FOCUS—A remote access file handling system on-line to a CDC 6000 series computer

D. Ball, P. M. Blackall*, Valerie Gerard, G. R. Macleod, P. J. Marcer†, and E. M. Palandri

CERN, Geneva, Switzerland

The paper describes a system to provide multi-access facilities for file creation, storage and manipulation on a medium sized computer, together with remote job submission to a large batch processing computer. Within this framework, the system also provides file facilities for the accumulation of sample data from a number of remote process control computers which acquire data in real time in physics experiments and which are connected to the file handling machine by very high speed data links. The system has been entirely implemented at CERN on a CDC 3100 computer connected to a CDC 6600 computer and a CDC 6500 computer.

(Received December 1969)

1. Introduction

Over the last few years the use of on-line computing techniques has become rather widespread in high energy physics experiments. This, combined with the use of new particle detectors has led to a considerable increase in the amount of data which can be recorded in an experiment. The new detectors (such as, for instance, counter hodoscopes and spark chambers) enable direct digital measurements to be made of points on the trajectory of nuclear particles. The digital data is recorded in real time on magnetic tape for subsequent processing by a computer. A small process computer (such as, for instance, a PDP8, PDP9, IBM 1800 etc.) is often used to buffer the data and to carry out control and monitoring functions on the performance of the detection equipment. A recent review (Lord and Macleod, 1969) gives a more detailed description of the computing techniques and the data rates involved, so they will not be further elaborated here.

The FOCUS system was designed to allow data recorded by the small data acquisition computers to be transmitted over high speed data links and processed by a large scientific computer during the course of the experiment.

2. Outline of the system

FOCUS provides multiple access I/O facilities on a CDC 3100 computer with limited disk storage. A CDC 3100 computer has channel-to-channel connections to the CDC 6600 - CDC 6500 computers and has direct data-link connections to each experimental hall at CERN. The configuration is shown in **Fig. 1.**

To its users FOCUS is a file handling system allowing several users simultaneously to create, accumulate, and manipulate information files. Its potential use covers a much wider range of application than the on-line physics experiments whose needs initiated the system design, and it has been implemented as a general purpose multi-access remote job entry facility for CERN's central computer system. Facilities are provided for:

1. On-line storage of active users' program and data files.
2. On-line creation, manipulation and modification of users' program and data files from teletype and storage tube display consoles.
3. Transmission of data between the 3100 and the remote data acquisition computers via data-link connections.
4. Priority access to the central computer system for job input files created through FOCUS.
5. The transmission of job output files from the central computer to remote users.

In addition, all the facilities of the central computer operating system are available.

Users access the system via a variety of consoles which at present comprise 17 KSR33 teletypes, one KSR37 teletype and three Tektronix T4002 storage tube displays and which are connected to FOCUS by means of a Hewlett Packard 2116B computer. This computer acts as a 'front end' message processor for the consoles and other slow-to-medium speed I/O equipment, relieving the 3100 of the detailed device control while at the same time enabling the system to easily handle a variety of device speeds ranging from 10 characters per second KSR33's through 1000 characters per second T4002 displays to 500 kilobaud serial connections to remote computers. In addition to the software, all hardware for connecting the HP2116 both to the 3100 and to its own peripherals has been developed at CERN. (Bruins, Olofsson and Slettenhaar, 1970).

There are two additional classes of computer-computer connections in the system. The central computers are linked to the 3100 by standard CDC channel couplers. Connection to remote computers for data acquisition purposes is made by a network of fast data-links designed and constructed at CERN (Joosten, 1968, 1969). This network is controlled at the 3100 by a Data-Link Synchroniser (DLS). The DLS can control four high speed links to remote sites (up to 3 km distant) and each link is terminated by a Data-Link Terminal (DLT) which in turn can be connected via a standard interface to four remote computers located at distances up to several hundred metres from the DLT. Thus one DLS can connect as many as 16 remote computers to the 3100. The Data-Link System transmits data and status as 12-bit parallel bytes and has a design speed of up to one million bytes per second. Only one remote computer can transmit data at any one time.

