OFFLINE COMPUTING AND NETWORKING

J. A. Appel, P. Avery, G. Chartrand, C. T. Day, I. Gaines,
C. H. Georgiopoulos, M. G. D. Gilchriese, H. Goldman, J. Hoftun,
D. Linglin, J. A. Linnemann, S. C. Loken, E. May, H. Montgomery,
J. Pfister, M. D. Shapiro, and W. Zajc

INTRODUCTION

This note summarizes the work of the Offline Computing and Networking Group. The report is divided into two sections; the first deals with the computing and networking requirements and the second with the proposed way to satisfy those requirements.

In considering the requirements, we have considered two types of computing problems. The first is CPU-intensive activity such as production data analysis (reducing raw data to DST), production Monte Carlo, or engineering calculations. The second is physicist-intensive computing such as program development, hardware design, physics analysis, and detector studies.

For both types of computing, we examine a variety of issues. These included a set of quantitative questions: how much CPU power (for turn-around and for through-put), how much memory, mass-storage, bandwidth, and so on. There are also very important qualitative issues: what features must be provided by the operating system, what tools are needed for program design, code management, database management, and for graphics.

PRODUCTION DATA ANALYSIS

The total CPU power for data reduction has been estimated by extrapolating the experience of UA1 as described by D. Linglin in these proceedings. The offline computing takes 4 seconds per event on an IBM 3081K. This is mostly for the central tracking and is approximately half track-finding and half track-fitting. We have tried to extrapolate accounting for the increase

in multiplicity and the expected increase in the number of detector channels. Another issue is that the high-level triggers may be expected to increase the complexity of events but at the same time may provide some of the processing on the sample that does pass the trigger requirements.

With these uncertainties, the time per event is expected to be 60 seconds on the 3081K or 1200 seconds on a VAX-11/780. The computing group at Snowmass 84 calculated a similar number but the result is not completely independent since the UA1 was also used as a starting place. Assuming a recorded event rate of 1 per second, we would need the equivalent of 1200 VAX to process the raw data. Because the event sample can easily be divided into parallel processors this load specifies the required through-put but does not constrain the speed of the single processor units.

We have taken 1 Megabyte as the average event size although the data acquisition group indicates that this may be a factor of 2-3 high depending on how much the data are packed in the front-end processors. The data-recording system must have an average recording rate of 1 Megabyte per second. However, peak rates of 3-5 Megabytes per second capacity may be needed to avoid deadtime. Assuming 30% duty-cycle, a single large experiment will record 10 Terebytes a year. The output of the data processing will be comparable.

The typical analysis code for event analysis will require approximately 16 Megabytes of memory in each processor. The event record and output are included in this estimate. We take this as the minimum memory size for all the processors. Single processor-multiple job systems will require 2-3 times this for efficient resource utilizaton. This does not include memory for any diagnostic summaries or other large arrays.

Network links will be required to move sample events to collaborating institutions for study during detector installation and testing or during

data-taking if there are problems with some component. To move sample events to other institutions we require a network bandwidth of at least 100 kbits/sec. At that speed, a single event (1 Megabyte) can be transfered in 80 seconds, assuming full line utilization. The question of high performance networks is discussed in a note by Greg Chartrand in these proceedings.

The problem of data processing emphasizes the CPU power aspects of any computer system. In fact, many experiments have even managed to analyse all their data with such CPU powerful, but "user-hostile" computer systems as the CDC 7600 or CYBER. For experiments with a large volume of data running over a long time, it is necessary to have a number of features to improve the efficiency of production. These include a flexible command language to automate job submission and to provide a summary of errors in the production job stream. The computer must provide a full database management system to track the production history and data storage situation, to allow for changes in running conditions and to include calibration data for all runs. The calibration time-intervals for detectors will be varied and the calibration database must track changes for all devices.

The computing environment must provide operator support for system back-up and tape and disk management. In addition, the system must support running all analysis programs in multiple locations. The code may be tested and run at any collaborating institution and will also migrate to the high level trigger as the experiment matures and the trigger becomes more sophisticated.

