Computer network system with security for a protein data collection system at the Photon Factory

K. Sasaki,[a]*† N. Watanabe,[b]† N. Sakabe[b]† and K. Sakabe[c]†

[a]*School of Informatics and Sciences, Nagoya University, Furo-cho, Chikusa, Nagoya 464-8601, Japan,* [b]*Photon Factory, Institute of Materials Structure Science, High Energy Accelerator Research Organization, Ibaraki 305-0801, Japan, and* [c]*Department of Chemistry, Faculty of Science, Nagoya University, Furo-cho, Chikusa, Nagoya 464-8602, Japan. E-mail: sasaki@info.human.nagoya-u.ac.jp*

In 1997 the prefabricated house of the TARA Sakabe project was constructed very near to the Photon Factory ring, and many computers were installed for crystallographic data handling. A data server with high speed and large capacity was required to improve the efficiency of the protein data collection system which integrated a 'high'-security computer network. The new network, based on a 100 Mbps Ethernet, consists of a DEC AlphaServer 4000 with a 115 Gbytes RAID disk, DLT as a backup device, CISCO PIX-32 as a firewall between the TARA private network and KEK, and a 100 Mbps switching hub to be linked to imaging-plate readers and workstations. Therefore, the digital output data from the imaging-plate reader are directly recorded on the server disk resulting in higher efficiency of the users' beam time. In contrast to recording on tape, there is very little problem with backup resulting in a high confidence in the data-collection system.

1. Introduction

The protein data collection system at the Photon factory (PF) was first constructed using a Weissenberg camera combined with an imaging plate (IP) of dimensions 20×40 cm^2 at BL6A in 1986 (Sakabe, 1991). In this system, a Fuji BA100 was used as the IP reader. This reader was operated by a PC9801 personal computer and digital output data were stored on open-reel magnetic tapes; each user took the tapes back to their own laboratory to process the data. In this period all data were stored on magnetic tapes, thus the user had no worries about the safety of the data. In 1996 a new image reader, a Fuji BAS2000, which could read the same size of IP, was introduced for protein data collection on BL6A.

In 1993, BL18B, the second protein crystallography station, was also constructed, and a time-resolved Laue camera, which could be used as a Weissenberg camera, was installed in the hutch (Sakabe *et al.*, 1995). Two R-AXIS DS 40L image readers, which

---

† Guest researcher for the Sakabe project of the Tsukuba Advanced Research Alliance (TARA), University of Tsukuba, Japan.

are a commercially available type of IPR4080 (Sakabe *et al.*, 1997) modified by Rigaku Co. Ltd, with large-formatted IPs of dimensions 40×80 cm^2 and the BAS2000 were installed, and the intensity data were recorded using the disk space of the image reader's computer. The user transferred these data to DAT or to an 8 mm tape which they took back to their own laboratories.

In 1996 an additional protein data collection station, BL6B, was constructed by the Sakabe research project to study structural biology at the centre for the Tsukuba Advanced Research Alliance in the University of Tsukuba (TARA Sakabe project). Here, a dedicated Weissenberg camera with a single cassette of radius 575.7 mm and two R-AXIS DS 40Ls were installed. With this camera, one or two large-formatted IPs can be attached to the cassette and one large-formatted IP is usually used for data collection. Users often use two large-formatted IPs at once which greatly increases the volume of data recorded (Sakabe *et al.*, 1997). At around the same time, a prefabricated house was built very near to the PF ring by the TARA Sakabe project and many computers were installed in this house for data processing and crystal structure analysis. Thus, the construction of a new private computer network system between the PF experimental hall and the TARA house (Fig. 1) was necessary in order to improve the speed of the data transfer and the security and handling of a large volume of data. Here the construction of the computer network and its evaluation is described, and future plans, including a new beamline BL6C for the TARA Sakabe project, is discussed.