The hardware thus exists to enable the 3100 to act as a node machine of a network connecting many remote computers to the central computing facility and the FOCUS system is the software designed to implement this capability.

*Present address: Data Processing Associates, 1 High Street, Guildford.
†Present address: The Computer Unit, Bristol University.

3. Structure of the system

The system time-shares its activity between several users, their number being limited in the current configuration by the machine storage capacity. There are two classes of user within the system:

1. Physical users, each associated with a console.
2. Pseudo-users, each of which is designed to take care of a background system activity.

From the point of view of the executive, all users are treated identically. There are currently three pseudo-users. The first transfers files between tape and disk—a background system activity automatically initiated when a user logs in or logs out. The second receives output files from FOCUS jobs executed in the central computer. The third handles user initiated line printing, and the transfer of jobs from FOCUS to the central computers.

Data-link use is always associated with a console, and hence with a physical user.

The central processor is shared between:

1. The system executive.
2. User activities.
3. An idle program.

Fig. 2 gives a simplified block diagram of the system structure.

3.1. *The system executive*

The system executive consists of the following modules:

1. Processor scheduler.
2. Memory scheduler.
3. Interrupt processor.
4. Central I/O control.

User activities and the idle program operate with interrupts enabled, while the executive operates with the interrupt system disabled. Communication between user activities and the executive takes place in two ways:

