Ln 0, Col 0 Sys 0:Z:\lsi\_ Proc 000:f0f0f0f0 Thrd 000:1 ASM OVR CAPS NUM

# 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. Disk is a system volume. Disk is 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. Disk 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 does not support clustering. Disk partition style is MBR. Disk 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. Disk is a system volume. Disk is 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. Disk 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 does not support clustering. Disk partition style is MBR. Disk partition type is BASIC.

[Back to Summary](#)

[Back to Top](#)

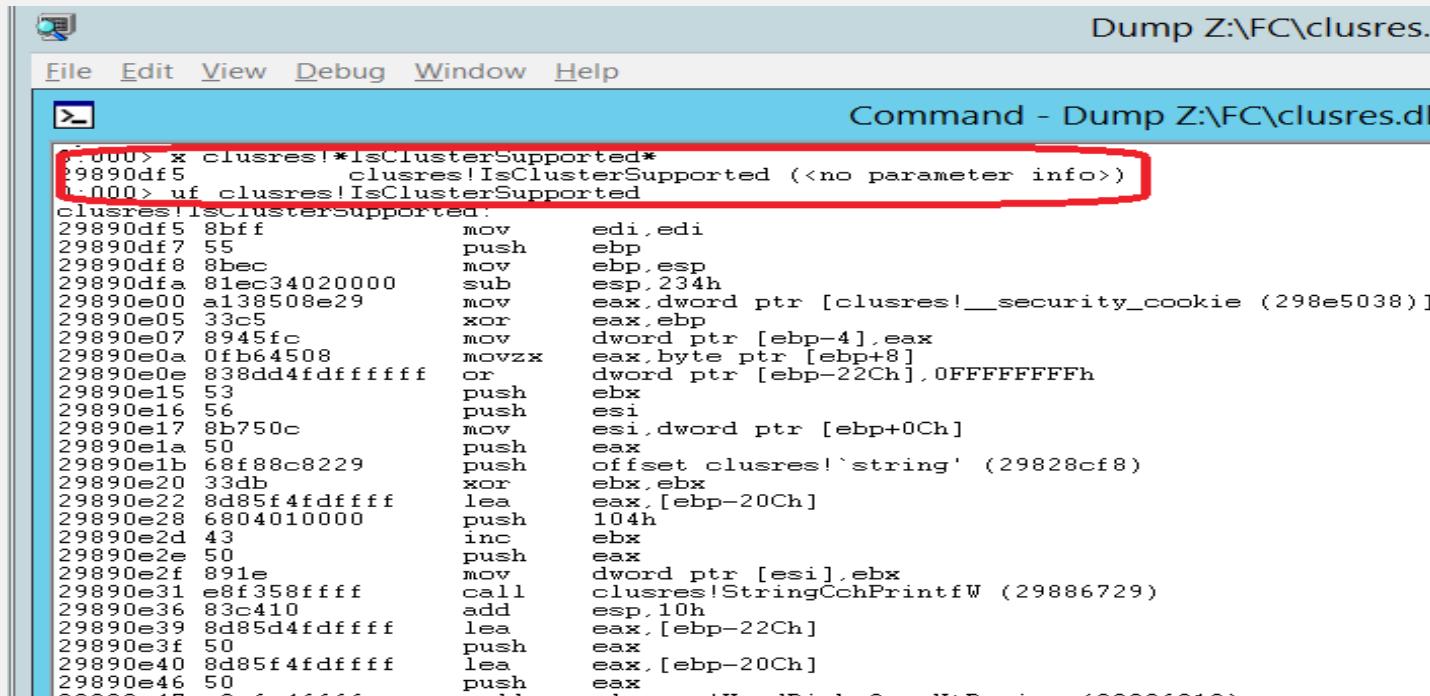
# Failover Cluster Manager (cont.)

## List All Disks log file

```
Z:\FC\Reports\ValidateStorage.log          1252 Ln      127/1717 Col 17 C
000009bc.00000288::19:03:56.989  DoIoctlAndAlloc: ControlCode 0x70050, retCode 1, status 122
000009bc.00000288::19:03:56.989  DoIoctlAndAlloc: ControlCode 0x70050, retCode 1, status 122
000009bc.00000288::19:03:56.989  IsDynamicDisk: Exit IsDynamicDisk: DynamicDisk 0, status 0
000009bc.00000288::19:03:56.989  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 GetAdadpterBusType: 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_#5&17595
```

