Computing at KEK

Y. Morita [a]

[a] *KEK, High Energy Accelerator Research Organization,*
1-1 Oho, Tsukuba, Ibaraki 305 Japan

In 1996 and in 1997, KEK Computing Research Center installed two large UNIX workgroup cluster systems. Computing and data storage resources are partitioned and allocated to each research group with independent DCE cell administration. Data transfer between different workgroup clusters are established with inter-cell communication. These systems are the largest installation of DCE and DFS which are being used in production, in combination with the hierarchical storage management software on 20-TB and 160-TB tape library mass storage systems.

Key words: DCE; DFS; HSM; MSS; Tape Library; DVD

1 Introduction

In April 1997, National Laboratory for High Energy Physics (former KEK) and Institute for Nuclear Study (former INS) are merged into one entity, High Energy Accelerator Research Organization (new KEK). Research activities at the new organization cover wide range of fields, such as Belle experiment at KEK B-factory, proton synchrotron fixed target experiments, LHC/Atlas collaboration, Kamiokande long baseline neutrino experiment, heavy ion beam experiments, and various material structure science using Photon Factory and Meson and Neutron Facilities.
It is the aim of the KEK Computing Research Center to support these wide range of research activities with different computing requirements.
Two large UNIX systems are installed to provide computing and data storage resources to these research groups. Central Computer System is replaced from a Hitachi mainframe system to UNIX workgroup clusters in January 1996, and Belle Computer System is installed in January 1997, replacing Fujitsu mainframe system for the former TRISTAN experiments. There also is a Fujitsu VPP 500/80 super computer system (128 GFlops).
Each of the above system is replaced with a bidding process in every four or five years. [1]

[1] There is another computer system at KEK Tanashi Campus, former INS.

Computing Research Center also supports several public login systems, and domestic and international computer networks.
Figure 1 shows the configuration of KEK Computing Research Center.
To provide high speed data storage capacity for large number of users, we have chosen helical scan high performance tape drives using video broadcast technology as our primary mass storage system [1].

2 Central Computer System

Total CPU power of the Central Computer System is 16,000 SPECfp92, with a total disk capacity of 1 TB. The system resources are divided into ten workgroups (Fig. 2).
Each workgroup cluster consists of several computing servers and data servers, which belong to a single DCE cell. User home directories and group disks are shared within the cluster with DFS [2].
There are two sets of SONY DTF tape library systems for the storage capacity of 12 TB and 8 TB. These systems are connected to four clusters. Hitachi JP1/OmniStorage hierarchical storage management software is used in conjunction with DCE and DFS. Users or user groups can also mount their tapes from the tape library directly in their analysis processes, to make use of the high I/O speed of the SONY DTF tape drives (12 MB/s) [3].
Computing servers are mixture of Hitachi, Hewlett Packard, and DEC server workstations. Data transfer between different workgroup clusters are achieved with intercell communication over FDDI network.

3 Belle Computer System

The Belle Computer System has a total CPU power of about 50,000 SPECint92 Fujitsu AP3000 system with eight SMP servers. A data server with 35 CPU nodes interconnects 4.2 TB Gen5/XL RAID disks and 160 TB SONY helical scan tape drive robotics. Belle raw data from the experimental hall will be transmitted to the tape library with 3 km optical fiber link [4]. SONY PetaServe, a modified version of Computer Associate's OSM, will be used as hierarchical storage management software [5].
Servers in this system are interconnected with each other with ATM switch networks and Fujitsu APnet. Desktop workstations of Belle collaborators on site or at domestic and foreign universities is connected to the system with DFS running on ATM or FDDI networks, or on wide area networks.

