Extended Operating Configuration 2 (EOC-2) Design Document

David Barkai[1]

RND-94-003 January 1994

NAS Systems Development Branch
NAS Systems Division
NASA Ames Research Center
Moffett Field, CA 94035-1000

barkai@nas.nasa.gov

Abstract

This document describes the design and plan of the Extended Operating Configuration 2 (EOC-2) for the Numerical Aerodynamic Simulation division (NAS). It covers the changes in the computing environment for the period of '93-'94. During this period the computation capability at NAS will have quadrupled. The first section summarizes this paper: the NAS mission is to provide, by the year 2000, a computing system capable of simulating an entire aerospace vehicle in a few hours. This will require 100 GigaFlops (10^{11} Floating point operations per second) sustained performance. The second section contains information about the NAS user community and the computational model used for projecting future requirements. In section 3, the overall requirements are presented, followed by a summary of the target EOC-2 system. The following sections cover, in more detail, each major component that will have undergone change during EOC-2: the high speed processor, mass storage, workstations, and networks.

Contents

[1] Computer Sciences Corporation, NASA Contract NAS 2-12961, Moffett Field, CA 94035-1000

1. Summary

1.1 Mission

The NAS mission is "to provide the Nation's aerospace research and development community by the year 2000 a high-performance, operational computing system capable of simulating an entire aerospace vehicle system within a computing time ranging from one to several hours" (from Ref.1—the NAS Program Plan). This mission supports the objectives of the NAS program, which are aimed at assisting U.S. aeronautics maintain dominance of the world's aircraft market. These objectives are:

- Act as a pathfinder in advanced, large scale computer capability through systematic incorporation of state-of-the-art improvements in computer hardware and software technologies.
- Provide a national computational capability, available to NASA, DoD, other government agencies, Industry and Universities, to insure continuing leadership in computational fluid dynamics and related computational aerospace disciplines.
- Provide a strong research tool for the Office of Aeronautics.

Fulfilling these objectives, and thus advancing the NAS mission, has a number of practical and significant benefits. NAS provides for:

- Accelerated vehicle development cycles
- Critical design data that cannot be obtained from experiment
- Reduced development costs and risks
- Reduced need for ground and flight testing
- Full aerodynamic optimization of an entire vehicle
- Increased performance, safety and stability through integrated vehicle design
- Increased vehicle operational efficiency

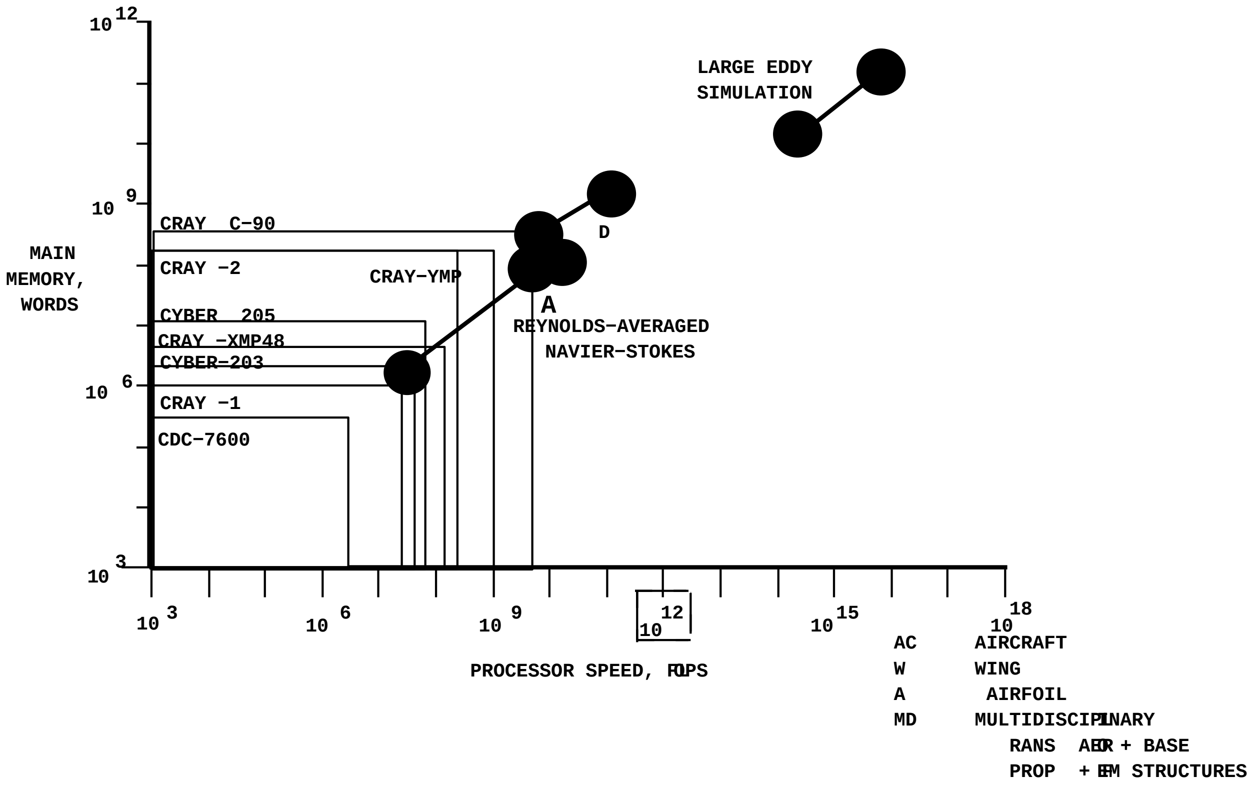
The computational tasks are defined in terms relevant to the designers and engineers (e.g., Controls, Stress, Propulsion, etc.). The characteristics of the applications are, in turn, expressed in terms of computing loads on the system. That is to say, answering the question, "What are the sustained computing speed, memory size, storage, and network capacity required to deliver the results within one to several hours?" This analysis, combined with projections of future technologies, is the basis for a multiple-phase plan for NAS to achieve its mission.