1. A user activity can request executive action (e.g. an I/O

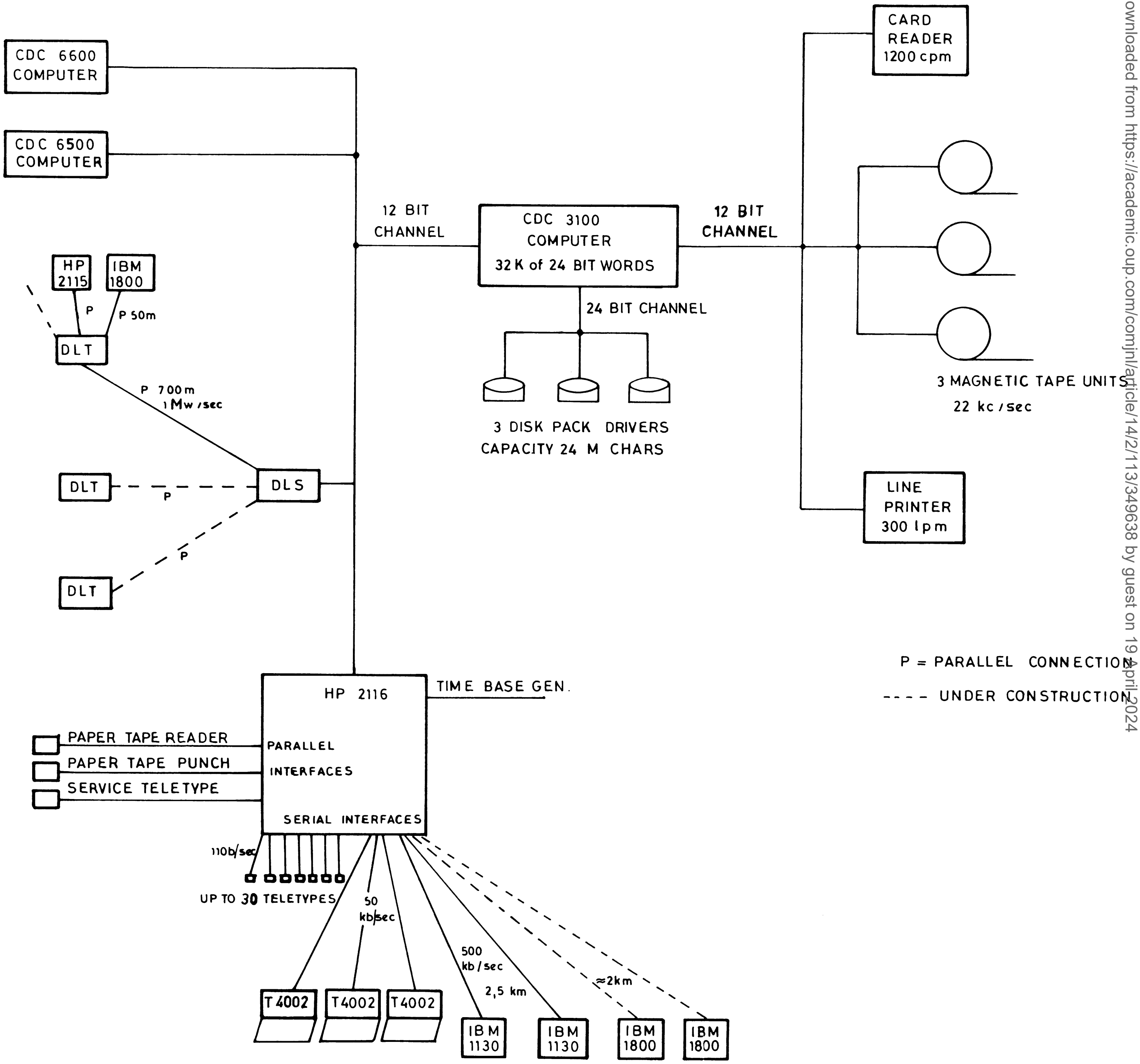

Fig. 1. Block diagram of hardware configuration.

