Improving availability of VMware vSphere 4's virtualization environment with ExpressCluster X

July 31, 2009

1. To begin with

In a server integrated environment, due to adoption of virtual environments, system failure of physical servers will lead to the stop of all the services running on the virtual machines. Therefore, the improvement of availability in a virtual environment will become a key element to maintain business continuity.

Although VMware vSphere™4 has its own availability functions, such as VMware HA and VMware FT, the monitoring target of these functions are limited, so for example, they can't detect failure of business applications running on the virtual machines. So in this document, we will verify some ways to realize further high availability, by installing clustering software, ExpressCluster X, and availability improving software for single servers, ExpressCluster X SingleServerSafe(SSS), in VMware vSphere 4 environments.

In this document, we will use VMware HA as a basic configuration, and install ExpressCluster in each layer (VMware ESX host, and guest OS). Then, we will compare and verify the difference between VMware HA basic configuration and ExpressCluster installed configuration, in perspective of availability, and describe the specific advantages that can be received by installing ExpressCluster.

The verifying configuration (A~E) are the following.

(A) VMware basic configuration
(B) Installing ExpressCluster XSSS in the ESX host
(C) Installing ExpressCluster XSSS in the guest OS
(D) Installing ExpressCluster X in the guest OS
(E) Installing ExpressCluster XSSS / X in the ESX host /guest OS (recommended configuration))

2. Comparative verification with VMware HA

Here, we will compare and verify the business continuity when failure occurs in physical server levels or virtual machine levels, in each configurations (VMware HA / ExpressCluster installed). The errors assumed are noted below.

- Physical server level
  - network error
  - disk error
- Virtual machine level
  - guest OS load stall
  - guest OS stop[1]
  - business application abnormalities

2.1. (A) VMware HA basic configuration

The following two are VMware HA's functions. In this document, we will verify in an environment with all these functions activated.

ESX host monitoring	Will monitor whether the ESX host is dead or alive, by network heartbeat between two ESX hosts or between ESX host and a vCenter server, using a service console LAN. And when the ESX host goes down, will restart the virtual machines running on that host on another healthy ESX host (which is called "virtual machine failover").
Virtual machine monitoring	Will run heartbeat between ESX host and guest OS (VMware Tools), and restart the virtual machine on the same host when the guest OS stops.

Figure 1 shows the detail overview of VMware HA basic configuration. Details of network configuration and disk configuration will be stated in figure 2 and figure 3. The detail figure shows only one ESX host, but the second host will be the same configuration.