PRODUCTION MONTE CARLO

The problem of generating large samples of events by Monte Carlo is, in many respects, similar to the production requirement. The number of events will scale with the size of the real event sample. The events can be distributed to many parallel processors. In addition, the same tools are

needed to manage a large production run. The Monte Carlo does not define any new system requirements.

ENGINEERING

We consider here engineering calculations which require significant CPU power for a single calculation. These include the calculation of magnetic-field or stress calculations for a detector. In many of these, the calculation may be efficiently run on a vector-supercomputer. At the present time there does not seem to be any easy way to divide these problems so that they can run on a large array of small processors. We assume that the central facility at the SSC must include a very fast computer with vector capability. This system must have at least the speed of a CRAY 1 (approximately 20-40 times the scaler speed of a VAX-11/780). This system will also be used for physics analysis where the turnaround is as important as the through-put.

PROGRAM DEVELOPMENT

The program development effort for a large SSC experiment will be very large. Experiments with large detectors today have developed approximately 500,000 lines of code each. The SSC experiments will have well over 1,00,000 lines of code. If we assume an average productivity of 1000 lines/man-year, the cost of software will be a large part of the cost of detectors.

It is extremely important to improve the productivity of the developers and to improve the quality and the reliability of the resulting software. To do this, the computer systems must provide a broad range of tools to assist in the design and management of software. These issues were discussed extensively and are reviewed in two separate reports in these proceedings.

In addition to tools for design, code management, and automated software testing, the systems must provide standard libraries for graphics and database

management. To reduce the over-all cost of software development, it may be possible to share common programs among experiments.

It is extremely important to have network links between all the institutions where this development effort is taking place. These links will be used to transmit design documents, program libraries, and data bases. The network software must be integrated with the code and library management tools so that software at all institutions can be maintained simultaneously.

While the development effort emphasizes the qualitative aspects of the computers, there are also significant needs for computing power. Each developer requires the equivalent of approximately 20% of a VAX-11/780 in terms of through-put. Assuming 1000 users, the total capacity must be approximately 200 VAX. Many of the problems in the analysis stage cannot be efficiently run in parallel processors and would be directed to a central computer with significantly faster serial turn-around.

PHYSICS ANALYSIS

This aspect of the computing problem is highly interactive and it is crucial to maintain high productivity. The system must provide libraries to allow easy access to the data. Because of the large volumes of data for each experiment, it will be necessary to use new database structures to extract events or variables of interest. The use of new large-volume, random-access devices such as video disks may allow significant improvements in data storage and access.

To maintain high productivity, the computers must provide fast response. We have assumed that the system specified for engineering and development is sufficient to support the analysis effort.

The analysis effort will also require extensive network links to share data bases and programs. It will be necessary to transmit large graphics

files and documents. The use of standards for documents and graphics will significantly improve productivity.

HARDWARE DESIGN AND STUDY

These problems are very similar to those of development and analysis. We assume that they do not significantly change the system requirements as described above although they will require that the computer systems and network links be installed during the early stages of experiment design and well in advance of the commissioning of the Collider.

A MODEL FOR CENTRAL COMPUTING AT THE SSC

In this section we will describe a model for central computing facilities at the SSC which satisfies the general specifications given above. A schematic picture of the central facility is given in Fig. 1, assuming that the permanent primary data recording from each experiment occurs locally at each intersection region, as discussed below.

RECORDING THE RAW DATA

There are two options for permanent primary data recording. Permanent recording of the data may be done at each intersection area and the recording media physically transported to the central facility. Alternatively, high speed links from the intersection regions to the central facility could be used to transfer data in realtime to be recorded at the central facility.