2. Construction of the network system with 'high' security

The intensity data recorded on an IP using a Weissenberg camera or Laue camera produces output to disk as digital image data. The volume of digital image data per sheet of large-formatted IP (80×40 cm^2) is 64 Mbytes. It is normal that one data set for one crystal consists of more than 20 sheets of IP and the total volume of the data set becomes more than 1 Gbyte. In order to reserve these data, a DEC AlphaServer 4000 was introduced into the TARA house which was built at a distance of ~35 m from the PF ring, as shown in Fig. 1. The private network system was constructed as shown in Fig. 2. The first priority of the network system demanded by the project was high security.
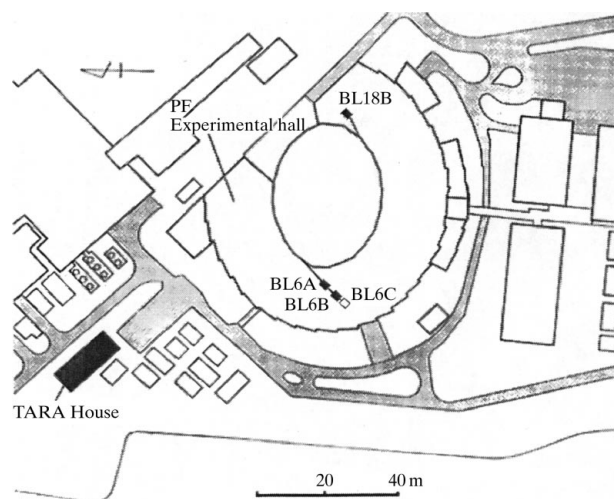

Figure 1
Location of the beamlines for protein crystallography (BL6A, BL6B, BL6C and BL18B) at the PF experimental hall, and TARA house where the private computer network with 'high' security was constructed.

The digital data from the R-AXIS DS 40L are written to the RAID disk of the DEC AlphaServer 4000 using NFS. The reserved data are backed up automatically once a day on DLTs. The backup files on the DLT are preserved for one year. In case of any problems concerning the backup data, a service to restore the user's files from DLT to DAT is provided for the user by the TARA Sakabe project. The capacity of the RAID disk of the server was 115 Gbytes to be unified as one volume and the RAID disk system at level 5 was introduced in consideration of performance, safety and availability because small computations are carried out at the same time as data transfer. Regarding the calculation of the disk space, we assumed that the minimum period of reserving data on the RAID disk is 3 d. During user beam time almost all data on the disk are produced by large-formatted IPs. It takes 12 min to read out one large-formatted IP by one IP reader. Thus 120 sheets of the IP can be read in 1 d. The number of IP sheets from five R-AXIS DS 40Ls is thus 600 sheets, so that the volume of digital data becomes 38.4 Gbytes. A 115 Gbyte RAID disk fills up completely in 3 d. One of the good characteristics of this disk system for this data-collection system is that if any problems on the RAID disk occur it is automatically replaced by an assistant disk, and in this system we can change the disk without stopping the server. Furthermore, most of the workstations in this new private network system are connected to RAID disks with NFS, so that the handling of the data is very easy. Preservation of a large volume of data on the RAID disk of the server requires security. In order to raise the security of the network, a CISCO PIX-32 firewall was introduced (Comer & Stevens, 1995).