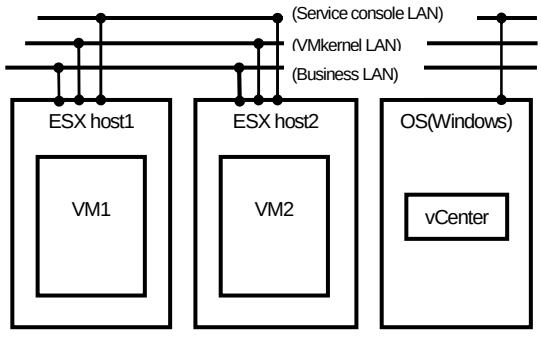

Figure 1 : VMware HA basic configuration

[1] Here, we will imagine bluescreen for Windows, and kernel panic for Linux

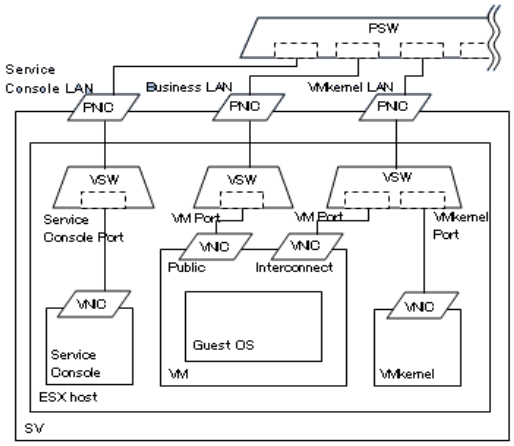

Figure 1 : Network Configuration

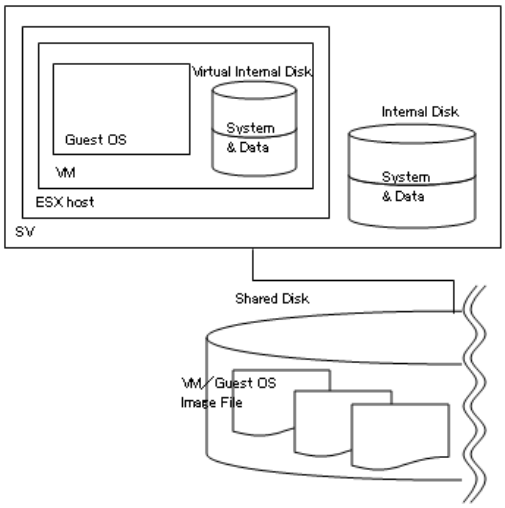

Figure 2 : Disk Configuration

The abbreviations in the figures are the following.

SV	Physical Server
VM	Virtual Machine
PNIC	Physical NIC
VNIC	Virtual NIC
PSW	Physical Switch
VSW	Virtual Switch

As for the network configuration, we assigned the service console port, the VMkernel port, and the virtual machine port to different physical NICs, considering the availability and performance in virtual environments. In the following, we will describe the LANs, which are composed of physical NICs connected to each port, the service console LAN, the VMkernel LAN, and the business LAN. Also, please note that we added a virtual NIC (noted as "Interconnect") of a virtual machine port, under the VMkernel LAN, due to the configuration of installing ExpressCluster X in guest operating systems. We will use this as an Interconnect LAN when configuring a cluster between guest operating systems. This will not be used if not configuring a cluster between guest operating systems.

As for the disk configuration, since we are using VMware HA, the virtual machine (and the virtual disk that the virtual machine uses) and guest operating system files will be stored in the shared storage.

In this environment, the combination use with VMotion and VMware DRS is possible.[2]

The verification results of the system behaviors, occurring in case of failure in VMware HA configuration, are organized in the chart below. This chart will be the base of this comparative verification hereinafter.

Error section	Location of failure	Result of verification
Physical server	(a) Service Console LAN	Can be detected by VMware HA's ESX host monitoring →virtual machine failover
	(b) VMkernel LAN	Undetectable
	(c) business LAN	Undetectable
	(d))shared disk[3]	Undetectable →guest operating system hung-up mode
	(e) internal disk[4]	Undetectable
Virtual Machine	(f) OS load stall	Undetectable
	(g) OS stop	Can be detected by VMware HA's virtual machine monitoring →virtual machine restart
	(h) business application	Undetectable

The following notes the system action (especially in perspective of business application) after failure occurs.

- Enables business continuation
  As for (a) and (g), it is possible to continue business, after restarting the virtual machine by VMware HA's monitoring function.

- Enables business continuation with limitation
  As for (b), there is no effect on business applications, but the virtual machines VMotion will not be able to work

[2] It is required that VMotion and VMware DRS are included in the vSphere package that you are using

[3] Please assume that a single error is occurring on the ESX host. Double error (which errors occur on multiple places) is out of target this time.

[4] Please assume the disk which the ESX host is installed in to.

- Disable to continue business

  As for (c), you cannot continue business, since you can't access to the guest OS.

  As for (d) and (e), you cannot continue business since the guest operating system is in hung- up mode, due to ESX host's defective performance or virtual machine / guest operating system's access failure to the image files. In addition, there are cases that the operating system works partially, such as replying to ping requests to the guest operating system, so, from the perspective of protecting business data, it's hard to say that the machine is in a healthy situation.

  As for (f), the performance of business applications will decline, until the guest operating system's high loaded situation is resolved.

  As for (h), you cannot continue business, since it is not included in VMware HA's monitoring target.

2.2. (B) Installing ExpressCluster XSSS in the ESX host

Here, we will install ExpressCluster XSSS in the ESX host (install in the service console), shown in VMware basic configuration (figure1), and enhance the monitoring functions at physical server level.

The configuration outline is described in Figure 4.

