AT/25.1/007
24 FEB 84
COMPUTERS AND PROCESSING - FILE NOTE

COMPUTING REQUIREMENTS FOR AT SOFTWARE DEVELOPMENT

R.WAND
M.KESTEVEN
P.RAYNER

1.0 INTRODUCTION

As set out in the computer workshop report (AT/10.5/002), the AT computing requirements at Culgoora were to be met by two VAX 11/750's; one for synchronous tasks and one for asynchronous tasks. So as to provide a measure of protection against failure of a single cpu , these two computers were linked by a CI (computer interconnect) bus. The proposed purchase date for the first of the 11/750's was early in 1985, with purchase of the CI and the other 11/750 delayed until later. This schedule meant that all AT software development work in 1984 would have to take place on the existing VAX 11/750 (RP750) at Epping.

In practice the RP750 has already become saturated and cannot adequately support AT development work. We are therefore proposing that another VAX 11/750 (to be referred to as AT750) be purchased immediately and used primarily for this AT software development effort. AT750 would initially be installed at Radiophysics and be networked with RP750, eventually to be moved to Culgoora early in 1987 where it would become the synchronous computer.

2.0 THE EXISTING SITUATION

The RP750 on which AT development work presently takes place is overloaded during the day. It frequently takes of order one minute to log onto the machine. It can take the order of 15 minutes simply to read an image from tape and to display it on the DEANZA. During the day the ratio of elapsed-time /CPU-time can reach 8:1; the same task in the evening shows 1.3:1 . This means that a task worth waiting for in the evening - say a five minute wait - may not be worth waiting for during the day.

Note that the times given are average times and are by no means the worst experienced. Moreover AIPS plays little role in this dismal state of affairs; most AIPS heavy tasks (map, clean, mem, etc) run on AIPS batch - no more than 2 concurrent tasks, and these at priority 3. This means that they obtain essentially no CPU time during the day.

Appendix A presents results of some diagnostic tests which indicate that the computing power of the existing VAX 11/750 is in fact being utilised quite efficiently and there is little scope for improving the situation without adding more computing capacity.

It is important to recognize that the real testing of the image processing system can only be done by the astronomers - the AT group can implement and maintain a system, but the assessment of its acceptability must come from the astronomers. This means that better facilities must be provided. At the present time, the RP750 offers an unpleasant environment for the astronomer because:

a. Only in the evening does the response time approximate that of a computer.

b. During the day it is often the case that even such simple tasks as the displaying of an image can take an irritatingly long time.

c. Tasks such as CLEAN now take an average of 1 day.

d. During the day it is often difficult for the general user to gain access to a terminal. Access to a graphics terminal is now essentially impossible. (even though there are ostensibly 9 or so such terminals in the building).

e. Because disk space is at a premium, data files have a short half-life, and are in danger of evaporating before the analysis is complete.

It is true that the RP750 use is reduced in the evening and on weekends - indeed, it is during these hours that the bulk of the AIPS development and testing occurs; however the AT-computing group cannot be expected to maintain these hours as a long-term proposition.

3.0 AT SOFTWARE DEVELOPMENT REQUIREMENTS

The AT computing load has up until now been primarily restricted to evaluation of and enhancements to AIPS, and also for configuration studies. This load will certainly increase in the near future. The immediately predictable tasks:

-Simulation studies.

a. In conjunction with the configuration group, to assess the merit and/or drawbacks of various configurations.

b. On behalf of the processing group to investigate and assess different calibration/self-cal/processing schemes.

Both these projects involve the generating, cleaning and comparing of modest sized (eg. 256x256) maps. On the

present machine large spectral data cubes could not be handled in our lifetime.

-Image processing

a. Upgrading, repairing and testing of AIPS, GIPSY (and possibly others)

b. active use of AIPS/GIPSY by astronomers - VLA, FLEURS, MOST and TEST data;

c. investigating alternative algorithms and display schemes. (arguably, part of a.)