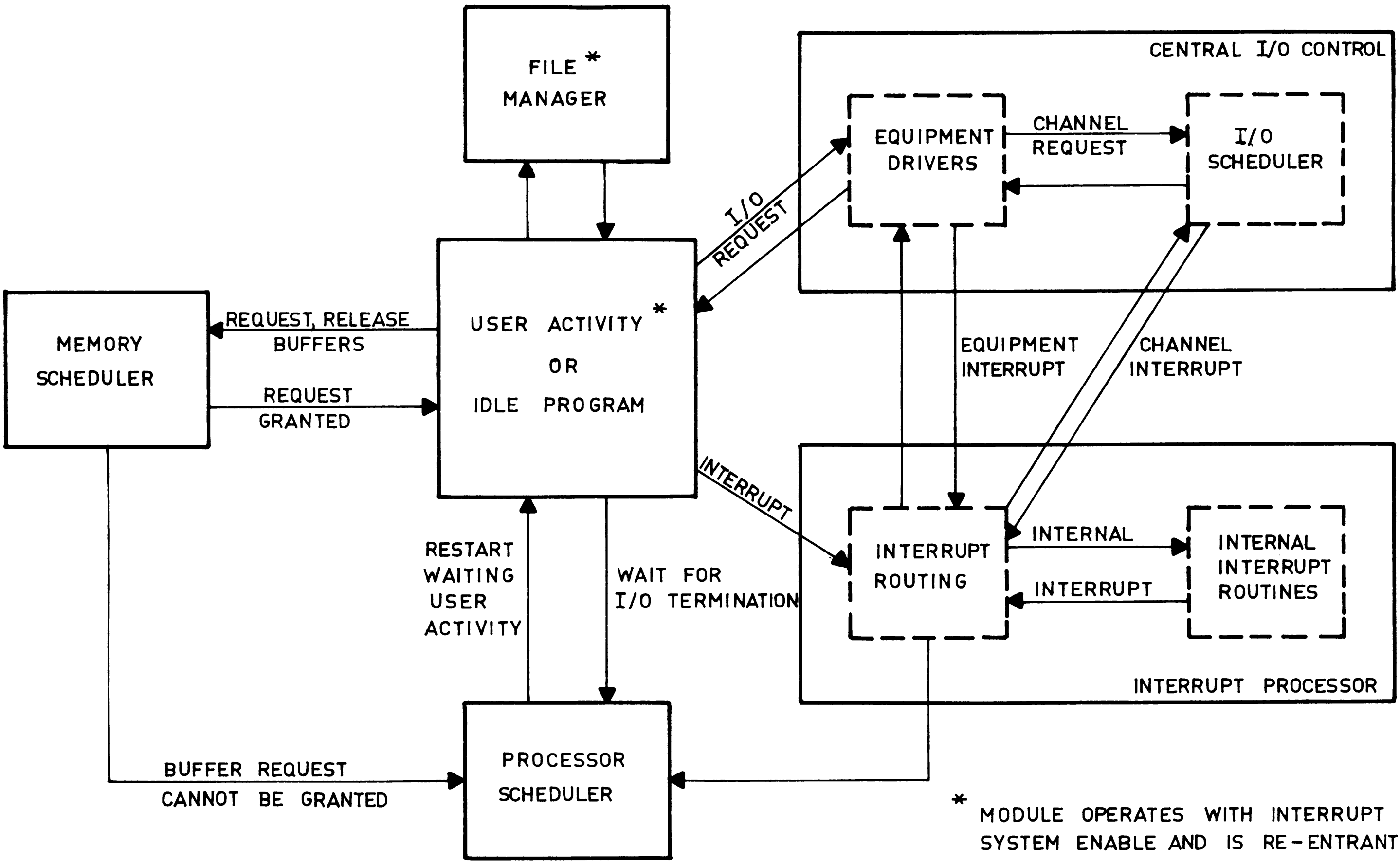

Fig. 2. Block diagram of system structure.

operation or the allocation of CM buffer space). If the request cannot be immediately satisfied, the user can be placed in a wait status and the processor switched to another user activity.

2. An interrupt can cause suspension of the current user activity and activation of an interrupt driven part of the executive such as an equipment driver or the central I/O controller. This interrupt can in turn signal completion of some action and cause a user activity to be rescheduled for the processor.

3.1.1. *Processor scheduler*

The processor is scheduled between user activities on the basis of priority and demand. Each user has an associated priority and a processor queue is maintained for each priority level. Whenever the executive is prepared to relinquish the processor it branches to the processor scheduler which scans each priority queue in descending order, and if it finds a user activity waiting, it restarts it. If no user activity is waiting the idle program (a one instruction loop) is given control until an interrupt occurs. No attempt is made to time-slice the processor between user activities.

3.1.2. *Memory scheduler*

Memory is divided into two parts, a permanently allocated area and a dynamic buffer area. In the resident part are the system tables, the interrupt handling routines, the file manager, essential equipment drivers and certain other frequently used subroutines. The buffer area is used for routines read from the disk and as work space. Buffers are assigned in units of 64 words (equal to one sector on disk), and their allocation is controlled by a fast non-interruptable routine.

If space requested is not available, the user is generally put into wait buffer state. Whenever space is released or moved all such users are reactivated and try again to obtain the space they need. To try to avoid all users going into wait buffer status, in certain cases a current user activity may be terminated.

3.1.3. *The interrupt processor*

Whenever an interrupt occurs, the interrupt processor:

1. Saves the current register contents of the interrupted routines.
2. Analyses the source of the interrupt and transfers control to the appropriate interrupt servicing routine.

All routines activated by interrupts return control to the interrupt processor on completion. The interrupt processor then transfers control to the processor scheduler which activates the highest priority user waiting for the processor.

3.1.4. *Central input/output control*

I/O control is shared between a centralised I/O scheduler and a driver for each device. I/O operations are initiated by a user request to an equipment driver via a trapped instruction, and the equipment drivers in turn formulate requests to the central I/O scheduler. This then supervises all physical input/output and schedules channel activity. Once initiated, all I/O operations are controlled by channel interrupts (signalling the end of I/O on a channel) and external interrupts (signalling completion of activity by an I/O device).