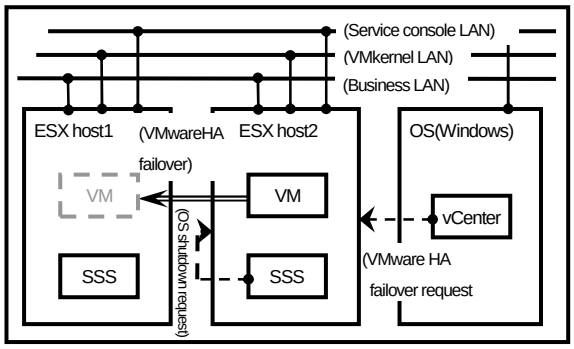

Figure 3 : VMware HA + host SSS configuration

The results of verification are organized in the chart below. The parts that differ from VMware HA basic configuration are underlined. However, the results of errors occurring at virtual machine level is the same as VMware HA basic configuration, since SSS on the ESX host cannot monitor errors occurring at virtual machine level, so we will omit that part.

Error section	Location of failure	Result of verification
Physical Server	(a) Service Console LAN	Can be detected by VMware HA's ESX host monitoring →virtual machine failover
	(b) VMkernel LAN	Can be detected by NIC Link Up/Down monitoring →Actions such as restarting the ESX host are possible in abnormal status
	(c) Business LAN	Same as above
	(d) Shared disk	Can be detected by disk monitoring →Prevents the guest OS to continue the application in an unstable status, by shutting down the ESX host when in abnormal status After shutting down the server, VMware HA will failover the virtual machine
	(e) Internal disk	Same as above

The advantages are listed below.

- NIC errors, other than the service console LAN

  Able to improve availability by restarting the ESX host or failing over the virtual machine, when SSS's NIC Link Up/Down monitor detects an error. The service console LAN will be monitored by VMware HA.

- Disk error

  Able to avoid the guest OS to continue running the application in a half-stalled status, by shutting down the ESX host when SSS's Disk monitor detects an error. Then, VMware HA's ESX host monitor will detect the server down, and failover the virtual machine.

2.3. (C) Installing ExpressCluster XSSS in the guest OS

Here, we will install ExpressCluster XSSS in the guest OS, and enhance the monitoring functions at virtual machine level. Also, this will enable to monitor the physical server indirectly from the virtual machine monitor.

The configuration outline is described in Figure 5.

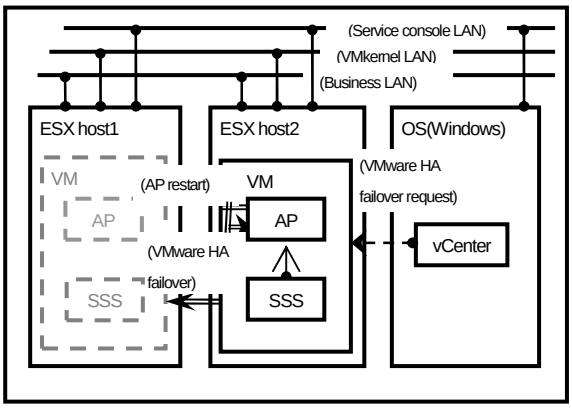

Figure 4 : VMware HA + Guest SSS Configuration

The verification results of this configuration are organized in the chart below. The parts that differ from VMware HA basic configuration are underlined.

Error section	Location of failure	Result of verification
Physical Server	(a) Service Console LAN	Can be detected by VMware HA's ESX host monitoring →virtual machine failover
	(b) VMkernel LAN	Undetectable
	(c) Business LAN	Can be detected indirectly by SSS's IP monitor from the guest OS. →There are no effective actions to take, when in abnormal status
	(d) Shared disk	(Linux version only) Can be detected indirectly by user space monitoring →Able to reset the virtual machine and stop the application when an error occurs.
	(e) Internal disk	Undetectable
Virtual machine	(f) OS load stall	Can be detected by user space monitoring (Linux version) / Disk RW monitoring (Windows version) →Resets the virtual machine and restarts the guest OS when an error occurs
	(g) OS stop	Can be detected by VMware HA's virtual machine monitoring →Restarts the virtual machine
	(h) Business application	Can be detected by monitor Agents[5] or monitoring whether the process is alive or not →Able to restart the application when an error occurs

The difference between VMware HA, is that in this configuration, you can improve the availability of virtual machines by restarting the guest OS or business applications when an error occurs on the virtual machine. The advantages are listed below.

