New computer and network system

H. Togawa[1] and T. Hotta[1]

[1]*Research Center for Nuclear Physics (RCNP), Osaka University, Ibaraki, Osaka 567-0047, Japan*

1. Introduction

The computer and network system at RCNP has been replaced with a new one. The rental period is from 1-MAR-2006 to 31-AUG-2010. The logical structure of the new system is shown in Fig. 1. Our three targets are large storage, redundant system, enough security. Our 200 TB of disk storage is intended to store all online data from all experiments in RCNP (including LEPS at SPring-8) and other required data in analysis and theoretical calculations in the half of the rental period (about 2 years). To maintain 99.7 % of availability, we introduce redundancy into any level of any system. For example, RAID, SAN, HA, multi-servers, dual power supply,,, etc. To keep secure environment, we introduce some seciruty system. For example, network authentication, fire wall and audit system.

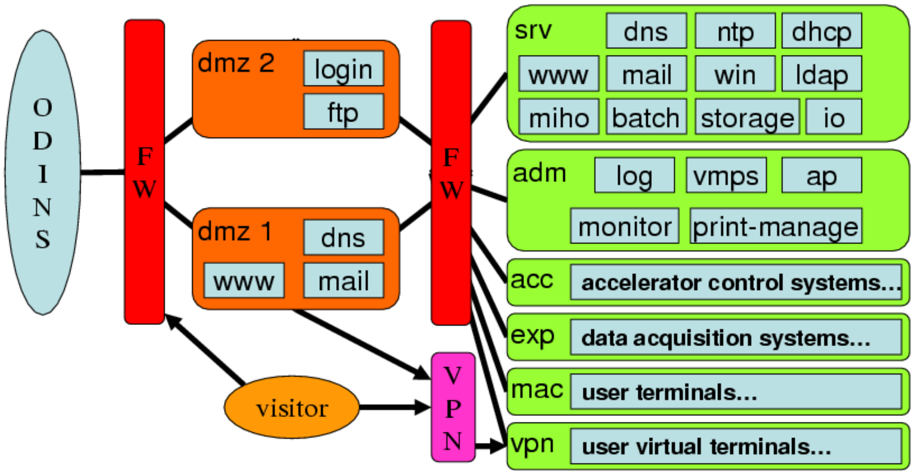

Figure 1: Logical structure of the new computer and network system.

2. Main Servers

2.1 Storage system

Figure 2: SAN switches and 200-TB storage disks.

2.1.1 Capacity and speed

Total capacity is 200 TB for users. The total bandwidth of 24 Gbps is observed in reading 16 files concurrently. The total bandwidth of 14 Gbps is observed in writing 16 files concurrently.

2.1.2 SAN

All user disk of 200 TB are placed under dual SAN switches (Fig. 2).

2.1.3 Redundancy

We configure four 26+P RAID under one controller with 4 hot spare disks. These hot spares are global within controller. Each RAID controller has dual path for SAN switches and dual path for each RAID. All hosts connected to SAN have each path to each SAN switch. If one path is failed, other path is used for disk access. If one SAN switch is down, other SAN switch can serve all disks. Each RAID has redundant power supply. Each controller has redundancy of power suppy and controller card itself. There is no single point of failure in the storage system.

2.1.4 Load balancing spreading on LUN

We create 13 logical disks in one RAID set. We configure one logical disk of about 500 GB to be corresponding to one LUN of SAN. These LUNs are visible from GPFS which is a file system of IBM. We configured one logical volume consist of these LUNs to be distributed maximum over 32 RAID so that DISK I/O should be distributed for maximum I/O performance.

2.1.5 Volume management

The GPFS supports dynamic configuration of logical volume so that the volume size can be increased or decreased by adding or subtracting LUNs to/from existing and operating volumes.

2.1.6 Hot repair

Each component can be repaired (replaced) during operations. In the case of single point of failure, the operation should not be stopped and the failed device can be repaired during operations.