-Programme development for the asynchronous computer:

Data editing, calibrating and display.

-Programme development for the synchronous computer.

4.0 RECOMMENDATION

That we purchase immediately the VAX 11/750 ultimately intended for the synchronous computer at Culgoora and install it at Epping. Thus we propose advancing the purchase date of a machine which would need to be acquired in the relatively near future in any event. This machine is already adequate for its intended purpose - ie. there is little point in delay in the hope of better hardware becoming available. (This same argument, for example, would not be true of the VAX 11/780 data reduction computer intended for Epping). In any case DEC presently do not appear to be developing any new computer likely to supercede the 11/750 at comparable cost.

5.0 PROPOSED COMPUTER SYSTEM

We propose that the new equipment shown in Table 1 be purchased to comprise the AT750 computer system. The manner in which AT750 will be integrated with the existing Epping VAX computers is shown in Figure 1.

The current proposal (see Figure 1) is for an Ethernet link between the AT750, RP750, and RP730, with a possible future connection to the AAO VAX. Ethernet operates at 10Mbits/sec and is supported by DECnet. It is considered important that some experience be gained with Ethernet as it is anticipated that this system will also be used between the synchronous and asynchronous computers at Culgoora. Note that the Table 1 does not include the cost of connecting RP750 and RP730 to Ethernet; this would amount to an extra $5K for each node (item 10 in the Table 1). The cost of the Ethernet cable given in the table is only for the RP VAX

computers and does not include connection of the AAO node.

In addition to the Ethernet link, the RP750 and AT750 will be linked by providing a new shared 475 Mbyte disk drive using a System Industries 9920 controller and associated SIMACS (System Industries Multiple Access Control System).The 9920 will support up to 4 disk drives in a radial configuration so it would be possible at some later date to transfer one of the existing RP750 Fujitsu disks to the 9920 to provide additional shared disk capacity.

SIMACS is under consideration as an alternative to the DEC CI bus and HSC50 as a less expensive means of interconnecting the synchronous and asynchronous computers at Culgoora. Its acquisition for the AT750-RP750 interconnect will permit a careful evaluation of its capabilities. Appendix B gives a more detailed description of the capabilities and limitations of the SIMACS system.

Three RK07 28 Mbyte disk drives are presently available at RP and it is proposed that one of these be used as the system disk for the AT750. The cost of a controller for this disk is included as item 5 in Table 1. In addition, one of the existing RK07 drives on the RP750 will be dual-ported to the AT750 so it can be used to backup the system disk on either the AT750 or the RP750 (item 6 in Table 1).

With the proposed configuration the RP750 should become a more viable interactive machine if the heavy AT work is transferred to AT750; this expectation is reasonable, as the AT associated computing currently accounts for a large fraction of the RP750 load.The AT750 is configured with only a few terminals, and neither tapes, nor graphics devices. AT750 needs for these peripherals would be satisfied via the 2-way links to RP750.The magnetic tape units and disks which will eventually be attached to an interconnect and shared by the synchronous and asynchronous computers are not being purchased at this time.

As development proceeds AT750 will be used for testing of special purpose AT hardware (such as the antenna link controller and correlator interfaces) which means that it may have to be withdrawn from general use from time-to-time.

Purchase of an array processor (FPS5205 or equivalent) is also currently under consideration. This unit may well be best attached to RP750, in order to expedite AIPS.

TABLE 1: New equipment purchases for AT750 computer.