Figure 1 shows the relationship between types of computations and the processing requirements—the speed of the computer system and the size of memory needed. Computational tasks have been defined that require up to 10,000 Teraflops, with main memory of about 100 GigaWords, for simulation runs done in 15 minutes (Large Eddy simulations of full aircraft). Even if the job turn-around is 10 times slower (a couple of hours) the need still exists for enormous capability. To put it in perspective, we are today at the level of 3-5 GigaFlops with a 16-processor Cray C90—a factor of 2,000,000 times below that required for the much more complex Large Eddy Simulations (LES).

Figure 1:

AERONAUTICS MODELING AND SIMULATION

15 - MINUTE RUNS



There are phases the NAS capabilities must step through before reaching the final goal. The current mission statement addresses the NAS goal through the year 2000. By that time it is hoped that the 1 Teraflops level of sustained performance for aeronautics computations may be reached. The current thinking is that, though TeraFlops-level applications will be demonstrated, the average sustained performance may be up to 10 times lower (at the 100 GFlops range). Nevertheless, this level will provide Reynolds-Averaged Navier-Stokes simulations capability for a whole aircraft, and the ability to run significant multi-disciplinary codes. At this level the numerical simulations will enable NAS, to a large extent, to realize the benefits stated above. It is also clear that there is a defined need for a many-