In order to speed up the backbone of the network, a 100 Mbps Ethernet was introduced. Regarding the estimation of the flow rate of the data files on the network, one of the main types of data files transferred on the network are digital data from large-formatted IPs produced from R-AXIS DS 40Ls. The rotation speed of the 100 cm-circumference dram of the R-AXIS DS 40L is 8.33 rotations s^{-1}. Thus the transfer speed of the data is 133.3 kbytes s^{-1}. There are five readers in the beamline. If all readers are used at once the flow rate of the data on the network is 5.3 Mbps, which is not enough for a 10 Mbps Ethernet because the efficiency of the network is normally less than 50%. Moreover, data checking and data processing (two images) are carried out during IP reading. At that time the total volume of the data on the network was 640 Mbytes. Assuming the efficiency of the network is 50%, the required time to transfer these image data is 102.4 s. Thus it is necessary to introduce a 100 Mbps Ethernet. In order to suppress the increment of traffic of the network (Comer & Stevens, 1994) we connected the computers of BL18B to one sub-network and those of BL6A and BL6B to another sub-network against the data server as shown in Fig. 2. There are two BAS2000s which are connected to SUN computers at BL6A and BL18B. These IP readers can read small IPs (20×40 cm^2) and the digital data are stored on the local disk space of the computer until they are transferred to the RAID disk of the server by FTP. The transfer of these image files does not effect the network much in comparison with that of large-formatted IP data files. Optical fibre cable was used for the connection between the PF experimental hall and the TARA house, which are connected by 100 Mbps FDDIs. The 100 Mbps switching hub was set at the point where a large volume of data flows. The network was designed so as to connect five IP readers controlled by SGI Indy and 35 computers including computers of private research rooms in the TARA house for data processing and structure analysis. This network
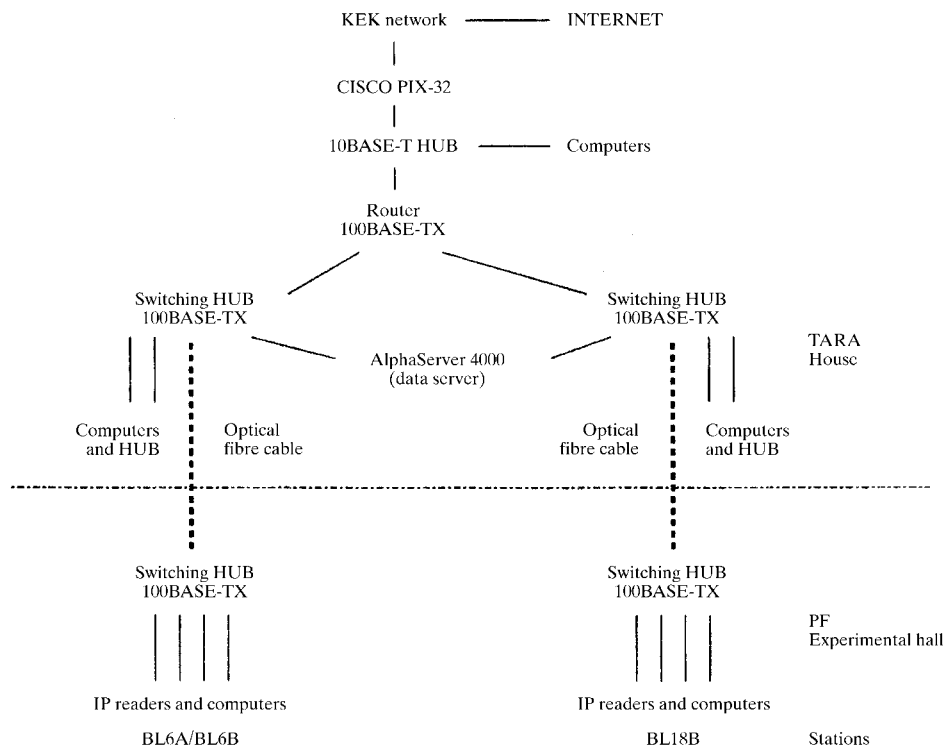

Figure 2
The computers and network instruments shown above the dotted line are installed in the TARA house and those shown below are installed in the PF experimental hall. Computers in the TARA house are connected to BL6A/BL6B and BL18B stations in the PF experimental hall by the optical fibre cable *via* a switching hub (100 Base-TX). Computers for BL6A/B and BL18B are connected to a DEC AlphaServer 4000 equipped with a 115 Gbyte RAID disk *via* a sub-network.

connects to the High Energy Accelerator Research Organization (KEK) network through CISCO PIX-32.

3. Results and discussions

This data-collection system has been in use since the users beam time started in October 1997. There have been no problems accessing the computer systems without authorization because of the use of PIX Firewall. Furthermore, the digital output from the IP reader is directly written to the RAID disk space of the server using NFS, so problems caused by backup using media by the user become very small. The efficiency of the synchrotron radiation beam usage becomes high because it is not necessary to transfer data to another computer or DAT during the experiment. The increment of the efficiency is very important because the domestic user can only occupy BL6B for half a day or a day and BL6A/ BL18B for a day.

The statistics of backup data by DLT per day in user beam time between November and December 1997 are shown in Fig. 3. The maximum of the data per day was approximately 40 Gbytes, the same value as that when all large IP readers are working for 24 h. In this period the volume of average backup data by DLT was 13 Gbytes day^{-1} and then it was possible to store data on the RAID disk for a week on average. The volume of the backup data was 7 Gbytes day^{-1} in the user beam time from January to March in 1998 and 6 Gbytes day^{-1} in the period from April to July in 1998 on average.