ITEM	COST $K
1.VAX 11/750,FPA floating point accelerator, 2 Mbytes memory, and VMS copy licence;	72
2.Additional 2 Mbytes memory;	7
3.Software copy licences for Fortran,DECnet;	11
4.Emulex CS11/F1 terminal controller(16 lines);	6
5.Disk controller for RK07 (system disk);	7
6.Dual porting of RK07 disk (system backup disk);	5
7.System Industries 9920 Disk Storage System (includes Fujitsu 475 MByte disk ,9920 controller,CMI adapter);	23
8.CMI adapter (to allow RP750 to access new disk);	5
9.SIMACS (software and firmware to allow sharing of new disk by AT750 and RP750);	15
10.DEUNA Ethernet connection and transceiver;	5
11.Ethernet cable and terminators;	0.5
12.LA100 hardcopy terminal.	3
TOTAL	159.5

APPENDIX A

USAGE OF THE PRESENT EPPING VAX 11/750 -(PTR)

These notes look into the question of current usage of the VAX 11/750 and whether it can be improved.

Usage of this VAX is extremely variable, ranging from simple edit, compile and run jobs to intensive compute bound batch jobs that consume hours of CPU time. Under such circumstances it is difficult to collect and analyse usage data which will lead to clear cut lines of action.

In this study data were collected from runs of the MONITOR utility. The runs were done hourly over a 24 hour period on several days. Data generally represent 1 minute averages. Table A1 gives a summary of data collected on the 9th February and Table A2 on the 15th. The contents of the columns are:

1. the time of the measurement.

2. the number of processes in common event flag waits. This generally means waiting on terminal input.

3. the number of processes hibernating. These are mostly system utility processes such as ACP's and print symbionts.

4. the number of computable processes. These are processes that are able to run immediately i.e. they are not waiting on I/O or other events. If there are more than a few processes in this state there will be no idle time.

5. the percentage of time that the processor spent executing interrupt handling code.

6. the time in kernel mode. This is time spent by the operating system handling such jobs as scheduling, paging and swapping. It is unproductive time as far as a user is concerned.

7. the time spent in executive mode. This is mainly taken by RMS (record management service) handling file accessing, reading and writing.

8. the time spent in supervisor mode. This is the time spent executing DCL.

9. the time spent in user mode. This is the time spent executing user code.

10. the time spent in compatibility mode. During this time the processor in executing RSX images, i.e. emulating a PDP 11.

11. the time spent executing the NULL process. This process is only run when the system has absolutely nothing better to do.

Note that the processes in LEF, HIB and COM account for all user processes. No user process was swapped out. It would appear that our system does not need to swap, we have sufficient physical memory to keep all working sets in memory. So one source of performance degradation, namely swapping, does not afflict our VAX.

Looking at Table A1, we can see that, except for the period around 9 am, there was essentially no idle time. So the VAX was doing "something" all the time. The question is - was it doing something useful? From the table we can see that, generally, the time spent in user mode was above 60%. There was a low period around 11:00, the reason for which is not clear. The average time spent in user mode over the 24 hours was ~80%. This high percentage was helped by the 92% times between midnight and dawn when the system was running big batch jobs.

Looking at Table A2, we see a similar pattern of daytime utilization. There is idle time during the first part of the morning then essentially none throughout the rest of the working day. The last batch job completed around 22:30. From then on the system was idle till ~08:30. Note that there was a meeting of the "Parkes User Group" held on 15 February between 13:30 and 16:00. Amongst the attendees at that meeting were most of the heavy batch users. So, perhaps, with less time to produce that nights batch command files the batch queue emptied before the next morning.

From this, admittedly scanty, statistical base we can draw some conclusions:

1. there is very little that can be done to improve the daytime utilization. The VAX is "doing its best" to support the daytime load.

2. there would be better utilization over a 24 hour time scale if users moved their bigger interactive jobs onto the batch queue. This would soak up any idle time during the night.

3. the gain from (2) would be small and, perhaps, just result in slightly better interactive response during the day.

4. the only way to make a significant increase in computing power available to the current user population would be to buy a second processor. This processor could be used for interactive work but might be better employed as a batch processor for daytime jobs. In any event, if the second processor were available, we will need to investigate ways of sharing the computing load between the two processors.