2.1.7 Snapshot

The GPFS supports snapshot function. We take snapshot every night so that user can recover their files which are deleted or modified accidentally. The number of snapshot generation is 7, so user can recover their files up to 7 days ago.

2.2 CPU server

There is one IBM p595 server (Fig. 3) partitioned to two virtual machines. One is an interactive node which has 8 CPUs and 10 GB of memory. Other is a batch node which has 32 CPUs and 122 GB of memory. This partitioning can be changed dynamically during operation. The performance of each CPU is 2796 SPECfp2000. Each node has 4 fiber channel connections to each SAN switch for disk access. Major hardware components can be repaired (replaced) during operation.

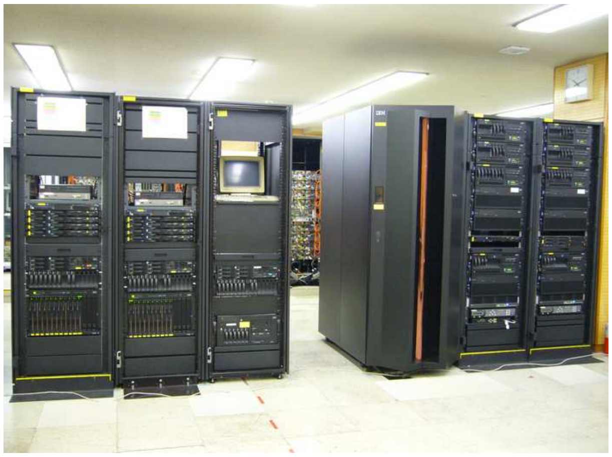

Figure 3: (from left) Two application servers with redundant structure, server for accelerator control, main computing server, and two (login, ftp, and storage) servers with redundant structure.

2.3 Storage server

There are dual storage servers for computers which are not connected to SAN switches. The served protocols are nfs2,3,4 and cifs (samba). Theoretical bandwidth is 6 Gbps for each server. The dual servers constitute a HA cluster. If one server fails, other server will function as both servers.

3. Access servers

3.1 File transfer server

We placed special servers only for file transferring to/from outside RCNP because high-speed file transfer is much heavy load. The file transfer servers support protocols for scp (hpn-ssh) and bbftp. Theoretical bandwidth is 2 Gbps for each server. For redundancy and load balance, we operate two servers but user should select each server.

3.2 Login server

For protect main servers from intrusion, we operate login server to separate main servers from outside RCNP. The login server relays slogin connection from outside to inside servers including interactive node and super computers.

3.3 VPN server

We also operate Remote Access VPN servers. Using CISCO VPN client software, users can get the environment almost identical to inside network of RCNP. Only exception is that only default workgroup "WORKGROUP" can work properly and other workgroups cannot be accessed over the VPN tunnel. For redundancy there are dual servers. The client software can automatically retry to connect other server if one server fails.

4. Application Servers

4.1 WWW server

4.1.1 Inside and outside separation

Because the www server is always to be targeted by intruder, we configured the www server very carefully. The first, we separate www servers to inside servers and outside servers. The outside www server can be accessed from the world wide through the fire wall. The inside www server can be accessed from the inside only. The inside www server also accept reverse proxy connection from the outside www servers. The contents reside on storage servers and accessed from inside www servers. Thus, the contents cannot be altered even if the outside www server is intruded.

4.1.2 Redundancy and load balancing

For redundancy and load balancing, both outside and inside servers are dual systems. The load balancers distribute connections to two sets of www servers. If one www server fails, the load balancers do not route connections to the server any longer.

Because free database softwares MySQL and PostgreS we used do not support load balancing system, we should specially set URL to be routed only to the server which the database software runs.

4.2 Mail server

4.2.1 Inside and outside separation and send and receive separation

For performance and security and various reasons, we separate mail servers to inside and outside, send and receive servers. Then we have inside-send, inside-receive, outside-send, outside-receive servers.