All channel interrupts are routed directly to the I/O scheduler. External interrupts, on the other hand, are routed directly to the drivers which check equipment status after each operation and issue equipment control functions. The drivers also notify the users after each requested I/O operation is complete and replace the user activity in the queue for the processor, if necessary.

3.2. *User facilities*

The facilities of FOCUS are accessible to users by a series of commands which are entered at a teletype console. A partial list of commands is given in Appendix A. The system maintains a list of authorised users who are identified by name, number and group code in a manner compatible with conventions for batch users of the central computer system. To obtain access to the system the user enters the LOGIN command on his console after which all facilities of the system except those concerned with system control and accounting, are available to him. Commands are provided for system access, file management, file manipulation, file editing, for job submission to the central computer and for viewing of graphical output generated in the central computers.

Each command corresponds to a service program which is stored permanently on disk. Commands permitting concurrent execution by several users are coded in a re-entrant form. The commands are typed in a free format and consist of an 'action verb' followed by one or more parameters which are usually file names, e.g. RENAME JOHN MARY. The command 'renames' the file named JOHN by the new name MARY. Users are informed of illegal parameters and depending on the command, either the correct parameter can be retyped or the command re-initiated. Similarly, the user is prompted with short messages for missing parameters. The action of each command is almost self-evident from the command names which follow closely those originating in the Compatible Time-Sharing System (Corbato *et al.*, 1963).

3.2.1. *Communication with the* 6600 *Series computers*

The facility for transmitting files between the central computer and the 3100 consists of two parts:

1. A user command, SEND, for initiating the transfer of a job file consisting of a number of FOCUS files to the central computer.
2. A FOCUS pseudo-user for carrying out the transfer.
3. A FOCUS pseudo-user for receiving the output files in FOCUS when they become available and transferring them to the appropriate user.

To transmit a job the user specifies, by using the SEND command, the sequence of FOCUS files making up his central computer job file, and which output files he wishes returned to FOCUS. After initiating transfer the user is free to carry on any FOCUS activity, including sending other jobs to the central computers.

On job completion the central computer interrupts FOCUS, causing it to activate the FOCUS pseudo-user to input the results files to FOCUS and notify the user of their availability.

The command STATUS enables a user to check on the progress of any job he has sent to the central computer for execution. Facilities for interrogating the status of any job, whether remotely submitted or not, also exist.

The central computer side of the interface is a program called REMOTE which runs under the CERN version of the CDC 6000 series operating system. REMOTE consists of two parts; a small program (1·5K words) which resides permanently in the central memory, together with a larger program which operates in one of the peripheral processors (PP's) of the machine. When not active, the PP program resides on a system disk and is called into activity at approximately 10 second intervals to check for requests from the 3100 for job transfers, or for the existence of files waiting to be transferred back to FOCUS.

The I/O interface is designed so that files can be input and output simultaneously, with the block transfers in either direction interleaved on the I/O channel (Gerard, Marcer, and Palandri, 1969).

3.2.2. *Communication with remote computers*

The interface between FOCUS and the remote computers connected by the CERN data-link network is implemented very simply and consists of two parts:

1. User commands DATIN and DATOUT for respectively inputting and outputting a file between FOCUS and the remote computers.
2. A data-link driver for controlling the I/O operations for both commands.

From the point of view of FOCUS the remote computers function purely as another I/O device. DATIN allows a user at a console near the remote computer to accumulate a file of data from his machine to FOCUS, prior to sending it to the central computer for processing. The complementary command enables an output file to be sent back to the remote computer.

The main complication in implementing the system lay in the fact that most remote computers to be connected to FOCUS are data acquisition computers operating in a stringent real time environment. Thus the FOCUS data-link driver is designed to operate in conjunction with the remote computer software in such a way that the link to FOCUS operates only when more important activities in the remote computer permit, and that the FOCUS transfers can be interrupted and if necessary terminated, at any time a higher priority activity at the remote site requires it.

