Latest Cluster System Technology

EGETA Shusaku, KAITE Ikuo, JINNO Eri

Abstract

In order to prepare for the advent of the ubiquitous society, computer systems are required to offer a higher reliability than ever before. One of the solutions to this requirement is the CLUSTERPRO, software for improving server system availability by introducing clustering technology. This paper is intended to describe the function and performance of the latest version, CLUSTERPRO X, together with details of NEC's efforts aimed at assimilating the latest technologies and an appraisal of future perspectives.

Keywords

clustering, failsafe, forced shutdown, failover, high availability, high reliability, HA

1. Introduction

As the real ubiquitous society approaches, the importance of IT infrastructures is being recognized more and more, so that a higher reliability or a higher "safety" may be offered than ever before. At NEC, we offer clustering software "CLUSTERPRO" as "a first rate technology for implementing 'dependability'" in the framework of our IT platform vision "REAL IT PLATFORM."

Since its first shipment in 1996, CLUSTERPRO has positively assimilated market needs as well as the latest technologies in order to improve system availability, securing thus, a prominent position in the Japanese market.

This paper describes the functions and performances of our latest version, CLUSTERPRO X and outlines our efforts aimed at incorporation of the latest technologies as well as offering an appraisal of future perspectives.

2. Basic Functions and Configurations of CLUSTERPRO

A cluster system is composed of two or more servers. The roles of failover type clustering software including the CLUSTERPRO are roughly categorized into the following two functions. It is these functions that improve the system availability and achieve the failsafe property.

- Detection of irregularities produced in the cluster system.
- Passing-over of job services from an abnormal server to a healthy server for resuming the services (Failover).

CLUSTERPRO is compatible with the Windows and Linux platforms, and is categorized into the shared disk type (**Fig. 1**) and the Data mirroring type (**Fig. 2**) according to the type of the shared data storage.

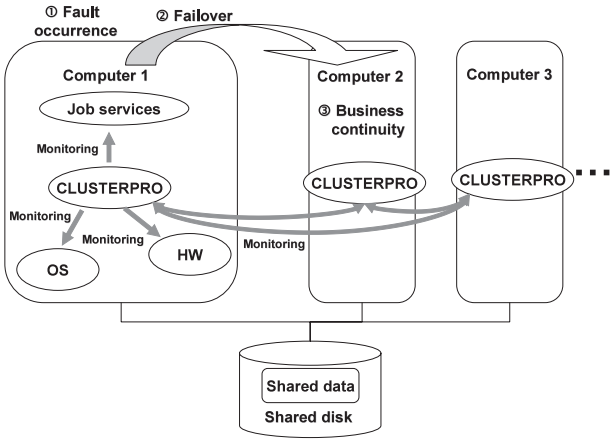

Fig. 1 Shared disk type.

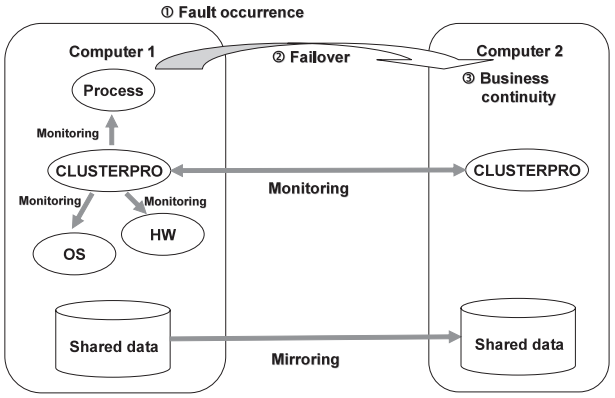

Fig. 2 Data mirroring type.

In October 2006, we released the latest version of CLUS-

TERPRO, which we named CLUSTERPRO X. Below we review the following topics by referring to the new functions and enhanced points of CLUSTERPRO X.