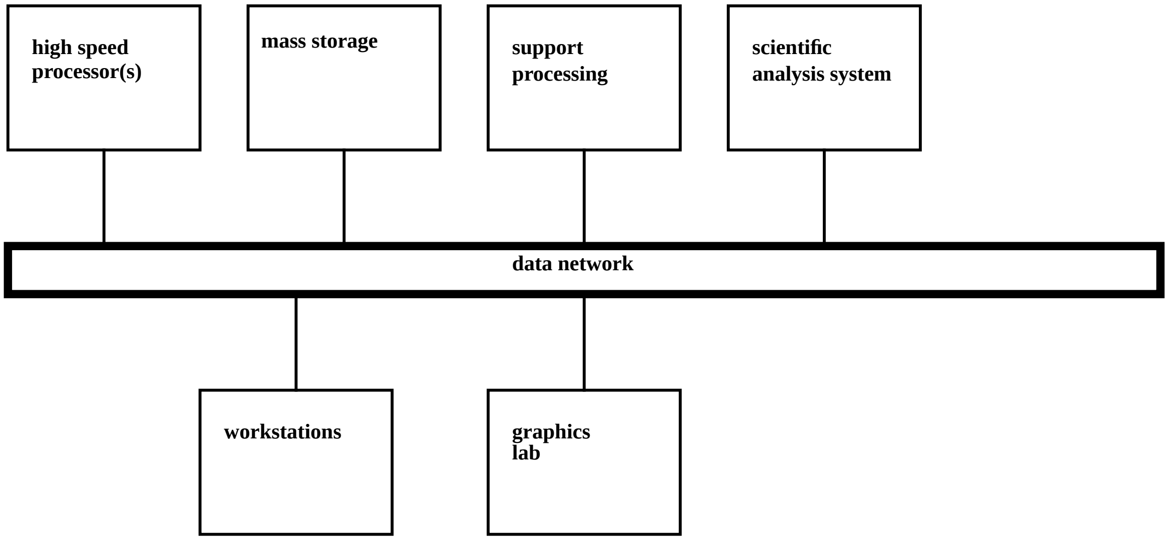
Page 3

fold improvement in computing power, beyond the end of this century, before the simulations requiring thousands of TeraFlops can be accomplished.

1.2 History

The main components of the NAS Processing System Network, or NPSN, are depicted in Figure 2. The high performance computational components are the 'high speed processors.' This has so far been a multi-vector-processor shared-memory system, or what we may call a 'traditional supercomputer.' This is the production machine for the compute intensive CFD codes. The 'scientific analysis system' is made up of a computational engine in conjunction with workstations. Its function is to provide a platform for post-processing of the numerical results from the supercomputers. It transforms raw numerical data into meaningful visuals, which provide insight to the numerical simulations. The other components support and connect these computers and make it a complete and consistent computing environment. The 'mass storage' subsystem contains processors that manage the central on-line file system and tape archival system. The 'workstations' include the desktop computers, and other workstations clustered, locally and remotely, with the scientific analysis processor. The 'support

Figure 2. NPSN Functional Components
(production systems only)



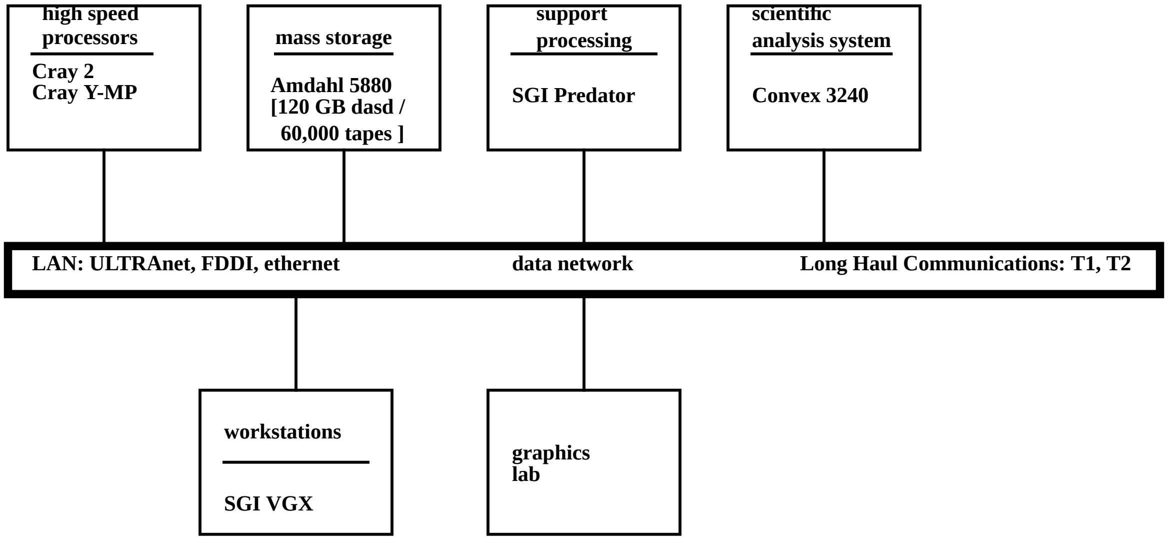
processing' component is a collection of large workstations used to manage common software and files. And the 'graphics lab' serves to create advanced visuals (e.g., videos), and to examine new graphics products. All these components are connected together by the 'data network,' which includes both local area network and long haul communications.

As technology and requirements evolve, these components are upgraded in a phased implementation fashion, which preserves the balance between them, and within the total system.