TABLE A1: MONITOR data collected on 9-February

	STATES			MODES						
TIME	LEF	HIB	COM	INT	KERN	EXEC	SUP	USER	COMP	IDLE
10:14	10	11	3	6	3	1	0	89	0	0
11:16	8	10	7	16	38	8	5	24	5	0
12:19	13	13	4	10	16	8	2	61	0	0
13:21	5	11	4	9	13	5	2	69	0	0
14:23	7	13	2	9	7	1	0	81	0	0
15:26	6	12	2	14	11	2	0	71	0	0
16:28	7	12	3	7	4	1	0	69	0	0
17:31	12	9	5	7	10	3	0	77	0	0
18:33	6	10	5	5	2	0	0	91	0	0
19:36	5	11	5	5	1	0	0	91	0	0
20:38	5	11	5	5	2	2	0	88	0	0
21:40	6	11	5	8	8	4	1	76	0	0
22:43	7	11	6	7	8	4	0	79	0	0
23:45	4	11	5	4	1	0	0	92	0	0
00:47	4	11	5	5	1	0	0	92	0	0
01:50	4	11	5	5	1	0	0	92	0	0
02:52	4	11	3	5	1	0	0	92	0	0
03:54	4	11	3	5	1	0	0	92	0	0
04:57	4	11	3	5	1	0	0	92	0	0
05:59	4	11	3	5	1	1	0	91	0	0
07:14	5	11	2	6	3	2	0	87	0	0
08:16	6	11	2	7	4	1	0	87	0	0
09:19	11	10	1	8	13	9	4	7	0	55

TABLE A2: MONITOR data collected on 15-February

	STATES			MODES						
TIME	LEF	HIB	COM	INT	KERN	EXEC	SUP	USER	COMP	IDLE
09:36	16	11	1	12	31	3	4	13	0	36
10:36	13	11	2	6	16	7	1	34	0	32
11:36	16	10	6	10	24	15	0	49	0	0
12:36	10	12	5	5	8	11	0	73	0	0
13:36	9	15	6	7	20	4	0	67	0	0
14:36	9	13	7	4	7	3	0	83	0	0
15:36	7	13	7	14	33	3	2	44	0	0
16:36	15	11	6	5	13	11	0	68	0	0
17:36	14	12	5	4	4	1	0	88	0	0
18:36	10	12	5	3	3	2	0	90	0	0
19:36	8	12	5	3	1	0	0	93	0	0
20:36	8	12	5	2	2	0	0	94	0	0
21:36	7	11	3	3	2	1	0	91	0	0
22:37	7	11	3	2	2	1	0	93	0	0
23:37	7	10	1	1	1	1	0	3	0	92
00:37	8	10	1	1	1	0	0	3	0	92
01:37	7	10	1	0	1	0	0	3	0	92
02:37	8	10	1	2	1	0	0	3	0	91
03:37	7	10	1	1	1	0	0	3	0	92
04:37	8	10	1	1	1	0	0	3	0	92
05:37	8	10	1	2	1	0	0	3	0	91
06:37	8	10	1	0	1	0	0	3	0	92
07:37	7	10	1	0	1	0	0	4	0	92
08:37	9	10	1	3	8	2	0	11	0	73

APPENDIX B

DISK SHARING WITH SIMACS -(PTR)

B.1 INTRODUCTION

These notes are intended to give a brief discussion of the System Industries product named SIMACS (System Industries Multi-Access Control System). In broadest outline SIMACS allows read/write sharing of a disk volume by multiple VAX processors and it may provide an economical means of establishing an interconnect between the synchronous and asynchronous computers at Culgoora. Installing it on the existing RP750 and the proposed AT750 would provide a means of carefully evaluating its performance.

B.2 STANDARD DISK I/O

To understand how SIMACS works we need to look at how VMS normally does I/O to a disk. There are three types of I/O that can be done to a device.

1. Physical - data is read from and written to the actual physically addressable units accepted by the device. For a disk this is a sector. The part of the operating system responsible for handling the data transfer is called a device driver or just driver.