- Complication and diversification of configurations
- Advance toward the DR (Disaster Recovery) domain
- High availability at the mainframe level
- VM-level switching
- Process-level switching
- Image of effective operation in the future

3. Complication and Diversification of Configurations

While the basic configurations of CLUSTERPRO include the shared disk and data mirroring types as mentioned above, CLUSTERPRO X features a configuration that extends the performance of the data mirroring type. The innovative N:1 data mirror type configuration uses a single, centralized server to which all data is to be mirrored from multiple source servers (**Fig. 3**).

In addition, as a further enhancement for dealing with the complications and diversification of configurations, we have also added a mixed configuration[*1] of the shared disk and data mirroring types. This enables the shared data of site A to be mirrored at site B by installing a shared disk type cluster in site A and a standby server in site B.

[*1] Currently under study.

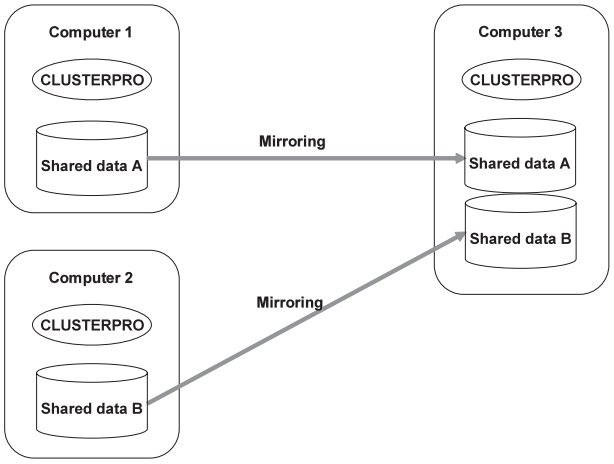

Fig. 3 N:1 data mirror type.

4. Advance Toward the DR Domain

CLUSTERPRO X features the three following enhancements related to data mirror type clusters.

1) Mirroring protocol changes from an original protocol to TCP/IP.
2) Increased mirroring speed.
3) Incorporation of the async I/O function.

The use of TCP/IP protocol has made it possible to place clustered servers in different networks, and the increase in mirroring speed and incorporation of async I/O function have facilitated the placement of servers at remote locations.

The increase in mirroring is found to be as high as about 150% in a full-scale copy processing test of the mirroring source area (**Table**).

Fig. 4 shows the results of an experiment for the verification of the async I/O function, which was conducted by connecting Tokyo and Hiroshima through IP-VPN. The data shows that, when the data write rate in Tokyo (Write KB/sec.) is increased gradually, the communication rate for transferring the write data to Tokyo (Send KB/sec.) is about 1,500KB/sec (12Mbps), which is adequate for the data write rate. However, when the

Table Time required for full-scale copy.

Product	Time Required	Speed Increase Rate
CLUSTERPRO Ver. 8.0	0:16:38	–
CLUSTERPRO X Ver. 1.0 (Sync I/O)	0:09:38	173%
CLUSTERPRO X Ver. 1.0 (Async I/O)	0:11:34	144%

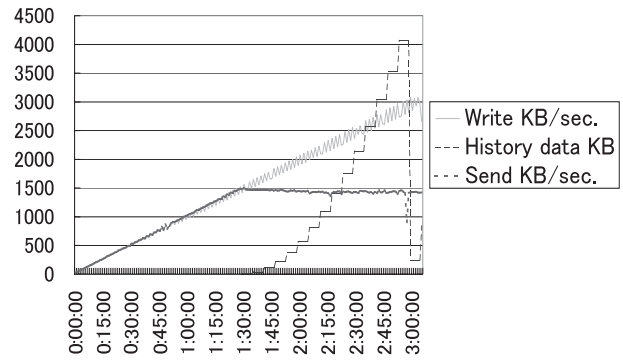

Fig. 4 Async I/O performance (Tokyo–Hiroshima).

data write rate exceeds 1,500KB/sec., the communication rate begins to be delayed and the non-transferred data starts to be stored as history data. This means that the environment used in the verification can withstand updating type jobs of 1,500KB/sec., or nearly 13GB per day.