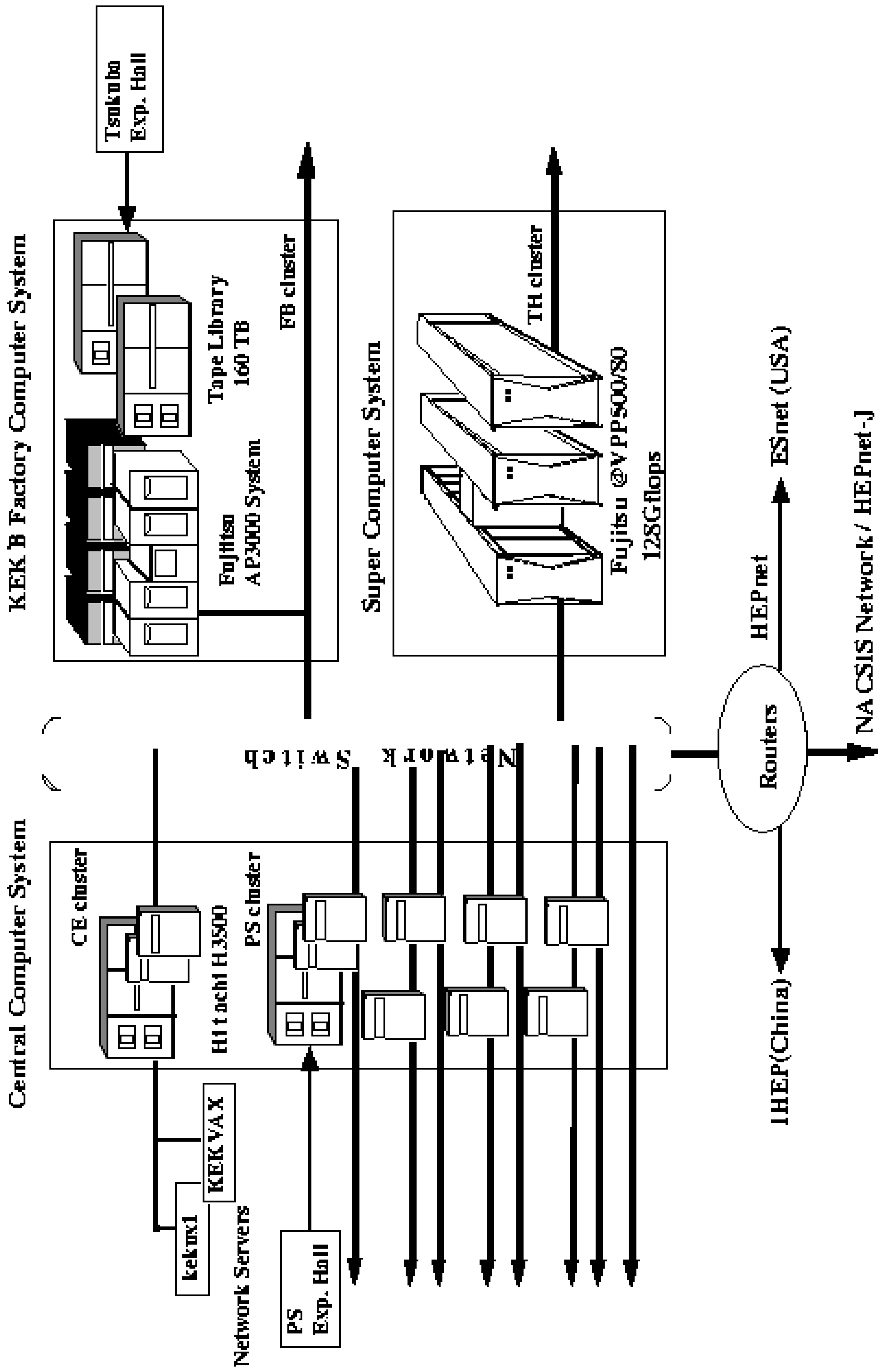

Fig. 1. System Configuration at the Computing Research Center (April, 1997)