- Shared disk error (Linux version only)

  Able to avoid the guest OS to continue running the application in a half-stalled status, by resetting the virtual machine when SSS's user space monitor detects an error.

  If the guest OS is Windows version, the business application might continue in an unstable status, likewise the VMware HA configuration[6].

- Errors occurring at virtual machine level

  As for (f) and (g), the OS stall/stop monitoring function will be complemented by combining SSS and VMware HA.

  As for (h), it is possible to monitor business applications that VMware HA doesn't support.

Also, (c) can be detected indirectly by SSS's IP monitor[7], since the business LAN is assigned to a virtual NIC of the virtual machine. But no effective action can be made toward the physical server, since SSS is running on the guest OS on a virtual machine. However, it is possible to stop the application if you don't want to continue the application with the network unusable.

2.4. (D) Installing ExpressCluster X in the guest OS

Here, we will install ExpressCluster X in the guest OS, and enhance the monitoring functions at virtual machine level, and configure a cluster between guest operating systems. You can reduce business downtime by the cluster configuration, and also, take in the advantages of configuration C (2.3). Please refer to 6.1 for more details of business downtime.

The configuration outline is described in Figure 6. The cluster type used for verification is mirror disk type[8].

[5] Able to detect application's hung-up status and result error, by a specialized monitoring for specific applications

[6] This is because RW monitor (Windows version) cannot detect the guest OS stall that happens when a shared disk error occurs.

[7] Here, we will assign a system external address as the monitoring target. (Ex; the business LAN's default gateway IP address used on the guest OS)

[8] You cannot use VMotion and VMware DRS when using the shared disk type cluster. This is because VMotion cannot be used when choosing something else than "No" for the virtual machine's SCSI controller's "SCSI bus share", (when you share the disk between the virtual machines), due to VMware's specifications.

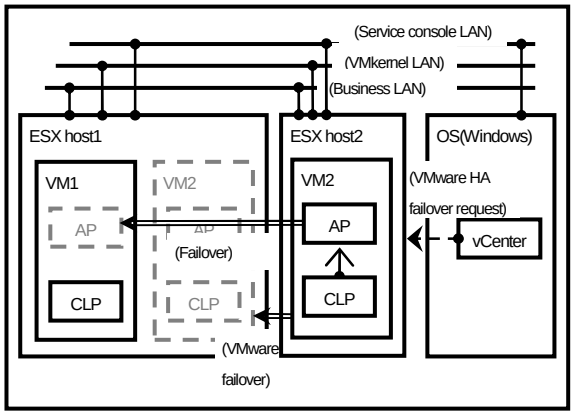

Figure 5 : VMware HA + Guest OS Cluster Configuration

The verification results of this configuration are organized in the chart below. The parts that differ from VMware HA basic configuration are underlined.

Error section	Location of failure	Result of verification
Physical Server	(a) Service Console LAN	Can be detected by VMware HA's ESX host monitoring →virtual machine failover The business application can be failed over during the virtual machine's failover
	(b) VMkernel LAN	Undetectable
	(c) Business LAN	Can be detected indirectly by Express Cluster's IP monitor from the guest OS →Able to failover business application when an error occurs
	(d) Shared disk	(Linux version only) Can be detected indirectly by user space monitoring →Resets the virtual machine when an error occurs. Business application can be failed over while resetting the virtual machine
	(e) Internal disk	Undetectable
Virtual machine	(f) OS load stall	Can be detected by user space monitoring (Linux version) / Disk RW monitoring (Windows version) →Resets the virtual machine and restarts the guest OS when an error occurs. Business application can be failed over while restarting the guest OS
	(g) OS stop	Can be detected by VMware HA's virtual machine monitoring →Restarts the virtual machine. Business application can be failed over while restarting the virtual machine
	(h) Business application	Can be detected by monitor Agents or monitoring whether the process is alive or not →Able to failover business application when an error occurs

The difference between VMware HA, is that in this configuration, you can reduce business downtime by failing over business applications when an error occurs on the virtual machine. The advantages are listed below.

- Service console LAN's NIC error

  VMware HA's ESX host monitoring will failover the virtual machine. Meanwhile, the guest OS cluster will detect heartbeat timeout, and failover the business application to the standby guest OS, without waiting for the virtual machine's failover to complete. When the virtual machine's failover is completed, a guest OS cluster will be rebuilt on the new server (which the virtual machine failed over to).