2. Logical - data is transferred in blocks relative to the "start" of the disk. The mapping of block numbers to physical cylinders, tracks and sectors is handled by the driver. The mapping will be made in such a way as to make the transfer of contiguous logical blocks as efficient as possible.

3. Virtual - data is transferred in blocks but relative to the "start" of an open file. The mapping between virtual and logical blocks is made by taking groups of blocks from a pool of available free blocks.

User programs will perform virtual I/O almost exclusively when doing disk transfers. Physical and logical I/O have no knowledge of files, whereas virtual does. The task of managing

B.4.1 Disk Sharing

In a SIMACS disk system the disk drives are single ported and the controller is multi ported. The controller is then used as a communication link between the processors. The information passed over this link is the ownership of each disk volume connected to the controller. Ownership takes the form of a semaphore which is used to lock a volume so that I/O initiated to the volume through the ACP prevents access from an ACP on another port. Before the semaphore is released by a port the releasing system updates the copies of the control files on the disk.

If a processor wishes to start transfers to or from a given file it first requests ownership of the disk. When this is granted the ACP reads the disk copy of the control files. It is then free to modify them however it wishes. The ACP can check if the disk control files have been modified since the last time that it accessed the volume. If there has been not modifications the read is skipped.

The owning ACP will not necessarily relinquish control as soon as requested, but the delay can be selected dynamically. An ACP can be instructed to relinquish the volume after a given number of IRPs (I/O Request Packets) have been processed or after a given number of clock ticks. The time that a requesting processor will wait before re-requesting ownership of a volume can also be selected.

B.4.2 File Locking

SIMACS also provides a mechanism for locking individual files against multiple writes and mixed reads and writes. This locking can be enabled or disabled when a volume is mounted. This locking prevents two processors from corrupting a file by attempting simultaneous writes and a reader getting invalid data when another processor is updating the file.

This is a locking mechanism between attached processors. The normal locks provided by RMS within a single processor are unaffected.

B.5 CONCLUSIONS

SIMACS does offer a neat solution to disk volume sharing. However, it does have some limitations:

1. a processor will have exclusive ownership of a volume for virtual I/O until it relinquishes it. This will hold up any virtual I/O from any other processor. This restriction does not apply to physical or logical I/O, since they do not use the ACP.

file structures is delegated to an ACP (Ancillary Control Process). This process is called directly by RMS (Record Management Services) and by the driver. The ACP is called by RMS when a file is first accessed, extended, truncated or deleted and for directory look up. The driver calls the ACP when it needs to know the blocks assigned to a file (i.e. the translation between virtual and logical block numbers). The driver does not call on the ACP at every transfer but keeps translation information in a data structure called a WCB (Window Control Block). The ACP is then only called when the information in a WCB is exhausted. The ACP itself manages a set of disk control files that enable it to allocate and manage space for other files. These files include the index file, storage bitmap file and the master file directory. Sections of these files reside in memory while the ACP is doing file operations and are written back to disk when files are closed or the disk is dismounted. So at any given time the information on the blocks allocated to a particular file resides partly in memory and partly on the disk.

B.3 DUAL PORTING WITH STANDARD VMS

Now consider a disk drive that can be connected to two controllers which are in turn connected to different processors. If the two controllers are doing physical or logical I/O then both can have read/write access to the disk, provided they agree between each other which blocks each will own. When it comes to virtual I/O things don't work as well. There can only be one set of control files, so if both processors are to have write access to the disk they must ensure that each is working with the current copy.

Standard VMS ACP's, drivers and controllers have no mechanism for controlling multiple write access to the control files. This means that only one processor can have write access to the disk and all other processors have only read access. The system manager must decide, when the disk is mounted, whether a disk is to be read/write or read only. Dual porting of DEC disks is then possible within these restrictions.

B.4 SIMACS

SIMACS overcomes the problems of dual virtual I/O by providing a mechanism that allows sharing of the disk control files. To allow this sharing SI has modified the standard disk ACP (F11BACP) (calling it BRBACP), and the standard disk drivers. Special firmware is also needed in the SI 9920 disk controller.