We have assumed that the peak data rate will be about 5 Mbytes/sec in approximate agreement with the estimates of the data acquistion subgroup. Curt Canada has provided an introduction to some of the current methods of mass storage in his paper submitted to this Workshop. Present relatively cheap tape storage such as the IBM 3840 tape cartridge is limited to 3 Mbytes/sec and stores 200 Mbytes/cartridge at 14$ per cartridge. It seems

likely that this will be improved substantially in storage capacity (perhaps by 10) and somewhat in speed by the 1990's. Canada also describes more expensive, faster and higher density magnetic tape systems which would already meet our specifications. Optical disks which already satisfy the capacity requirements need to be improved in writing speed for our needs. Although it is difficult to predict the future in this area, it seems likey that tape/optical disk drives of reasonable cost will be available for the SSC. If this is so, then data recording at the intersection region becomes possible.

Centralized recording (such as is done at KEK and DESY) is also a possibility. High speed, presumably optical fiber, links would be required from each intersection region to the central facility. The advantage of this system is less duplication of facilities and the possibility of better dynamic allocation of facilities as demand changes. The disadvantages are the cost of the high speed links, although such links, perhaps at a lower speed, will be required for communications between the central facility and the IRs. The exact choice will depend on the cost of data recording hardware and high speed links.

DATA REDUCTION

In our model we assume that essentially all of the data reduction (data reduction of raw data to data summary files) will be done by farms of microprocessors or equivalent. By farms of microprocessors we mean very cheap parallel computing of the one event per processor type. At present this is the only potentially feasible, cost effective means to handle the enormous demand.

We have furthermore assumed that by the early 1990's a single microprocessor will have the power of 8 mips (1 mip = a VAX11/780) and memory sizes up to 16 Mbytes at a cost equivalent to today's technology, about 1 mips and 4

Mbytes. As we will see in the cost estimates at the end of this report, if this is optimistic by a factor of two it has a small impact on the overall cost of the central facility.

Each farm of microprocessors is driven by a dedicated host computer (presumably of a VAX 8600 class or less) with dedicated tape/optical disk drives for access to the raw data and for output recording. In our model there are at least 8 production farms, each of 125 nodes or 1000 mips total.

In addition to the production farms, there is a development farm for code development and testing new microprocessors. This might have about a 200 mips capacity.

It is also possible that cheap vector processors will be available on the time scale of the SSC. If so, a vector farm capacity could be provided at a reasonable cost. Such an option may become increasingly important if parts of code used in HEP become vectorized to enhance speed.

To manage the farms and for program development related to data reduction and some Monte Carlo, as discussed later, one needs a facility with at least 100 mips capacity and extensive software tools. Logically this is equivalent to today's mainframes or clustered superminis. This program development facility and farm manager would have access to peripherals including mass storage, tape/optical disk drives and high speed disks; essentially everything on site.

MONTE CARLO

As indicated above we assume that the microprocessor farms are also employed for the bulk of Monte Carlo simulations of experiments. Again it is crucial to the final cost of the facility that such farm capability exists for Monte Carlo needs. The earlier these Monte Carlo studies begin, the more effective can be the detector design.

PHYSICS ANALYSIS AND ENGINEERING SUPPORT

Although we expect the collection of microprocessor farms to provide the enormous capacity for data reduction and Monte Carlo, we do not anticipate that they would be used for physics analysis or for many aspects of engineering support. The one event-one processor system is not designed for fast turnaround. It is crucial for physics analysis to have high speed as well as the capacity to handle hundreds of users. For this reason we have included in the central facility an element with

- > 40 mips/ processor

- > 200 mips total

- vector as well as scalar ability

In today's terms this element would represent the top of the line mainframe but assumes that vector capability will be added without reducing the relatively user friendly environment of today's mainframes. It seems likely that this element will be a small collection of very powerful processors, each with the speed given above. Compatibility with the farm manager/program development machines(s) would clearly be useful. Compatibility with off-site computing must also be considered. This would likely rule out radical but possibly more effective architectures that may be developed in the future.

TERMINALS AND WORKSTATIONS

A considerable number of the engineering functions may be performed on workstations with advanced graphics capabilities. In our model such workstations are "driven" by a separate computer, the CAD computer, which in turn communicates, if necessary, with the large analysis computer. A need for about 75 such workstations is anticipated.

To handle the anticipated number of users, about 2000 terminals will be required, extrapolating from the present Fermilab number of terminals and users. About 10% of these might be color and they all will likely have graphics capabilities.