There is one workstation for each beamline and six workstations in the TARA house which are now controlled by NIS/ NFS, and software for data processing are installed on all workstations. The general programs for protein crystal structure analysis are installed on three workstations in the TARA house. However, all workstations are used for extensive on-line data analysis and data processing using *Denzo* (Otwinowski & Minor, 1996) and *WEIS* (Higashi, 1989) in user beam time, and other
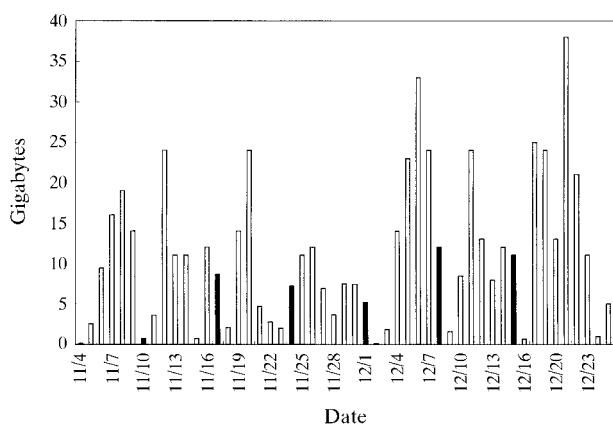
computations are not generally carried out during user beam time. Normally, at first, image file data from the R-AXIS DS 40L are written to a RAID disk and extensive on-line data analysis and data processing are carried out with these data files on the RAID disks. Thus the condition of the network system can be estimated by the backup data by DLT (Fig. 3) and this information will be very useful for the design of advanced network systems in the future. With regard to restoring users' backup data from DLT to DAT, there have been three demands per year.

A fully automatic data-collection system (Sakabe *et al.*, 1997) with two IP cassettes will be installed on BL6C (Fig. 1) soon. In this system the radius and width of the IP cassette are 400 mm and 450 mm, respectively. Thus the size of the IP is 2512 mm × 450 mm and digital data from one IP amounts to 226 Mbytes. According to the time schedule of this camera system, digital data from one cassette will be able to produce 0.226 Gbytes every 30 min on average. If this data-collection system work for 24 hours a day, an additional 10.8 Gbytes of RAID disk space is required per day. We anticipate that once this is fully in operation, including data processing, the case may arise where the assigned data-processing parameters are not suitable, resulting in useless structure factors. Thus it is necessary to reserve the image data on the RAID disk for at least 3 d for extensive analysis of the data, and for processing again with the same image data, because image data are very valuable, especially in protein crystallography, and there may not be another chance of obtaining a crystal of the same quality. Owing to the above reasons, an additional 32.5 Gbytes of RAID disk space is necessary. The disk space taken up by the resulting structure factors is very small, so we can send them to the user's laboratory *via* the internet. In this stage, Patterson map calculation, Fourier map calculation, the assignment of three-dimensional structures using electron-density maps and least-squares refinement will be carried out using these computers during user beam time, and thus a higher speed of data transfer is important for these calculations. We are planning to construct an advanced network system with an additional higher-speed data server in the financial year 2000.

Figure 3
The backup data by DLT from the RAID disk of the DEC AlphaServer 4000 every day is shown in the period between 4 November 1997 and 25 December 1997 in beam time. There is no beam between 9:00 in the morning of the day shown by the black bar (Monday) in the figure and 9:00 in the morning of the following day. This backup data by DLT is data recorded on the RAID disk after the last backup. In this period the average backup data by DLT is 13 Gbytes day^{-1} and then it is possible to store data on the RAID disk for one week.

References

Comer, D. E. & Stevens, D. L. (1994). *Internetworking with TCP/IP*, Vol. II, *Design, Implementation and Internals*, 2nd ed. Englewood Cliffs, New Jersey: Prentice-Hall.
Comer, D. E. & Stevens, D. L. (1995). *Internetworking with TCP/IP*, Vol. I, *Principles, Protocols and Architecture*, 3rd ed. Englewood Cliffs, New Jersey: Prentice-Hall.
Higashi, T. (1989). *J. Appl. Cryst.* **22**, 9–18.
Otwinowski, Z. & Minor, W. (1996). *Methods in Enzymology*, edited by C. W. Carter Jr & R. M. Sweet, p. 276. New York: Academic Press.
Sakabe, K., Sasaki, K., Watanabe, N., Suzuki, M., Wang, Z. G., Miyahara, J. & Sakabe, N. (1997). *J. Synchrotron Rad.* **4**, 136–146.
Sakabe, N. (1991). *Nucl. Instrum. Methods*, A**303**, 448–473.
Sakabe, N., Ikemizu, S., Sakabe, K., Higashi, T., Nakagawa, A., Watanabe, N., Adachi, S. & Sasaki, K. (1995). *Rev. Sci. Instrum.* **66**, 1276–1281.