- Business LAN's NIC error

  This can be detected indirectly by ExpressCluster X's IP monitor, likewise configuration C (2.3). In addition, it is possible to failover business application to the standby guest OS when the error occurs.

- Shared disk error (Linux version only)

  Able to avoid the guest OS to continue running the application in a half-stalled status, by resetting the virtual machine when ExpressCluster X's user space monitor detects an error. After resetting the virtual machine, the guest cluster will detect a heartbeat timeout, and failover business application. However, there are a few reminders about ExpressCluster's configuration. (Noted at the end of this section)

  If the guest OS is Windows, the half-stalled status of the guest OS will not be resolved. So in some cases, business application will activate on both servers (active/standby), and eventually, the business will stop[9].

[9] It is ExpressCluster's feature to urgently shutdown both servers when the business application activates on both servers, in perspective of protecting the data by exclusive control of resources.

Therefore, to address this error when using windows as a guest OS, it is required to install SSS in the ESX host (Configuration E (2.5)), and force the ESX host to shutdown when an error occurs.

- Errors occurring at virtual machine level

  As for (f) and (g), configuration C (2.3) only restarts the guest OS, but in this configuration, it is possible to failover the application to the standby guest OS while restarting the guest OS, by detecting heartbeat timeout between the guest cluster.

  As for (h), business application will be failed over immediately after the error is detected.

Reminders of ExpressCluster when shared disk errors occur

Failover failure will occur, if the business application's failover process is performed before the reset of the virtual machine is completed from the user space monitor. However, this is limited to cases where the floating IP (FIP) resource is contained in the failover group that the business application belongs to.

This happens because the guest OS (where the error is occurring), responds to the ping in a half-stall status when shared disk error occurs. ExpressCluster is designed to fail failover if the FIP address is already in use (if there is ping response) when starting the failover process, in order to avoid the FIP resource's double activation.

For the reasons above, it is required to make the timing to reset the virtual machine shorter than the failover time, by following the methods below.

- Change the FIP resource's ping retry times and ping interval value larger
- Change the user space monitor's monitoring timeout value smaller

2.5. (E) Installing ExpressCluster XSSS / X in the ESX host /guest OS

Here, we will install ExpressCluster XSSS in the ESX host, plus ExpressCluster X in the guest OS, and enhance the monitoring functions at both physical and virtual machine levels. This configuration provides the advantages of both configuration B (2.2) and D (2.4)

The configuration outline is described in Figure 7. Mirror disk type is used for guest OS clustering in the verification environment.

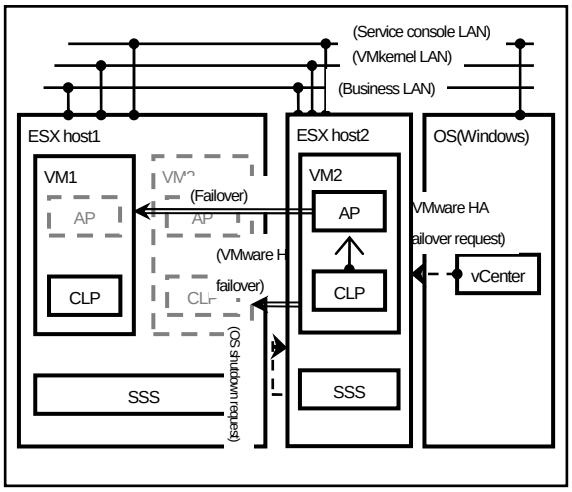

Figure 6 : VMware HA + host SSS／guest OS cluster's Combination Configuration

The verification results of this configuration are organized in the chart below. The parts that differ from VMware HA basic configuration are underlined.

