High-reliability, High-availability Cluster System Supporting Cloud Environment

Seishiro Hamanaka  Kunikazu Takahashi

Fujitsu is offering FUJITSU Software PRIMECLUSTER, a high-reliability software platform that supports the continuous operation of entire systems. With its highly reliable technology linked to Fujitsu servers (SPARC M10, FUJITSU Server PRIMEQUEST 2000/1000 series, and FUJITSU Server PRIMERGY), PRIMECLUSTER ensures fast detection of malfunctions and secure operational continuity not only in a physical environment but also in a cloud environment, unlike other companies’ cluster software. Also, due to a link to the live migration function of the server provided by virtualization software, PRIMECLUSTER enables the server unit to undergo maintenance without changing redundant configurations, or active-standby configurations. This realizes even greater availability for platform systems in a cloud environment. In these ways, PRIMECLUSTER supports continuous operation on a 24/7 basis, even in a cloud environment that integrates a customer’s business systems. This paper introduces the technology of PRIMECLUSTER, cluster software that supports platform systems in a cloud environment and business systems running in a cloud environment.

1. Introduction

Through the effective use of system resources such as servers, storage, and networks and advances in virtualization technology reducing operating costs, our customers’ business systems are migrating to a cloud environment. The prolonged suspension of business systems not only causes significant losses and missed business opportunities for the company involved, but it can also develop into a social problem. This is why there is a need for technology that quickly makes operation continue if a server malfunctions or an operating system (OS) crashes, and for technology that enables maintenance to be carried out on a server unit without suspending operation.

FUJITSU Software PRIMECLUSTER delivers the same high-reliability, high-availability technology achieved in a physical server environment to a business system operating in a cloud environment (referred to below as a virtual machine). It also brings reliability and availability to the cloud environment by providing new technology that does not suspend operation.

In this paper, we introduce the technology of PRIMECLUSTER, cluster software that achieves high reliability and availability in order to support platform systems in a cloud environment and virtual machines.

2. Issues facing operational continuity in a cloud environment

Shutting down business systems is unacceptable to our customers who require the provision of a constant service on a 24/7 basis. Recently, system shutdown has become equally unacceptable even in a cloud environment in which multiple operations are running with the use of virtualization software. In these circumstances, even in the event of an operation shutting down due to kernel panic\(^1\) of a virtual machine, the length of time from shutdown to resumption must be kept to a minimum.

Also, with a platform system in a cloud environment in which multiple operations are running, it is difficult to suspend all operations together to get enough time for maintenance. The operations

\(^1\) A state in which the OS completely stops due to abnormal operation on a server such as a CPU or memory failure or due to a fatal error with the OS (kernel part).
therefore cannot be shut down when conducting server unit and storage maintenance or applying OS modifications. New technology is required that will make it possible to carry out server unit maintenance without impacting the operations running on a virtual machine.

3. New high-availability technology supporting a cloud environment

To resolve the issues facing operational continuity in a cloud environment, PRIMECLUSTER achieves reliability and availability in virtual machines through new technologies aimed at a cloud environment in addition to the technologies of “server operational continuity,” “disk access continuity” and “network continuity” already developed in a physical server environment. (See “Four features for dealing with a cloud environment” below.)

In this section, we introduce two new technologies for resolving these issues.

3.1 Technology ensuring operational continuity even in case of malfunction

By linking to the system monitoring facility of Fujitsu servers (SPARC M10, FUJITSU Server PRIMEQUEST 2000/1000 series, and FUJITSU Server PRIMERGY), PRIMECLUSTER quickly detects malfunctions and securely switches servers to ensure operational continuity. We apply this technology to a cloud environment too, and here we explain the technologies that achieve operational continuity even when a virtual machine malfunctions.

1) Technology to immediately detect a system malfunction in a virtual machine

When a system malfunction such as a kernel panic occurs in a virtual machine, it is detected by the system monitoring facility of the server unit in which the malfunction occurred. The system monitoring facility immediately notifies PRIMECLUSTER, which is running on a normal virtual machine. Having received the notification, PRIMECLUSTER switches operation to the normal virtual machine.

Through this technology, PRIMECLUSTER ensures continued operation even if a malfunction occurs in a virtual machine.

Figure 1 shows the process of immediate notification of a malfunction in a virtual machine, taking as an example an environment using Oracle VM Server for SPARC.

This technology is also implemented in an environment in which the Red Hat Enterprise Linux virtual machine function (Kernel-based Virtual Machine, or KVM) is used on a PRIMEQUEST 2000/1000 series
2) Technology to securely stop a malfunctioned virtual machine

If a virtual machine in which a system malfunction such as an OS hang-up\(^2\) has occurred continues to run without being shut down, the virtual machine could unjustly access the disk drive being used in operation.

To prevent such unauthorized access, it needs to be ensured that a virtual machine in which a system malfunction has occurred is shut down.

It needs to be ensured that the virtual machine shuts down using technology that can check whether virtual machines are alive from their heartbeat via a cluster interconnect and that can forcibly shut down a virtual machine linked to the host OS (the OS running on the server unit).

**Figure 2** shows the process of securely shutting down a virtual machine that has malfunctioned, taking as an example the environment in which KVM is used.

In addition to the KVM environment, this technology is also implemented in the following two environments.

- Oracle VM Server for SPARC
- VMware vSphere 5 / VMware vSphere 4

### 3.2 Technology realizing better availability during system maintenance

In the platform system of a cloud environment, active-standby configurations need to be kept even during server unit maintenance, just in case there is a malfunction.