# Failover Cluster Manager (cont.)

## Clusters.dll

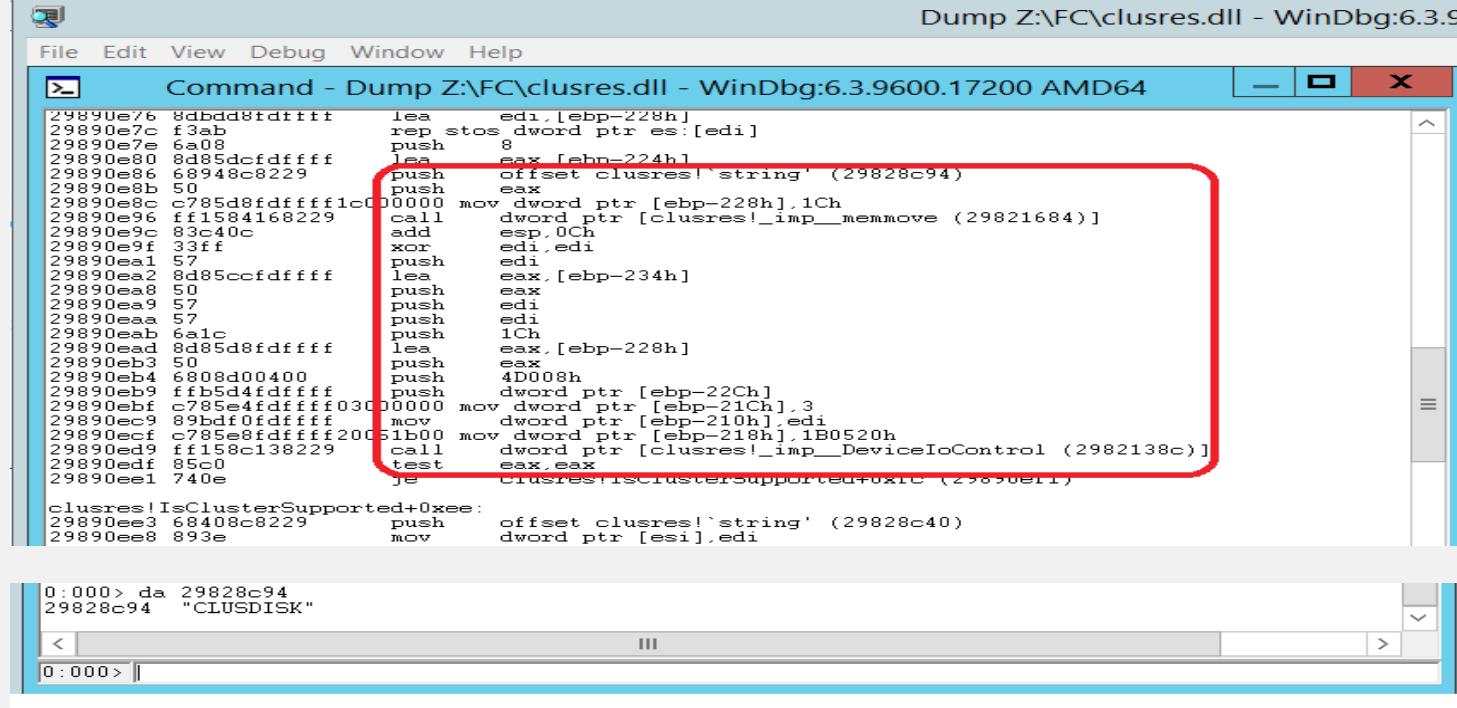


The screenshot shows a debugger interface with the title "Command - Dump Z:\FC\clusres.dll". The menu bar includes File, Edit, View, Debug, Window, and Help. The assembly code window displays the following:

```
0000> x clusres!*IsClusterSupported*
29890df5         clusres!IsClusterSupported (<no parameter info>)
1.000> uf clusres!IsClusterSupported
clusres!IsClusterSupported:
29890df5 8bff      mov     edi,edi
29890df7 55        push    ebp
29890df8 8bec      mov     ebp,esp
29890dfa 81ec34020000 sub    esp,234h
29890e00 a138508e29 mov    eax,dword ptr [clusres!__security_cookie (298e5038)]
29890e05 33c5      xor    eax,ebp
29890e07 8945fc      mov    dword ptr [ebp-4],eax
29890e0a 0fb64508 movzx  eax,byte ptr [ebp+8]
29890e0e 838dd4fdfffff or    dword ptr [ebp-22Ch],0FFFFFFFh
29890e15 53        push    ebx
29890e16 56        push    esi
29890e17 8b750c      mov    esi,dword ptr [ebp+0Ch]
29890e1a 50        push    eax
29890e1b 68f88c8229 push    offset clusres!`string' (29828cf8)
29890e20 33db      xor    ebx,ebx
29890e22 8d85f4fdffff lea    eax,[ebp-20Ch]
29890e28 6804010000 push    104h
29890e2d 43        inc    ebx
29890e2e 50        push    eax
29890e2f 891e      mov    dword ptr [esi].ebx
29890e31 e8f358ffff call   clusres!StringCchPrintfW (29886729)
29890e36 83c410      add    esp,10h
29890e39 8d85d4fdffff lea    eax,[ebp-22Ch]
29890e3f 50        push    eax
29890e40 8d85f4fdffff lea    eax,[ebp-20Ch]
29890e46 50        push    eax
```