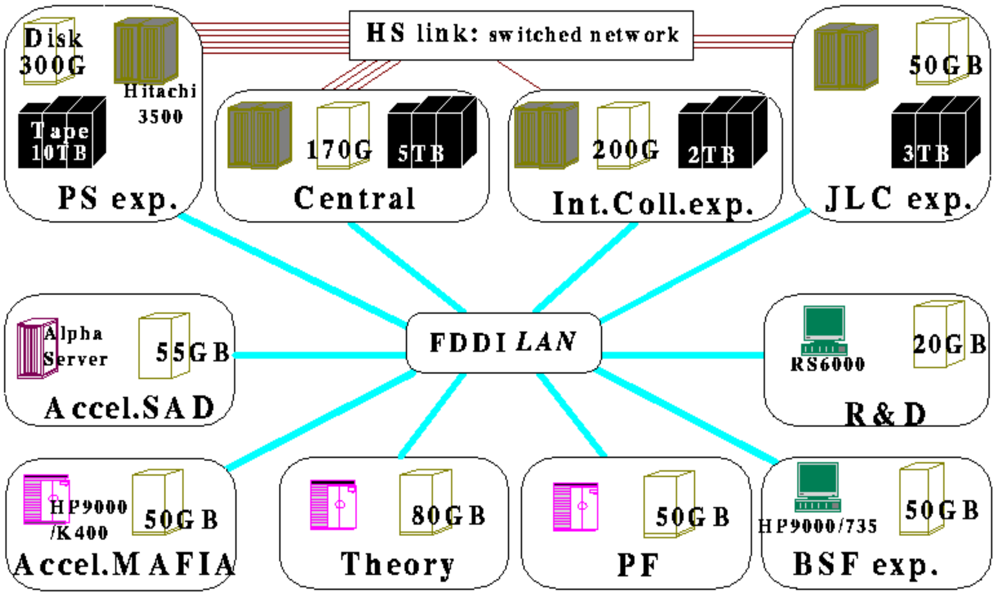

Fig. 2. KEK Central Computer System

4 Network Systems

We have 14 separate FDDI clusters covering the different part of the laboratory. They are interconnected with GIGAswitches. There also are several ATM switches covering local area LAN-emulation and Classical IP over ATM. Belle group makes an intensive use of the ATM local area network in their system, as described above. We also have the wide area ATM permanent virtual circuit links to several major Japanese national universities over NACSIS Network, a Japanese academic science network.

There are also smaller public service systems for network utilization, such as e-mail, ftp, file servers, WWW servers, and several kinds of public login services on VMS and UNIX. Telephone dial up services are provided for analog phone lines and ISDN lines.

5 Storage Devices and Future HEP Computing Model

As the cost performance ratio of recent personal computers are dramatically improved, it is becoming a quite popular practice to use several high-end personal computers for analyzing HEP data at several experiment groups such as Fermilab, DESY and CLEO. Recent explosion of the business PC markets, and serious competitions on prices and performance of the PC's, will have positive and favorable effects on HEP computing. Within a few years, most CPU cycles of the HEP data analysis will be given by PC farms.

It would be a potentially serious problem to handle large amount of HEP data when one wants to use PC farms for HEP data analysis, especially when the amount of HEP data reaches to the order of peta bytes per year. High-end UNIX server workstations may still play a role to provide these large amount of HEP data to PC farms from high speed storage devices, such as RAID disks, or high-end magnetic tape drive robotics. Or there is a possibility that a large business market would appear such that a mass storage system of order of peta bytes would be accessed directly from high-end PC servers themselves.

Whether running operating systems such as Linux, or running on commercial based operating systems such as Solaris or Windows NT, it would become one of the central issue to provide high speed data access to large and complex HEP data storage system from economical CPU cycles.

On the other hand, as the market of PC gets matured, we have to keep eyes on new kinds of storage technologies. High-end magnetic tapes will remain as high speed and large capacity archival media for next several to ten years. Tape robotics will provide peta byte hierarchical storage mechanism. However, a market is maturing very quickly for a new technology of storage media to handle relatively large amount of 'multimedia' data on PC's. One of the example is DVD, Digital Versatile Disk, an optical disk of the size of CD-ROM. This new media will store and play motion pictures and audio data on PC's and on consumer electronics products.

6 DVD as Alternative HEP Data Storage

One of the features of DVD is its random-access-ness to the data. Data on DVD can be accessed track by track, just like on CD-ROM or on magnetic disks. DVD media also has the same physical size as CD-ROM media, therefore the drive makers are able to make their DVD drives compatible with both DVD and CD-ROM media. There are three types of DVD standards currently proposed: DVD-ROM, read-only media like CD-ROM, DVD-R, write once media like CD-R, and DVD-RAM, rewritable media like magnetic disks.

DVD media can be made single sided or double sided. Storage capacities of DVD disks per side are 4.7 GB for DVD-ROM, 3.9 GB for DVD-R, and 2.6 GB for DVD-RAM. Capacities of DVD-ROM and DVD-R will increase by 50 to 100% in a year or so.