Error section	Location of failure	Result of verification
Physical Server	(a) Service Console LAN	Can be detected by VMware HA's ESX host monitoring →virtual machine failover. The business application can be failed over during the virtual machine's failover
	(b) VMkernel LAN	Can be detected by SSS's NIC Link Up/Down monitoring from the ESX host. →Actions such as restarting the ESX host are possible when in abnormal status. Business application can be failed over while restarting the ESX host
	(c) Business LAN	Same as above
	(d) Shared disk	Can be detected by SSS's disk monitoring from the ESX host →Avoids the guest OS to continue running business in an unstable status, by shutting down the ESX host when in abnormal status. Business application can be failed over during the

		ESX host's shutdown
	(e) Internal disk	Same as above
Virtual machine	(f) OS load stall	Can be detected by ExpressCluster's user space monitoring(Linux) /disk RW monitoring (Windows) from the guest OS →Resets the virtual machine and restarts the guest OS when in abnormal status. Business application can be failed over while resetting the virtual machine.
	(g) OS stop	Can be detected by VMware HA's virtual machine monitoring →Restarts the virtual machine. Business application can be failed over while restarting the virtual machine
	(h) Business application	Can be detected by monitor Agents or monitoring whether the process is alive or not, from ExpressCluster on the guest OS →Business application can be failed over when in abnormal status

The difference between VMware HA, is that in this configuration, it is possible to monitor from both physical and virtual machine level, and you can reduce business downtime by failing over business applications when an error occurs. The advantages are listed below.

- Disk error
  It is possible to continue business by failing over business application, regardless of the guest OS's type.

- Errors occurring at physical/virtual machine level
  The availabilities of configuration B (2.2) and D (2.4) will be applied

2.6. Summary of verification results

As for configuration A to E, the verification results of availability are summarized in chart 1. Availability will improve gradually from A to E.

The recommended configuration is E, since the availability of the business application running on the guest OS is high, and it can support errors occurring at physical server level, such as physical server's shared disk error. However, the installation cost will increase (figure 8), so it is necessary for customers to choose the configuration that matches their use best.

Chart 1: Availabilities of each configuration

Error section	Location of failure	A	B	C	D	E
Physical Server	Service Console LAN	2	2	2	3	3
	VMkernel LAN	1	2	1	1	3
	Business LAN	0	2	0	3	3
	Internal disk	0	2	0	0	3
	Shared disk	0	2	0	*	3
Virtual machine	OS load stall	0	0	2	3	3
	OS stop	2	2	2	3	3
	Business application	0	0	3	3	3

The marks above rated availability from the perspectives below.

3	Able to detect errors and continue business (The OS startup time is not included in downtime)
2	Able to detect errors and continue business (The OS startup time is included in downtime)
1	Cannot detect errors, but able to continue business with some limitations
0	Cannot continue business
*	Differs by the guest OS Linux→□、Windows→×

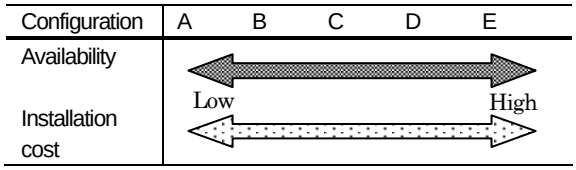

Figure 7: The relationship between availability and installation cost

3. Comparative verification with VMotion

VMotion can lively migrate virtual machines between the ESX hosts, and has high convenience in terms of operation, but it doesn't have any availability functions so you cannot prepare for disasters with VMotion only.

Therefore, in an availability emphasized system, the ExpressCluster installed configuration (verified in section 2) will be required.

Also, the point that you can carry out planned maintenance of the physical server (ESX host) can be mentioned as VMotion's advantage. You can update the ESX host and perform maintenance on the physical server's hardware while continuing business, by using VMotion to move the virtual machine running on the target ESX host to a different ESX

host. However, VMotion cannot move the business application running on virtual machines (guest OS) by application units, so business will have to stop when performing maintenance on virtual machines.

As for configurations D (2.4) and E (2.5), which install ExpressCluster in the guest OS, it is possible to continue business while performing the virtual machine's maintenance, since these configurations can move the applications by application units. In this case, the business downtime is the time that takes to move the application only.

As for configurations D and E, you can also use VMotion by configuring a mirror disk type cluster between the guest operating systems.

Chart 2: Business suspension when performing maintenance

Configuration	Physical server (ESX host)	Virtual machine (Guest OS)
VMotion	Suspend	Continue
Configuration D&E	Suspend	Suspend

4. Consideration of VMware ESX host cluster

It is possible to failover the virtual machine from ExpressCluster X, by installing ExpressCluster X in the ESX host and clustering the ESX hosts. This is realized by controlling the virtual machine's startup, stop, and monitoring from ExpressCluster X.