3.2.3. *File management*

The permanent file system was designed with the following points in mind:

1. Disk capacity was too small to store any files permanently. Thus the system must be tape based with only active files of users, who are actually logged in on the disk.
2. Disk accesses should be kept to a minimum.
3. Movement of files between tape and disk should be controlled automatically by the system, and proceed in parallel with manipulation of other files by the user.
4. Each user should be able to store within the system data in excess of the disk capacity allocated to him.
5. Some files would be large, frequently modified data files.
6. The design should allow for an increase in disk capacity and the number of simultaneous users to be made at a later stage.

These objectives were met by classifying files into three types:

(*a*) *Temporary*—these files are created, on the disk, during the run, and are destroyed when the user logs out (unless he has exercised his option to save them). They are thus work files.

(*b*) *Current*—these files are kept on tape, and automatically transferred to disk when the user logs in, and written back to tape when he logs out. Their total size is limited by the disk space allocated to the user. Current files should be high activity files.

(*c*) *Archive*—these files are stored on tape, and a particular file is only transferred to disk if the user explicitly references it. When the user informs the system he has finished with it for this run or he logs out, the file is written to tape only if its contents have been changed. In any case the disk space used by the file is released. A user may have archive files whose total size exceed his disk space allocation.

All files are created as temporary and commands exist to allow changing of mode. Access to files is either read only or read/write. No sharing of files between users is possible, however a user may transfer a file to another user who is logged in.

Space on the disk is allocated by the system in units of one track. Each track can hold 4,096 characters and is divided into 16 sectors. Information can be transferred in multiples of a sector, and since the hardware checksums each sector, it is not necessary to read information in the same units as it is written.

Files consist of an integral number of tracks, the last generally being only partially full. The tracks carry no linking information, all the space being used for data. Instead a list of the tracks constituting each file is maintained on a separate area on the disk. To manipulate a file it is necessary for it to have been opened. This is normally done automatically by the system. For such files a part of the track directory is maintained in core store to avoid disk accesses. To delete a large file it is only necessary to read one sector per 64 tracks instead of one sector per track if the linkage was carried in the tracks themselves. When a user has finished with a file he may close it, otherwise the system will, if necessary. To conserve space in memory, each user is limited to seven files open at any one time.

For each authorised user an entry is maintained in the Master File Directory (MFD). Among the information is the address of his User File Directory, which has one entry for each of his files, giving mode, access, which tape (if any) it is stored on etc. When a user is logged out this is the only information remaining on disk, all of his current and archive files being preserved on tape. These directories are constrained to be all on the same disk. When the system is run down a copy of the information on this disk is written to tape and a checksum computed. Before re-starting the system, this checksum is verified to ensure the integrity of the directories. If a permanent discrepancy is found, the disk is rewritten from the last dump tape. During a run the last dump tape is periodically updated with directories of users who have completed their session at a console, so if a system fault occurs it is possible to restart losing only directory changes of users actually logged in at the time of the failure. Each user has two current tapes which are used in turn to provide a possibility of recovery in the event of a system fault.

The transferring of files between tape and disk is handled by a pseudo-user known as the tape manager. This program is activated by a table of requests, each one of which is to transfer either all of a user's current files or an individual archive file. The 'tape manager' processes up to three requests in parallel. A user is prevented from manipulating a file which is in the process of being transferred by means of an 'inhibit' flag. A checksum is kept for each file on tape, and any errors in tape reading result in a message to the operator and one to the user warning him that certain files may be corrupt. A command exists whereby the user can instruct the system to re-read faulty files from either of his current tapes in order to recover them.

3.2.4. *Graphic display facilities*

A simple non interactive graphics facility is available in the system. Graphic files generated by a generalised display package in the central computers may be returned to FOCUS where a user command, TV, is available to interpret and display them on the T4002 storage tube displays. TV provides facilities for selecting individual frames from the file for display, expanding parts of a frame, displaying two frames side by side and superimposing different frames on the screen.