The inside-send servers are smtp servers used when PC and other client will send mails. The access from PCs are authenticated by LDAP server using SSL (smtps). Because our firewall deny direct smtp(s) connections from inside to outside, then all initiated mails should be authenticated.

The inside-receive servers are pop3 and imap4 servers. Which are also accessed from outside RCNP under authentication by using pop3s and imap4s protocols. This server also served mailing lists.

The outside-send server and outside-receive server is separated so that send-server can send mails normally in case that the receive-server is heavy loaded by processing many mails such as SPAMs or DDOS attacks.

4.2.2 Redundant system

Because the mail system is very important for our study, we configured them as fully redundant system.

The inside-send servers are configured as two independent SMTP servers. And we use DNS round-robin from single smtp server name to two SMTP servers. If one of SMTP servers is down, the clients can access successfully with 50 % probability. In case of accessing to failed server, the DNS resolves not-failed server address in a few minutes.

The inside-receive server is configured as active-standby redundant system.

The outside-send and outside-receive server is actually identical to its function. Actually send server can receive and receive server can send. The MX record of the outside DNS controls the receive server primary receives incoming mails and the send server as secondary. The MX record of the inside DNS controls the send server primary receives outgoing mails and the receive server as secondary.

Thus, we constituted the mail system by using easier technique as possible, rather than more complicated method.

4.2.3 Virus protection

All mails are checked by virus check software from trendmicro. Because we have several mail servers, we set the check point at inside-send servers and inside-receive servers. If mails polluted by the virus are found, that mails are silently deleted.

4.2.4 SPAM protection

All mails are checked by spamassassin and are added the headers which contain the "score". Because we are vary afraid of mis-discarded mails by spamassassin, all mail is classified into three kinds.

1. Not a spam mail, that score is less than 8, and delivered to user.
2. Doubtful mail, that score is between 8 and 40, and delivered to user.
3. Concretely spam mail, that score is over 40, and delivered to trash.

A user can select mails by setting filter of their MUA referring the "score". If important mail is delivered to trash, it can be recovered by SE on request from user in seven days.

4.3 DNS

Because our network is completely intra-net configuration, outside name space and inside name space are different. Then we should have 2 name servers for inside and outside. For redundancy there are two inside servers and two outside servers. To disclose inside servers, the query from inside servers are recursively transfered to outside servers.

4.4 dhcp server

The dhcp servers offer static IP address only to registered MAC address. For redundancy there are two dhcp servers. These servers offer the same IP address for the same MAC address. The DHCP relay agents are working at the FW module to service for multiple segments.

4.5 ntp server

There is a ntp server which have GPS time source (stratum 1). There are two ntp servers, which associated with stratum 1, for general clients and servers.

4.6 Authentication servers

Our main authentication method is LDAP. We manage almost accounting information on the LDAP. The other authentication method RADIUS is required for wireless access point and remote access VPN, etc where RADIUS is referring to LDAP to make single authentication domain. For redundancy there are two LDAP servers with load balancer and two RADIUS servers.

5. Sub-systems

5.1 Online station

For the last stage of the online data acquisition systems, we introduce two set of computers called Online Station. The purpose of the Online Station are

1. Collect all data from front-end online machines over the network.
2. Store all data to the storage system over fiber channel.
3. Analyze data to check the experiment.

There are two set of online station for redundancy and pre/post processing of the experiment. The two systems are sharing the same storage over the SAN (fiber channel) so that the other system can be used immediately if one system is broken.

5.2 I/O station

In spite of decreased use of the off-line media, the tape backup is necessary for some users. We introduce I/O station so that users can use DLT8000, DAT72 and LTO-3. We introduce exclusive device allocation system called "mtlock" to prevent accessing from others accidentally. We also introduce "chgmrt" command to set device mode such as density and compression.

5.3 Accelerator control sub-system

We introduce a part of the accelerator control system. This time we prepare the database server which can store all the accelerator control parameters and running data at real time and continuously.

5.4 Windows sub-system

This system is intended to serve applications running only on Windows. Currently Symantec Anti Virus management software and Vector works OPERA license server are running.