# Failover Cluster Manager (cont.)

## List All Disks log file



```
Dump Z:\FC\clusres.dll - WinDbg:6.3.9600.17200 AMD64
File Edit View Debug Window Help
Command - Dump Z:\FC\clusres.dll - WinDbg:6.3.9600.17200 AMD64
29890e76 8dbdd8fdfffff lea    edi,[ebp-228h]
29890e7c f3ab      rep stos dword ptr es:[edi]
29890e7e 6a08      push   8
29890e80 8d85dcfdfffff lea    eax,[ebp-224h]
29890e86 68948c8229 push   offset clusres!`string' (29828c94)
29890e8b 50        push   eax
29890e8c c785d8fdfffff1c00000 mov   dword ptr [ebp-228h],1Ch
29890e96 ff1584168229 call  dword ptr [clusres!_imp__memmove (29821684)]
29890e9c 83c40c      add    esp,0Ch
29890e9f 33ff      xor    edi,edi
29890ea1 57        push   edi
29890ea2 8d85ccfdfffff lea    eax,[ebp-234h]
29890ea8 50        push   eax
29890ea9 57        push   edi
29890eaa 57        push   edi
29890eab 6a1c      push   1Ch
29890ead 8d85d8fdfffff lea    eax,[ebp-228h]
29890eb3 50        push   eax
29890eb4 6808d00400 push   4D008h
29890eb9 ffbb5d4fdfffff push   dword ptr [ebp-22Ch]
29890ebf c785e4fdfffff0300000 mov   dword ptr [ebp-21Ch],3
29890ec9 89bdf0fdfffff mov    dword ptr [ebp-210h],edi
29890ecf c785e8fdfffff20051b000 mov   dword ptr [ebp-218h],1B0520h
29890ed9 ff158c138229 call   dword ptr [clusres!_imp__DeviceIoControl (2982138c)]
29890edf 85c0      test   eax,eax
29890ee1 740e      je    clusres!IsClusterSupported+0x1C (29828e11)

clusres!IsClusterSupported+0xee:
29890ee3 68408c8229 push   offset clusres!`string' (29828c40)
29890ee8 893e      mov    dword ptr [esi],edi
```

0 : 000> da 29828c94  
29828c94 "CLUSDISK"

# 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
<a href="#">List All Disks</a>		Success
<a href="#">List Potential Cluster Disks</a>		Success
<a href="#">Validate Disk Access Latency</a>		Success
<a href="#">Validate Disk Arbitration</a>		Cancelled
<a href="#">Validate Disk Failover</a>		Cancelled
<a href="#">Validate File System</a>		Cancelled
<a href="#">Validate Microsoft MPIO-based disks</a>		Success
<a href="#">Validate Multiple Arbitration</a>		Cancelled
<a href="#">Validate SCSI device Vital Product Data (VPD)</a>		Warning
<a href="#">Validate SCSI-3 Persistent Reservation</a>		Failed
<a href="#">Validate Simultaneous Failover</a>		Cancelled

## 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

The image shows two windows of the Failover Cluster Manager. The left window displays the 'Validate a Configuration Wizard' with a progress bar at 100% for 'Validating'. The right window shows the 'Storage' validation results table.

**Validate a Configuration Wizard (Left Window):**

- Before You Begin:** Select Servers or a Cluster, Testing Options, Test Selection, Confirmation, **Validating**.
- Progress:** 100% (Validating)
- Test:** Validate Disk Access Latency, Validate Disk Arbitration, Validate Disk Failover, Validate File System, Validate Microsoft MPIO-based disks, Validate Multiple Arbitration, Validate SCSI device Vital Product Data (VPD), **50% Validate SCSI-3 Persistent Reservation** (Issuing Persistent Res).
- Result:** The test passed, Pending, Pending, Pending, The test passed, Pending, The test passed, Pending.

**Storage Validation Results (Right Window):**

Name	Result	Description
List Disks	Success	
List Disks To Be Validated	Success	
Validate CSV Network Bindings	Success	
Validate CSV Settings	Success	
Validate Disk Access Latency	Success	
Validate Disk Arbitration	Success	
Validate Disk Failover	Success	
Validate File System	Success	
Validate Microsoft MPIO-based disks	Success	
Validate Multiple Arbitration	Success	
Validate SCSI device Vital Product Data (VPD)	Success	
Validate SCSI-3 Persistent Reservation	Success	
Validate Simultaneous Failover	Success	
Validate Storage Spaces Persistent Reservation	Success	



# THANK YOU



[plus.google.com/+RedHat](https://plus.google.com/+RedHat)



[linkedin.com/company/red-hat](https://linkedin.com/company/red-hat)



[youtube.com/user/RedHatVideos](https://youtube.com/user/RedHatVideos)



[facebook.com/redhatinc](https://facebook.com/redhatinc)



[twitter.com/RedHatNews](https://twitter.com/RedHatNews)