4. Experience with the system

The system has been operational since March 1969, initially for a few hours per day with a smaller number of consoles. The system is now operational for one shift and can accommodate up to 21 simultaneous users. Production runs are made on-line to the CDC 6600, approximately 800 jobs being submitted via FOCUS per week. A version of the system is also operational at Brookhaven National Library in the United States.

Two remote computers, an IBM 1800 and a Hewlett Packard 2115, are connected via the data-link network, and one experiment has run on-line to the central computers via FOCUS.

The current version of the system occupies permanently 14K of core storage leaving 18K to be dynamically allocated for transient routines and work space. Routines occupying in tota about 50K words of code are permanently kept on disk.

4.1. *Operational experience and system utilisation*

Operational experience with the system has made it possible to accumulate data on system performance. In particular data exists concerning:

1. System response to the user.
2. User response to the system.
3. Central computer response to jobs submitted from FOCUS.
4. System utilisation of hardware resources.

The nature of the user facilities available directly in the system is such that in most cases the system response to the user is apparently instantaneous even at the heaviest current loading. Delays can occur in operations involving much disk activity, such as copying and joining long files or the sparse editing of long files, but all other long operations, such as printing, transfer of files to the central computer or between disk and tape are buffered from the user. The delay in making available user current files at login (due to the necessity of reading them from tape) does not appear to greatly inconvenience the user, in view of the fact that other system facilities are available when waiting.

User response to the system has been good. Regular use is made of the system by more than 100 programmers (out of CERN programming population of around 450). All consoles are in regular use, and a booking system which limits a user to a maximum of one hour at a time on a T4002 had to be introduced.

Job turnaround in the central computer varies considerably according to the type of job, however for short jobs (up to one minute CP time) not requiring magnetic tapes it is rarely longer than 10 minutes.

An analysis of the accounting information for system utilisation of hardware reveals that the hardware is still only moderately loaded at present except for the central memory and with some re-organisation of memory use further expansion is still possible.

As an example of system utilisation, the accounting produced during a recent typical system run is as follows:

Number of users logged in during run	26
Number of jobs sent to 6600	151
Total run time	4 hours 20 min. 41 sec.
Total CP idle time	3 hours 24 min. 21 sec.
CP use for direct user activity	37 min. 18 sec.
CP used for executive and interrupt handling	19 min. 2 sec.

Channel 0, used for tapes, card reader and printer was active for 38 minutes 51 seconds, transferring 7931K words.

Channel 1, used for 6600, HP2116B, and remote computer connection was active for 1 minute 56 seconds, transferring 5883K words.

Channel 2 used for disks was active for 29 minutes 52 seconds, transferring 41579K words.

101239 disk requests were handled in a total of 2 hours 14 minutes 18 seconds of activity by the 3 disk drives.

4.2. *Operational problems and reliability*

Operational experience of system reliability has been good. After initial teething problems system failure due to software error has dropped to a low level apart from inevitable errors occasionally introduced during system modification. Considerable initial problems were also experienced with the hardware and in particular with the very vital system components, tape units and disks. Experience has shown that when the disks are operational they are extremely reliable and no known cases of the disk drives corrupting data have been recorded during testing and operational experience, despite some very thorough checks built into the system disk driver. On the other hand the drives occasionally developed faults, mainly mechanical, causing them to be down for several days at a time.

More serious were the early problems experienced with tape units. These form a vital part of the system mass storage as all user files are kept on tape between production runs. The tape units initially attached to the machine were inexpensive units of a semi-obsolescent design and from time to time unit compatibility problems developed. Thus a unit would appear to be working perfectly, but the next day it was found that all tapes written on that unit could not be read by the other units and occasionally not even by the same unit. Since the units were replaced with units of more modern design tape problems have dropped to a negligible level and in general the hardware now gives very little trouble.