6. Printers and scanners

We introduce 10 sets of daily use printers which can print full color images to A4 sheet by double sided mode with speed of 24 ppm. We also introduce 2 sets of high-speed and multi-sheet printers which can print full color images to A4, B4, A3 sheet by double sided mode with speed of 35 ppm. The "Poster Printer" is also introduced which can print A0 and B0 roll sheet. There are three multi-function machines which can be used as a copy machine and a scanner not only as a printer. One of them can be used as a facsimile.

7. Network

Physical structure of the network is shown in Fig. 4 and a picture of network equipment is shown in Fig. 5.

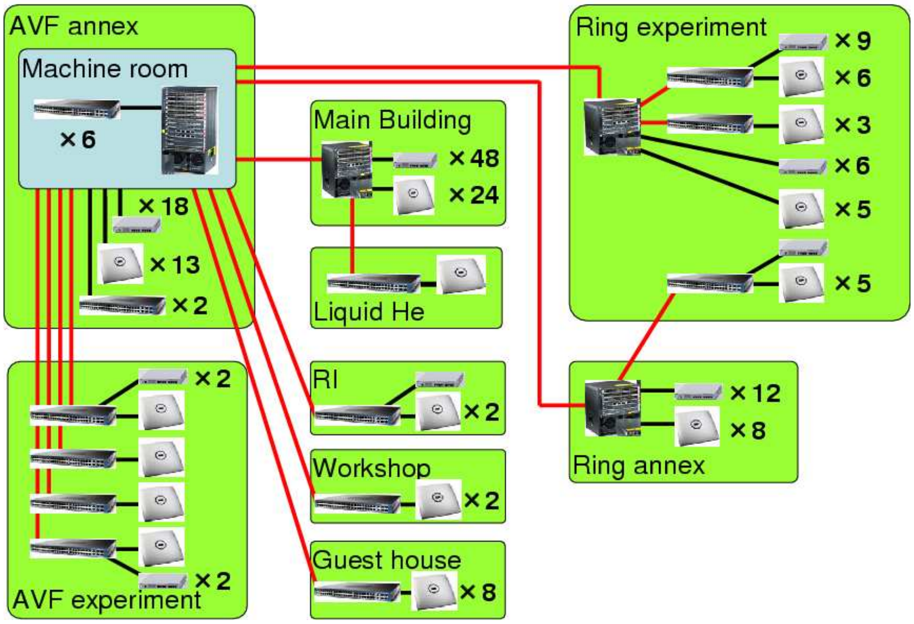

Figure 4: Physical structure of the network system.

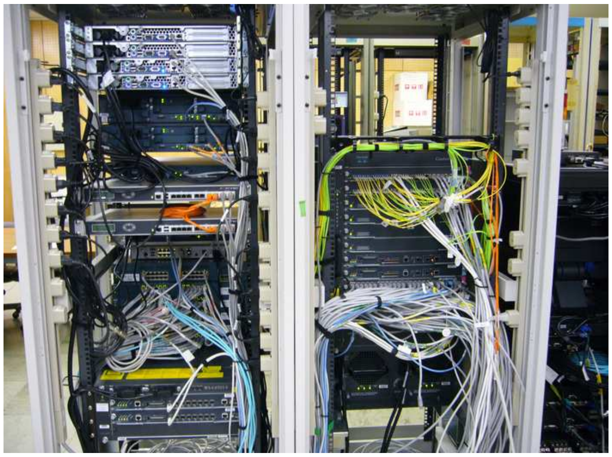

Figure 5: Core network equipment.

7.1 Firewall

We designed that the firewall should be exist between any networks of defferent security level. The external firewall separates DMZ network from outside network (i.e. campus network "ODINS"). The internal firewall separates all inside networks including DMZ each other. Only exception is VPN server. But VPN server is a security device itself which is equivalent to a firewall.