In addition to terminals, some workstations and more sophisticated graphics ability will be required. Perhaps 50 workstations with state of the art graphics would be required.

OTHER PERIPHERALS

Data and program storage is provided by a variety of means. For program storage one needs fast access devices (disks) with about 200 Gbytes of capacity. In addition one will clearly need magnetic tape and/or optical disk drives with a total capacity equivalent to about 100 6250 tape drives of today. A capacity of about 10 Tbytes will be needed in a mass storage device. All of these would be available to both the farm manager/program development computer and the analysis computer.

NETWORKING

It is crucial to provide within the central facility access to users and their computers off-site as well as connection to the interaction regions. Remote logon and at least modest ability for file transfer are required. Assuming a peak demand of 250 users then the link to an external network (such as HEPNET) would require at least a 1.2 Mbaud line.

It would also be desirable to have higher rate capability either via a satellite link and/or land lines to transmit graphics information and possibly data such as summary "tapes". Clearly the latter could also be done, at a much slower rate, but a lower cost via conventional shipping means. The ability to transmit graphic information, however, requires very high speed

links. It may be more cost effective to provide the graphics hardware at a central location and the appropriate links, rather than distributing the hardware.

COSTS

We have quickly and crudely made a cost estimate of the central facility shown in Fig. 1. The input to this cost estimate was the collective experience of the members of the working group and assumes some extrapolation of performance/cost. A summary of costs appears below.

ITEM	COST (M$)
Eight farms including drivers and peripherals	3.0
Farm manager/program development	2.0
Analysis computer(s)	20.0
Mass storage (10 TB)	5.0
"Disk" (200GB)	4.0
= 100 "tape drives"	1.2
2000 terminals and connections	4.0
CAD computer	0.5
CAD workstations	1.5
Physics workstations	1.0
HEPNET gateway	0.2
Earth station	0.8
Internal networking	1.5
TOTAL	45M$

These costs do not include any personnel costs.

PERSONNEL

We also made estimates of the number of personnel needed to operate such a central facility. This is discussed in more detail in the contribution of Jack Pfister in these proceedings.

TIME SCALE

An estimate of the required time development of our model is given in Fig. 2. Much of the initial programming effort for proposal writing and such must be done at existing or near future facilities.

It is of considerable importance to provide enough computing early on at the SSC site such that program development under a known system may begin quickly. For large detectors there is a very large investment in programming which will be primarily done by physicists.