The ESX host cluster's configuration diagram is described in figure 9.

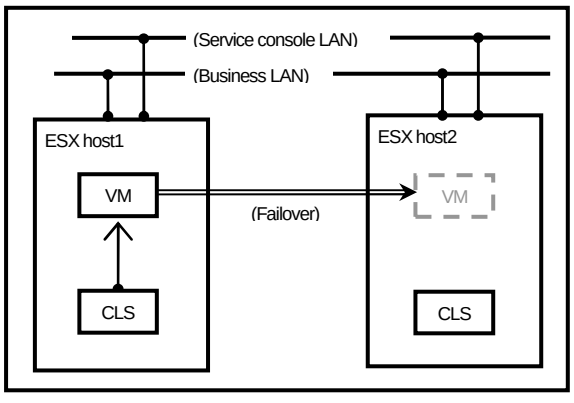

Figure 8 : ESX host Cluster

The advantages and disadvantages of this configuration are listed below.

- Advantages
  - You can configure the virtual machine's failover policy by ExpressCluster X. The failover policy enables to prioritize the ESX host's order to failover the virtual machine.
  - You can carry out the virtual machine's failover when the monitor resources detect abnormality, and configure the failover trigger circumstantially. (VMware HA can only detect heartbeat disconnection using the service console LAN)
- Disadvantages
  - It can't be used in combination with VMware HA, VMotion, and VMware DRS, because the virtual machine's start and stop is controlled from ExpressCluster X.

5. At last

In this document, we installed ExpressCluster X / XSSS in various configurations in VMware vSphere 4's virtualization environment, and compared and verified each configuration in perspective of availability with the VMware HA basic configuration. As a result, it has been proven from each configuration that availability can be improved by installing ExpressCluster in the virtualization environment and complementing VMware's functions. Specifically, they can be widely divided into the three points below.

- Improving availability of the business applications running on the virtual machines
- Improving availability toward errors occurring at physical server / virtual machine level.
- Enables to perform maintenance with the business continuing (Only in guest OS cluster configuration)

6. Appendix

6.1. Business downtime comparison

We compared the business downtime occurred by service console LAN's NIC error and OS stop error in configuration A, D, and E. As a reference, they are described in figure 10 and 11. The downtime of configuration A is longer compared to configuration D and E, since it includes the guest OS's stop/startup time in the virtual machine's failover and restart process.

- Service console LAN's NIC error

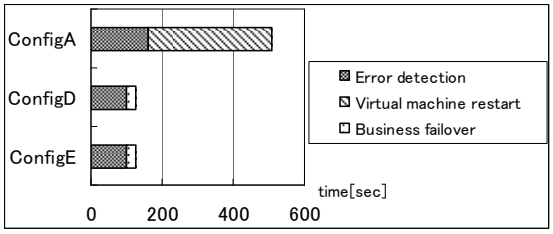

Figure 9: Business downtime comparison (Service console LAN)

- OS stop error

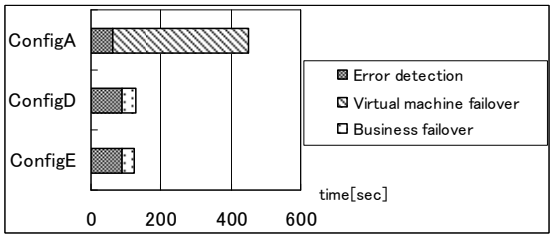

Figure 10: Business downtime comparison (OS stop)

Reminders of figure 10 and 11 are noted below

- The time required for the virtual machine's failover and restart depends on the guest OS's type, since it includes the guest OS's stop and startup. Red Hat Enterprise Linux 5.3 is used for the guest OS in the business downtime described above.
- As for VMware HA's virtual machine monitor and ExpressCluster's each monitors, it is possible to change the values of monitor interval and heartbeat timeout. Default values are used for the assessments above.

NEC Corporation
2nd Computers Software Division
ExpressCluster X Group

IT Platform Global Business Development Division
E-mail : info@expresscluster.jp.nec.com