Data I/O speed of a DVD drive is 1.2 to 1.4 MB per second, which is relatively slower compared to magnetic disks or high-end magnetic tape drives. However, the I/O speed can be increased by combining several DVD drives as a RAID drive. For

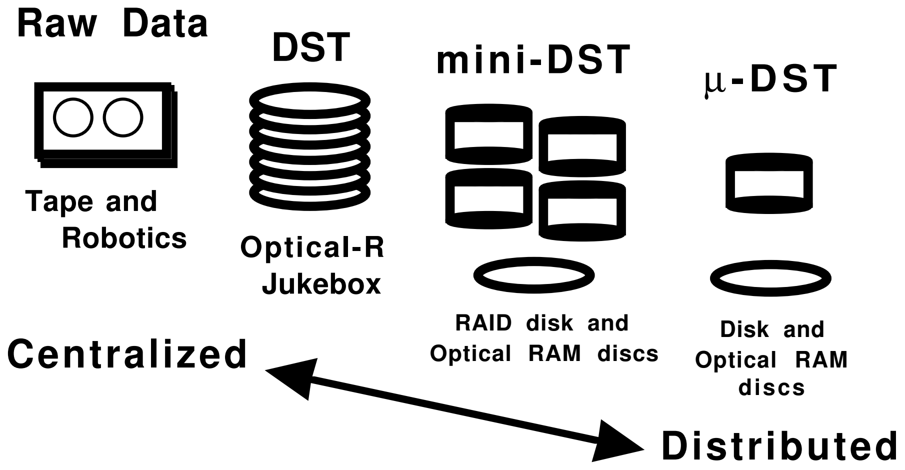

Fig. 3. Storage Media vs. Data Reduction Stages

example, RIKEN is investigating a 4-way DVD-R RAID drive on a DVD jukebox to handle RHIC experiments data in the regional computing centers [6].

There is a possibility for DVD drives to replace CD-ROM drives in the PC markets, as the needs for handling multimedia data increases. It is a common practice today for commercial PC magazines to provide CD-ROM media as a supplement of the magazine. Imagine that these magazines will issue DVD-ROM supplements instead to provide multimedia data in near future.

As HEP experiments become truly world-wide collaborations, it would become a potentially serious problem to provide experimental data to each collaboration member in each member country. At the experimental site, raw data and so called data summary tapes will be stored at a central and well-controlled location, to a high-end hierarchical storage system, such as tape robotics. While as the data reduction process proceeds, it would become favorable, or necessary, for downstream physics data to be distributed with economical random access media such as DVD disks (Fig. 3).

7 Conclusions

Computer systems at KEK will be replaced with a bidding process in every four or five years. That is, the system must catch up the advancement of the computer technology in the market in every four to five years. A first major transition from a center-supported mainframe system to the distributed UNIX workstations occurred at KEK in January 1996 as a new Central Computer System. The system consists from ten workgroup clusters and each cluster is administered as a separate DCE cell. The CPU, disk and hierarchical storage resources in the cluster are shared by DFS services. The system is proven to be a stable UNIX environment for large number of users with a simultaneous access to 20 TB of hierarchical storage sys-

tem. Direct tape drive allocation and tape file access mechanism is also supported. Belle Computer System is introduced in January 1997, and the 50,000 SPECint92 system is running with high speed hierarchical storage system consisting of 4.2 TB Gen5/XL RAID disks and 160 TB SONY helical scan tape drive robotics.
As the focus of the computer markets shift towards the business computing, a possibility to make use of cost effective CPU cycles on personal computers and new storage devices in the center-supported environment is being sought for.

References

[1] H. Fujii *et al.*, in: CHEP'95, ed. R. Shellard and T. D. Nguyen, (World Scientific, Singapore, 1995) p.36.

[2] S. Yashiro, T. Sasaki *et al.*, in this proceedings.

[3] Y. Morita, S. Yashiro *et al.*, in this proceedings.

[4] H. Fujii, R. Itoh *et al.*, in this proceedings.

[5] A. Manabe, H. Fujii *et al.*, in this proceedings.

[6] Y. Watanabe <watanaby@rikvax.riken.go.jp>, *RHIC experiments meeting*, private communication