4.3. *Experience with system implementation*

Initial design of the system began in March 1967, with implementation beginning in June 1967. Over the two years that elapsed before beginning production, a total of eight programmers contributed parts to the system, although no more than the equivalent of four full-time programmers have worked on the system at any one time. It is estimated that the work required to bring the system to production status has totalled eight man-years of effort and continuing development of the system has occupied the equivalent of two-and-a-half programmers over the last two years.

4.4. *Future system development*

Current plans in 1971 call for two shift running with occasional periods of 24 hours per day availability on demand from experiments serviced by remote computers connected to FOCUS. The number of simultaneous users will be increased to 30 and at least one fast alphanumeric display will be added to the current teletypes and T4002 displays. At the same time two remote batch entry stations consisting of a 4K IBM 1130 computer running a medium speed card reader and line printer will be connected to the central complex via FOCUS and two other high speed (500 kilobaud) serial data links will be added.

Acknowledgements

The authors wish to thank Dr. P. Zanella for his very valuable contribution defining the original design objectives of the system, and Mr. W. G. Moorhead for continuing help at all stages during the development work. Much of the design and implementation of more recent major improvements to the system has been the work of Dr. J. Gerard. We would also like to acknowledge the considerable and continued assistance of Mr. M. Baechler, Mr. G. Genova, Mr. L. Tausch and Dr. A. Yule in implementing various parts of the system.

Appendix A

1. FOCUS user commands

LOGIN	Log named user into system to begin a run.
LOGOUT	Terminate current run for user.
OPEN	Open named files.
CLOSE	Close named files.
LIST	List user's file directory on his console.
ACCESS	Change access mode of named file.
MODE	Change file type of named file.
RENAME	Rename named file.
RECOUP	Re-read named files from user's tapes.
COPY	Make copy of named file.
JOIN	Join named files.
DELETE	Delete named file from user's directory.
READ	Read named file from the system card reader.
PRINT	Print name file on system line printer.
INPUT	Input from user console to named file.
OUTPUT	Output named file to user console.
EDIT	Do context editing on named file.
SEND	Send named files to the CDC 6000 series computer for execution.
STATUS	Give current status of user job executing in the CDC 6000 series computer.
DATIN	Input named data file from specified remote computer.
DATOUT	Output named data file to specified remote computer.
SYSTEM	List commands available in system.
TALK	Send a message from user console to another console.
TV	Display of graphic display file on a T4002 console.

2. FOCUS system control commands

ADUSER	Add named user to list of authorised system users.
MFD	List authorised system users, with facilities for modifying disk space limits, console time limits etc.
LABEL	Write FOCUS label of magnetic tapes.
FINISH	Terminate current system run. Save all file directories on tape.

References

Barron, D. W., Fraser, A. G., Hartley, D. F., Landy, B., and Needham, R. M. (1967). File Handling at Cambridge University, *AFIPS Conference Proc.*, Vol. 30, p. 163.

Bruins, T., Olofsson, K. S., and Slettenhaar, H. J. (1970). Some hardware aspects of computer communication facilities at the CERN Computer Centre—CERN/DD/CO/70/17.

CDC (1968). 3100 Computer System Reference Manual (CDC Publication 60108400).

Corbato, F. J., Dagget, M. M., Daley, R. C., Creasy, R. J., Hellweg, J. D., Orenstein, R. H., and Korn, L. H. (1963). The Compatible Time Sharing System. A Programmer's Guide, MIT Press.

Gerard, V., Marcer, P. J., and Palandri, E. M. (1969). An interface for full duplex data transmission of files between a large multiprogramming operating system and an on-line file manipulation system.—CERN/DD/DH/69/15.

Joosten, J. (1968). High speed data transmission system for CERN data-links, CERN/DD/DA/68/5.

Joosten, J. (1969). The FOCUS data-links, a general hardware description, CERN/DD/DH/69/16.

Lord, D., and Macleod, G. R. (1969). The use of computers in high energy physics experiments, *J. Sci. Instrum.* (*J. Physics E*), Series 2, Vol. 1, pp. 1-9.