2. the overhead in flushing ACP data to the disk may be significant, especially if two processors are continuously requesting access to a volume.

3. it would introduce more non DEC software into the operating system. DEC shows no hesitation in changing standard system software when they wish to upgrade its capabilities. If you are using a non DEC version of a piece of software you may miss out on the upgrades. A good example of this is the DMAX 16 terminal controller and driver used at Parkes. When DEC introduced the terminal class driver under V3.0 the DMAX missed out on it. So, for example, you cannot use control T to interrogate the system.

4. a SIMACS disk cannot be the system disk. This probably just means that the disk cannot be booted. From a performance point of view I would be hesitant to put any page or swap files on a SIMACS volume.

5. power failure of a SIMACS controller or connected CPU is handled fairly crudely. We would need to write special command procedures to handle an after hours power failure and other times when manual intervention is not possible.

SIMACS may not be the last word in file sharing but it is much than the only other system available, the CI bus from DEC. It is recommended that we purchase a SIMACS system and evaluate its performance on the dual 11/750 system proposed for Epping. From this evaluation we can determine whether the CI bus will be suitable for the Culgoora system. A SIMACS system would seem to meet our requirements for file sharing in the dual system at Epping.

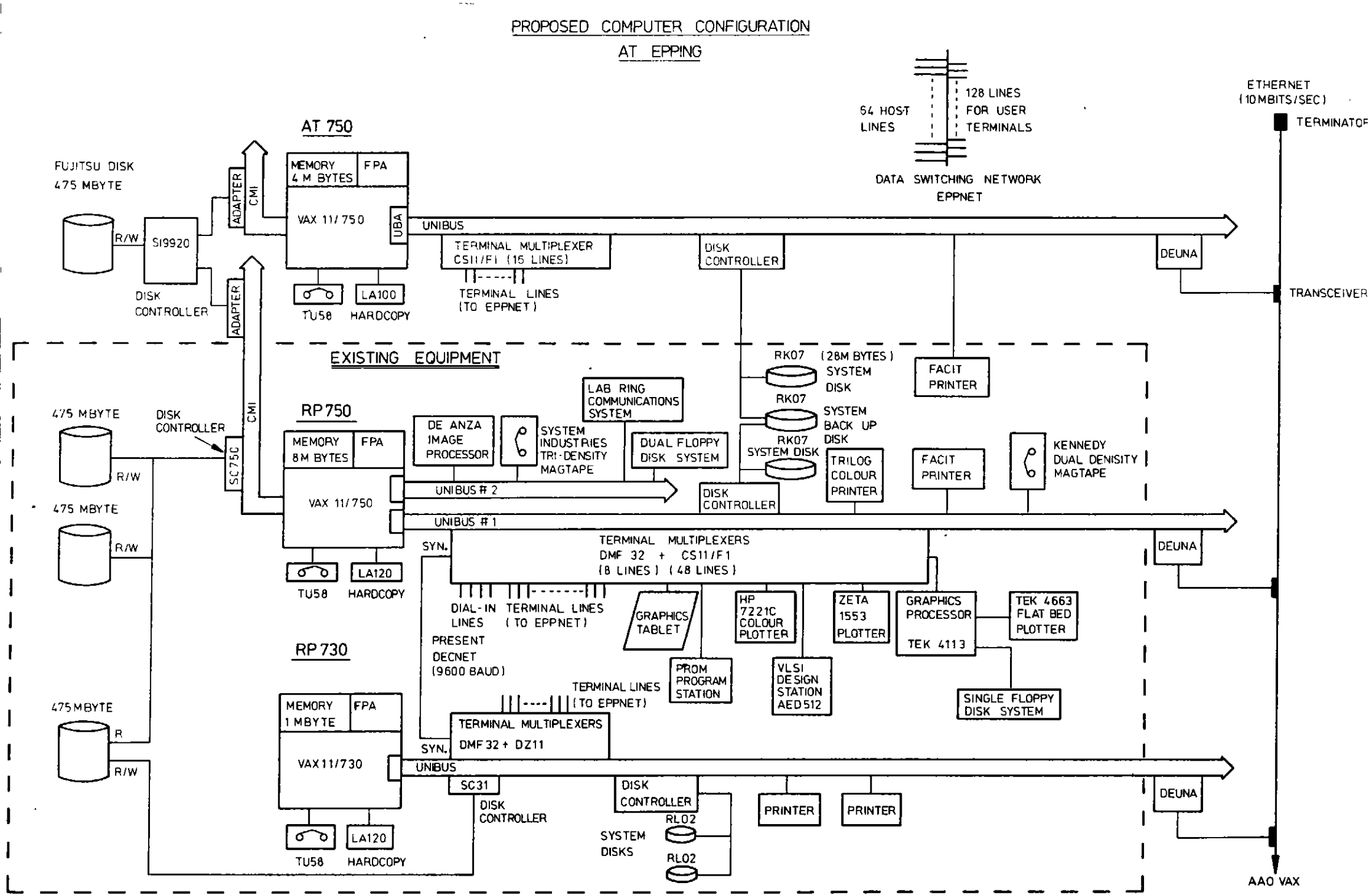
PROPOSED COMPUTER CONFIGURATION
AT EPPING
64 HOST LINES
128 LINES FOR USER TERMINALS
DATA SWITCHING NETWORK
EPPNET
ETHERNET (10MBITS/SEC)
TERMINATOR
AT 750
FUJITSU DISK 475 MBYTE
R/W
SI9920
DISK CONTROLLER
ADAPTER
CMI
MEMORY 4 M BYTES
FPA
VAX 11/750
UBA
UNIBUS
TERMINAL MULTIPLEXER CS11/F1 (16 LINES)
TERMINAL LINES (TO EPPNET)
DISK CONTROLLER
DEUNA
TRANSCEIVER
TU58
LA100
HARDCOPY
EXISTING EQUIPMENT
RK07 (28M BYTES) SYSTEM DISK
RK07
SYSTEM BACK UP DISK
RK07 SYSTEM DISK
FACIT PRINTER
LAB RING COMMUNICATIONS SYSTEM
RP750
475 MBYTE
DISK CONTROLLER
SC750
MEMORY 8M BYTES
FPA
DE ANZA IMAGE PROCESSOR
SYSTEM INDUSTRIES TRI-DENSITY MAGTAPE
DUAL FLOPPY DISK SYSTEM
TRILOG COLOUR PRINTER
FACIT PRINTER
KENNEDY DUAL DENISITY MAGTAPE
VAX 11/750
UNIBUS # 2
UNIBUS # 1
DISK CONTROLLER
TERMINAL MULTIPLEXERS DMF 32 + CS11/F1 (8 LINES) (48 LINES)
SYN.
DEUNA
TU58
LA120
HARDCOPY
DIAL-IN LINES
TERMINAL LINES (TO EPPNET)
GRAPHICS TABLET
HP 7221C COLOUR PLOTTER
ZETA 1553 PLOTTER
GRAPHICS PROCESSOR TEK 4113
TEK 4663 FLAT BED PLOTTER
PRESENT DECNET (9600 BAUD)
PROM PROGRAM STATION
VLSI DESIGN STATION AED512
SINGLE FLOPPY DISK SYSTEM
RP730
475 MBYTE
R/W
MEMORY 1 MBYTE
FPA
VAX11/730
TERMINAL LINES (TO EPPNET)
TERMINAL MULTIPLEXERS DMF32 + DZ11
SYN.
UNIBUS
SC31
DISK CONTROLLER
DISK CONTROLLER
RL02
SYSTEM DISKS
RL02
PRINTER
PRINTER
DEUNA
TU58
LA120
HARDCOPY
AAO VAX

Fig. 1.