Our systems have actually been achieving steady results in the DR (Disaster Recovery) domain in North America since our exhibit at CeBIT 2004 demonstrated clustering of ft servers installed in remote locations using CLUSTERPRO and have received the "Best of CeBIT America 2004" award.

5. High Availability at the Mainframe Level

Since the first shipment in 1996, CLUSTERPRO has been enhancing its clustering software functions based on NEC's DNA, cultivated through mainframe systems. Particularly, we are still continuing improvements with regard to the issue of how to promptly shut down a server where an irregularity has occurred. Our strategy here is adopted from the viewpoint of preventing data destruction resulting from a server being in a condition of irregularity.

At present, CLUSTERPRO X offers the following types of forced server shutdown functions.

- Forced shutdown, system panic and reset using a watchdog driver.
- Forced shutdown and reset interlocked with the hardware (H/W) watchdog timer incorporated in the BMC (Baseboard Management Controller).

These functions can be regarded as "methods for letting a server experiencing irregularity to shut itself down."

In addition to the above, we are also planning to provide "methods for letting an outside server shut down a server that experiences irregularity" as follows.

- Forced shutdown, system panic and reset using RDMA.
- Forced shutdown and reset based on remote control of BIOS.

6. VM-Level Switching

As seen with the expansion of applications of "VMware Infrastructure 3" released from VMware, Inc. among enterprise users and the recent increase in attention toward the OSS VMM (Virtual Machine Monitor) such as Xen, the VM (Virtual Machine) technology is becoming practical in the enterprise domain and there are already many instances of actual introductions.

CLUSTERPRO X can be used to implement various configurations when it is combined with VMware Infrastructure 3. Two of the typical examples of these configurations are as follows.

(1) VM-VM clustering

A cluster is formed by connecting VMs (**Fig. 5**). This is used to apply an ordinary cluster configuration to a virtual environment without modification. This configuration is sometimes used with the aim of retaining an existing system (legacy migration) after the H/W maintenance period has expired.

(2) PM-PM clustering

A cluster is formed by connecting PMs (Physical Machines)

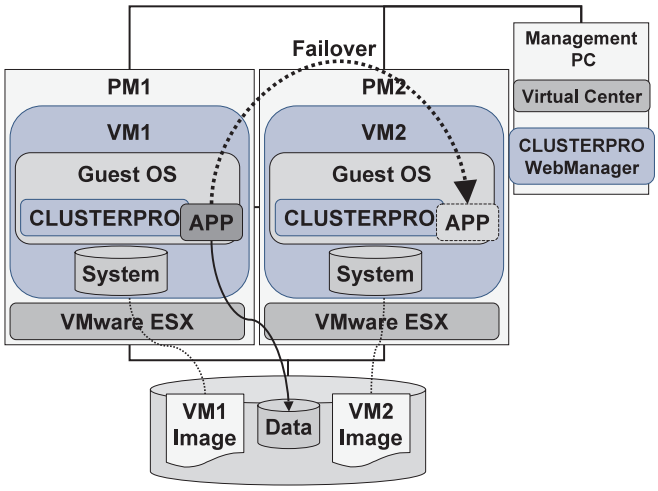

Fig. 5 VM-VM clustering (Shared disk type).

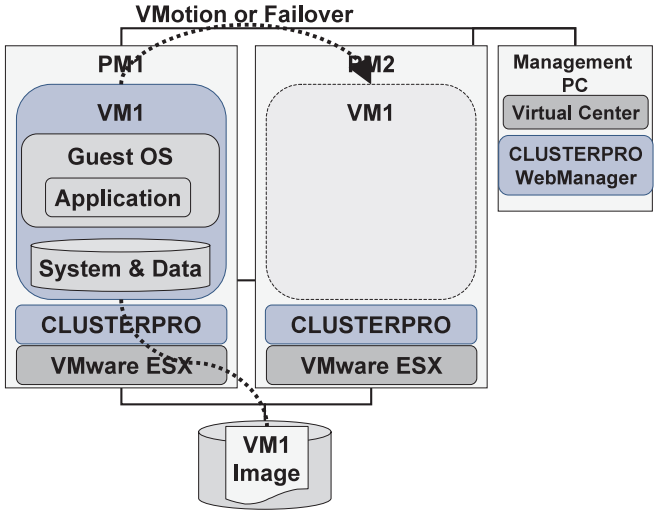

Fig. 6 PM-PM clustering.