With PRIMECLUSTER, due to a link to the live migration function of Oracle VM Server for SPARC, a virtual machine can be moved to a different server unit while maintaining operational continuity. Virtual machine migration is easily achieved by a Graphical User Interface (GUI) and linking to FUJITSU Software ServerView Resource Orchestrator (ROR) which manages the entire system including the server, storage, and so on. **Figure 3** shows the technology that achieves better availability during system maintenance.

This technology allows maintenance to go ahead while keeping active-standby configurations, and enhances availability in the platform system of a cloud environment.
4. Four features for dealing with a cloud environment

PRIMECLUSTER is a highly reliable software platform which, due to the redundant configurations of a server, storage and network, improves reliability and availability and minimizes the shutdown time of a business system.

In this section, we describe the four features for dealing with a cloud environment.

4.1 Server operational continuity

A cluster system uses multiple servers to increase availability. It is composed of an operational system on which business is conducted and a standby system for switching over to and using in case of malfunction. If there is no heartbeat response between servers, the system switches to the standby system and operation continues.

PRIMECLUSTER, as mentioned above, achieves reliability and availability by immediately detecting a malfunction on a virtual machine and securely shutting down a malfunctioning virtual machine. Here, we describe hot standby and standby patrol for further increasing availability.

1) Hot standby

We support a method called hot standby, which, after a malfunction occurred in the operational system, is ready in advance to resume operation on a standby system, unlike ordinary standby methods of transferring data and restarting work applications on a standby system. Compatible software (such as FUJITSU Software Symfoware Server) is standing by ready to immediately resume business processes on the standby system, by starting up software on the standby system and having a shared disk device open beforehand. This can shorten the time required to start up software when the server is switched to the standby system due to a server malfunction etc., enabling the rapid resumption of operation.

2) Patrol diagnosis

PRIMECLUSTER provides a standby patrol function, which prevents failure of operational continuity by monitoring the server, storage and network not only on the operational system but also on the standby system. When a malfunction occurs in the standby system, the resource status is indicated on the PRIMECLUSTER console screen and a message is output to the system log to prompt the system administrator to take action at the site of the malfunction.

4.2 Disk access continuity

Here we describe the features of PRIMECLUSTER GDS, which enables operation as one disk device, making multiple disk devices redundant.
1) システムボリュームミラーリング

システムボリュームディスクデバイスが故障した場合、ディスクデバイスの交換と復旧の作業が必要です。この場合、システムボリュームミラーリングを行うと、故障しているディスクデバイスを隔離し、通常のディスクデバイスで操作を行なうことが可能になります。また、ホットスペア機能も自動的に失効しているディスクデバイスを交換し、冗長なディスクデバイスの構成を自動的に復旧する機能があります。

2) レベル2 RAIDのミラーリング

レッジアリットーイディスク (RAID) デバイスは、ストレージの可用性を向上させるために広く導入されています。しかし、24時間7日間のミッションクリティカルシステムでは、さらに大きなデータ可用性が必要です。RAIDデバイスを冗長化し、ミラーリングを実行することにより、より信頼性の高いストレージシステムを実現することができます。

3) I/O応答時間保証

ハードウェアデバイスの故障（ディスクデバイス、ファイバチャネルスイッチ等）が発生すると、ディスクドライブ等の再試行が行われ、I/O応答は遅延します。I/O応答が遅延すると、システム全体の応答速度が低下し、運用の継続が影響されます。

PRIMECLUSTER GDSを用いてディスクドライブミラーリングを行い、ディスクドライブのI/O応答を監視します。応答が一定の時間帯を越えなかった場合は、ミラーリングを中断し、通常のディスクデバイスのみで運用を継続し、I/O応答時間を保証します。（図4）

4.3 ネットワーク継続

サーバが冗長化されても、通信相手が失われた場合や運用継続が影響される場合を含めて運用の継続を保証します。PRIMECLUSTER GLSは、ネットワーク経路の異常を検出し、異常経路を即座に切り替え、運用を継続します。また、ネットワーク経路の状態もモニターし、障害の発見を促進します。

Figure 4
I/O response time guarantee.
4.4 Virtual machine operational continuity

Technology developed in a physical server environment such as server operational continuity, disk access continuity, and network continuity has also been applied to virtual machines and enhanced reliability in entire systems.

PRIMECLUSTER is compatible with the following virtualization software.

1) Oracle Solaris environment
   • Oracle VM Server for SPARC
   • Oracle Solaris zone
   • Oracle Solaris Legacy Containers

2) Linux environment
   • Red Hat Enterprise Linux 7 virtual machine function (KVM)
   • Red Hat Enterprise Linux 6 virtual machine function (KVM)
   • Red Hat Enterprise Linux 5 virtual machine function (Xen)
   • VMware vSphere 5
   • VMware vSphere 4

5. Conclusion

The trend toward cloud services will accelerate in future in order for enterprises to utilize system resources such as servers, storage, and networks more effectively and reduce operational costs.

At the same time, mission-critical work handling valuable data requires the construction of private cloud services and the operation of platform systems, not migrating to an external cloud service. There is therefore a need for reliability and availability in private cloud services, too.

With PRIMECLUSTER, we will continue to support the stable operation of our customers’ business by linking to virtualization software and delivering technology that further enhances the reliability and availability of their business systems, in a cloud environment in which their mission-critical tasks and other important business systems are integrated, and in a virtualized environment.

References


Seishiro Hamanaka
Fujitsu Ltd.
Mr. Hamanaka is currently engaged in development and promotion of PRIMECLUSTER.

Kunikazu Takahashi
Fujitsu Ltd.
Mr. Takahashi is currently engaged in promotion of PRIMECLUSTER.