The external firewall and the internal firewall have taken the redundant configuration by two Fire Wall Service Module (FWSM), respectively, then there are four FWSM in Catalyst 6513 in total.

7.2 Core switch and Sub-core switch

There are 1 core switch at the machine room of AVF building and 3 sub-core switches at the Main building, RING experiment building and RING annex building. All core and sub-core switches are Catalyst 6500. For bandwidth requirements and redundancy, sub-core switches at the Main building and RING experiment building are connected to core switch with 6 of 1 Gbps fiber. The sub-core switches at the RING annex building is connected to core switch with 4 of 1 Gbps fiber.

7.3 Middle switch

For small building or large experiment room, there are 13 middle switches. These switches are function equivalent to core switches but service to small area. Middle switches are Catalyst 2960 or Catalyst 3550.

7.4 Edge switch and wireless access point

The edge switches are distributed in all area and directly connected to user terminals or equipments. The Allied Telesis GS908M supports 8 1000BASE-T ports for high speed requirements.

The wireless access point (CISCO AP1131AG) are also distributed in all area and supports IEEE802.3a, b and g concurrently. This AP also support multiple SSID corresponding to VLAN ID. It makes us to operate different purpose of networks in single AP. Because the multiple SSID shares single radio channel, the channel assignment is very easier.

7.5 Redundant network

All core and sub-core switches have redundant power supplies and controller modules. All connections between core, sub-core and middle switches use at least 2 fibers, which are connected to different modules in core or sub-core switches in case for the single module failure.

7.6 MAC address filtering

To construct secure network, we decided to introduce the MAC address filtering to exclude un-registered network equipment. The core, sub-core and middle switches have a CISCO-VMPS function. These switches carry only packets which MAC address have been registered. If a packet which MAC address is not registered comes at the port of the these switches, the packet will be discarded. For wireless network, the access point performs MAC Address authentication. In this way, unregistered equipment cannot communicate at all except in a single edge switch.

8. Management and Operation

8.1 User management

We have created the user management system which can manage all the life cycles of user management.

Anyone can apply for new registration using web interface. It is required for user to have a mail address. If not, the deputy can apply instead. Also required to specify the contact person who is a staff member of RCNP. The contact person judges whether his/her use of a computer and network is permitted. In case of the student, the permission of his/her supervisor is also required.

The user can update his/her own personal information registered on our database at any time using web interface.

All users are required to update his/her registration around the end of fiscal year. The user who don't update will be deleted soon.

The administrator can manipulate user application, inquiry for the contact person and supervisor if required. And can make, disable, enable, delete and revival the account. Of course can modify the account information and so on.

All the result of management system will be mailed to the user and the administrator so that they can check it.

8.2 Network management

All user who have account in our system can register MAC address of the network equipment using web interface at any time. This registration does not require the confirmation of the administrator, it takes a few minutes to take effect of the registration. The user can use the management system which can register, list and delete a MAC address.

8.3 Network traffic monitor

The network traffic monitor system is introduced to keep track the port base traffic (bps) for every minutes. This helps us to find a network trouble and security issue, and to report network usage. Almost all ports are monitored, so that we can find the traffic by each hubs (effectively equivalent to each room in the most case), each APs, main servers and the traffic across the firewall module.

8.4 Automated operation

Because our system consists of many components, it is much complex to turn on/off by human whole the system after/before power failure or other maintenance reasons. It is a problem especially in case of emergency, such as failure of air conditioner, a fire, an earthquake and power failure.

We designed two shutdown method. One is normal shutdown to insure all data to be safe. This method is intended to use in case of a failure of air conditioner or an electricity supplied by UPS in an accidental power failure.

Other is emergency shutdown to insure hardware not to be broken. This method is intended to use in case of a fire or an earthquake.

The “environment monitor system” has the fire sensor, the smoke sensor and the thermometer in several places of the machine room and monitor the UPS status including power failure. It will tell a normal shutdown or emergency shutdown to the “control PC” according to the condition. The control PC will stops each component according to pre-defined procedure.