in which VMware ESX is installed, and the movement and failover of VM are performed across the PMs (**Fig. 6**). This configuration can improve availability without the need to perform a special setting to the VM itself. However, as the VM is not clustered, detection of irregularities in job services and maintenance of the VM[*2] may pose problems.

7. Process-Level Switching

CLUSTERPRO X has already enabled VM-level switching as described above. From the viewpoint of "granularity of switching targets," we are at present also studying the possibility of providing process-level switching with more detail.

We have already implemented at the prototype level, the technology for migrating the process status as it is by suspending a process running on a server. This is achieved by transferring the image of the memory to server B and resuming the memory.

The system we are aiming at in the future is one that combines with the optimum resource placement technology. It will provide overall optimization, with which a process migrates seamlessly from a server running out of resource to a server with a resource headroom without the user being aware of it.

8. Image of Effective Operations in the Future

Like the linkage of the VM technology we have mentioned above, CLUSTERPRO X is implementing a mechanism for effective operation based on linkage with a blade server or a SSC (Sigma System Center) integrated management tool.

In the future, we are planning to link SSC's resource optimum placement function and H/W provisioning function in order to obtain the following effects (**Fig. 7**).

- Overall optimization, with which CENTERPRO X lets job services migrate automatically to an optimum server, based on the information received from the optimum resource placement function.
- Isolation of a server from the cluster according to the irregularity detection of CLUSTERPRO X, and automatic replacement/repair of the server by the H/W provisioning function.

[*2] Because of the impossibility of providing a rolling maintenance service that is capable of providing patches for a standby VM while at the same time continuing to provide an active VM service there may be occasions when a job should be shut down.

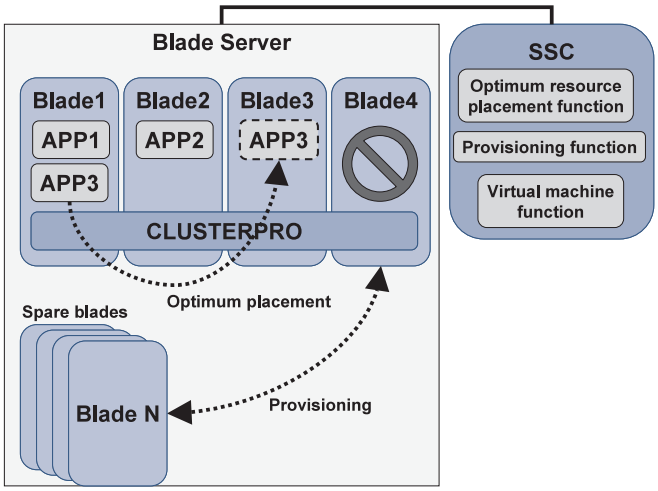

Fig. 7 Blade server/SSC linkage.

9. Conclusion

Since its release in 1996 CLUSTERPRO has continued to lead the PC server clustering field by adopting or linking the latest technologies.

In the future, we are determined to continue to adopt and link new technologies and emphasize our R&D efforts so that we can deliver "technology for implementing 'dependability'" to a wide range of customers and to eventually realize the vision of the "REAL IT PLATFORM."

* The corporate and product names mentioned in this paper are trademarks or registered trademarks of their respective owners.

Authors' Profiles

EGETA Shusaku
Group Manager,
2nd Computers Software Division,
Computers Software Operations Unit,
NEC Corporation

KAITE Ikuo
Manager,
2nd Computers Software Division,
Computers Software Operations Unit,
NEC Corporation

JINNO Eri
Network Software Division,
Platform Operations Unit,
NEC System Technologies, Ltd.