SUMMARY

The crucial ingredient in our model of a central computing facility for the SSC, is the reliance on cheap farms of microprocessors for most of the computing needs. It is clear that without the implementation of such farms, either within high energy physics or by industrial sources, SSC computing cannot be done without an enormous and unacceptable increase in the cost. We must have both the hardware and software ability to make microprocessor farms work. The other components of our model appear to be well within reasonable extrapolations of today's computing related technology. New ideas are not needed to satisfy the needs that microprocessor farms cannot supply.
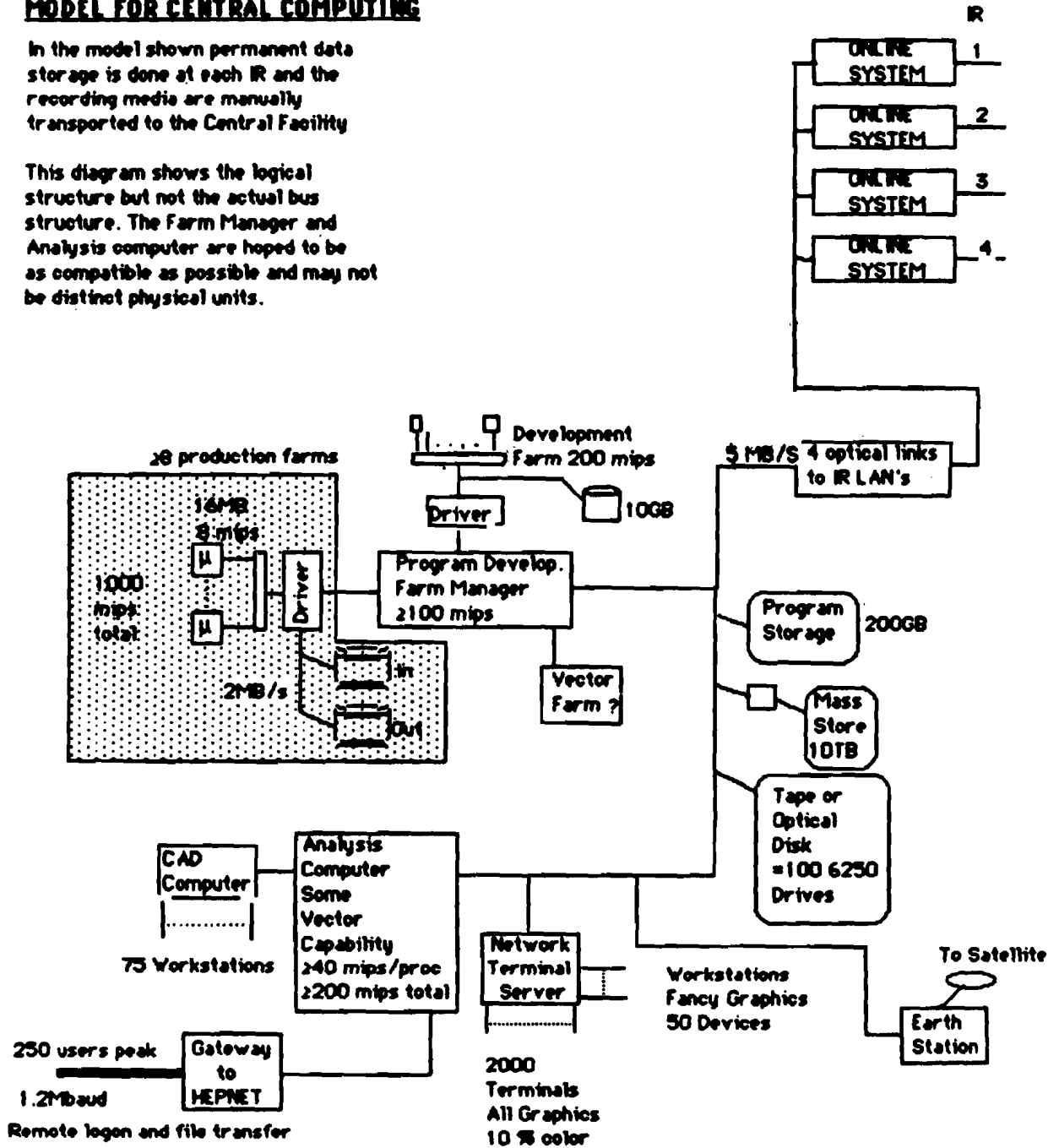
FIGURE 1

SSC

MODEL FOR CENTRAL COMPUTING

In the model shown permanent data
storage is done at each IR and the
recording media are manually
transported to the Central Facility

This diagram shows the logical
structure but not the actual bus
structure. The Farm Manager and
Analysis computer are hoped to be
as compatible as possible and may not
be distinct physical units.

IR

ONLINE SYSTEM   1

ONLINE SYSTEM   2

ONLINE SYSTEM   3

ONLINE SYSTEM   4

x8 production farms

Development
Farm 200 mips

5 MB/S 4 optical links
to IR LAN's

16MB
8 mips
μ

1000
mips
total

μ

Driver

10GB

Driver

Program Develop.
Farm Manager
≥100 mips

Program
Storage   200GB

2MB/s   In

Out

Vector
Farm ?

Mass
Store
10TB

Tape or
Optical
Disk
=100 6250
Drives

CAD
Computer

Analysis
Computer
Some
Vector
Capability
≥40 mips/proc
≥200 mips total

To Satellite

75 Workstations

Network
Terminal
Server

Workstations
Fancy Graphics
50 Devices

Earth
Station

250 users peak

1.2Mbaud

Gateway
to
HEPNET

2000
Terminals
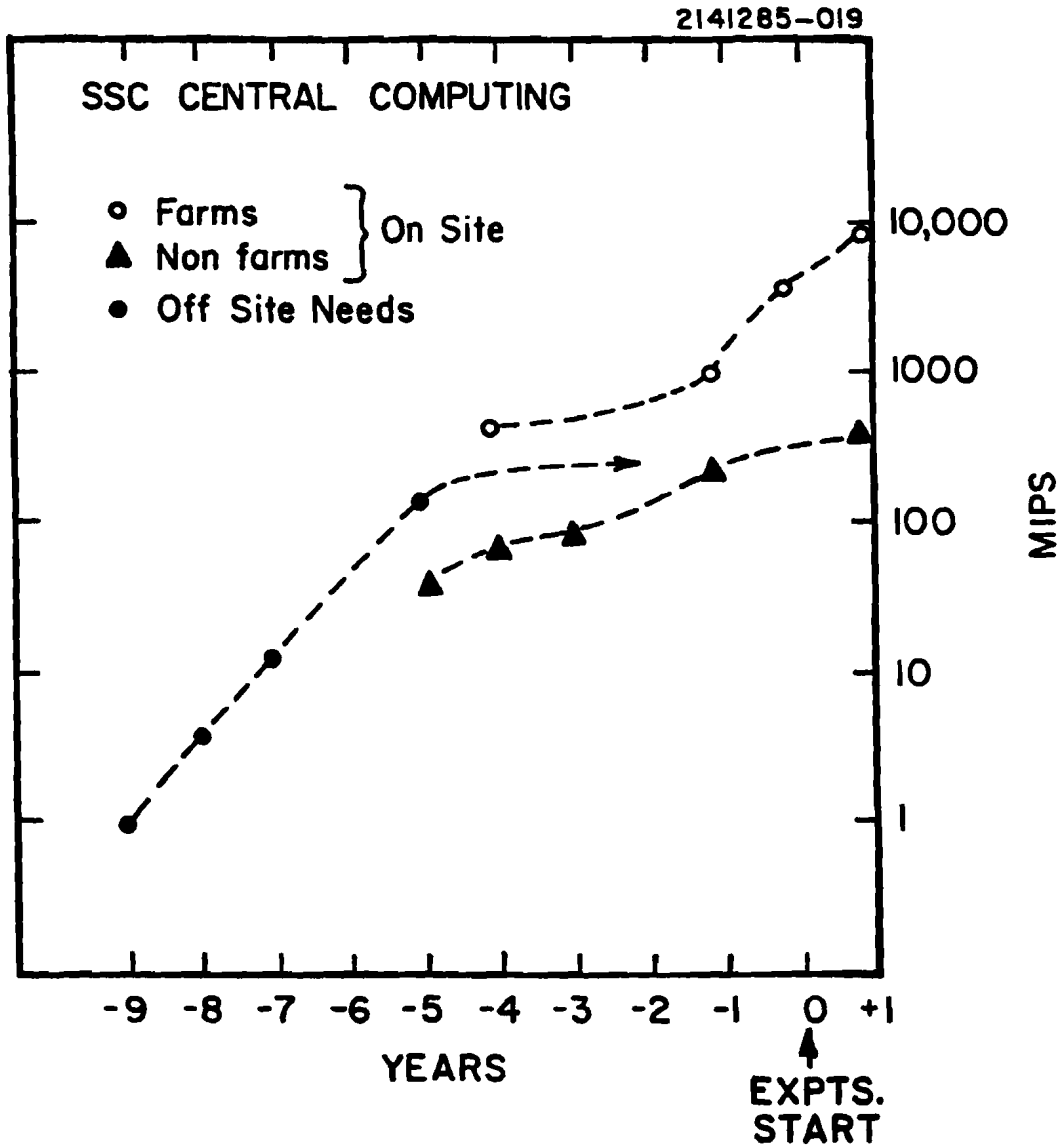All Graphics
10 % color

Remote logon and file transfer

FIGURE 2

TIMESCALE FOR DEVELOPMENT OF SSC COMPUTING