The NAS program started its full-capability operations in March 1987 when the Initial Operating Configuration (IOC) was declared functional. The performance goal set, and achieved, for the IOC was that of a sustained 250 MegaFlops on the NAS workload. The High Speed Processor (HSP) was a Cray 2, running UNICOS on its four processors, with 256 MWords of memory. An integrated Support Processor Complex provided 190 GBytes of mass storage and support processing functions through two Amdahl 5840 processors with a UNIX interface to the native operating system (VM), with communications via HyperChannel adapters and Ethernet adapters. The workstations consisted of 35 SGI 2500 Turbo and 3030 Iris workstations running SGI's UNIX operating system (IRIX) modified to include Berkeley-based networking features. The Configuration was complemented by four VAX 11/780 computers used for local and remote gateways, general production work by scientific users, and for software development and maintenance. They all ran 4.3 BSD UNIX. See Ref. 2 for more details.

The next phase, the Extended Operating Configuration (or EOC) was completed in the beginning of 1992. The performance goal for EOC was to operate at sustained rate of 1 GFlops. This was nearly achieved, with about 800 MFlops observed for a daily average. In its final form, EOC contained a Cray Y-MP (with 8 processors and 256 MWords of memory) in addition to the Cray 2, thus about quadrupling the compute power and memory size of the HSP component. These were connected to the next-generation workstations, Silicon Graphics VGX units, over high speed local networks—UltraNet, and FDDI and Ethernet. An Amdahl 5880 provided access and management of mass storage. Four SGI Predator class machines were used for general support processing, and a Convex 3240 was the hardware for the Scientific Analysis System (SAS) for pre- and post-processing CFD solution sets. Remote communication was accomplished using AEROnet, a routed network architecture, with multiples (4, 2, and 1) of T1 (1.5 Mbits/sec), and 56 Kbits/sec in the backbone circuits. Figure 3 shows a diagram of EOC.

Figure 3. NPSN EOC Functional Components



1.3 EOC-2 Summary

The Extended Operating Environment 2 (EOC-2) is the next phase in a multi-year plan towards a computing facility at NAS with sustained TeraFlops performance for NASA CFD codes. EOC-2 is currently in the process of being implemented. The plan for EOC-2 follows the same design assumptions adopted by NAS for the earlier period. Some of the important considerations are:

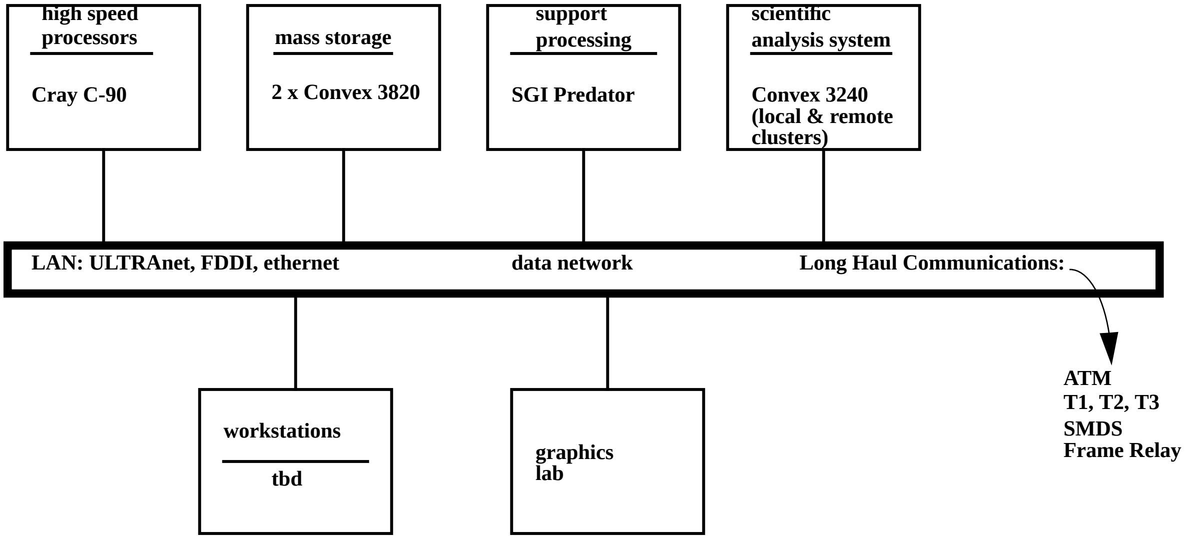
- Maintain two generations of HSPs concurrently.
- Follow the technologies available for the NPSN components such that state-of-the-art products are available to the NAS users.
- Use UNIX operating systems throughout the NAS complex.
- Use TCP/IP for file migration and management.

- Provide a level of service to remote users as close as possible to that given to local users.
- Maintain system balance as components are being upgraded.

These assumptions are consistent with achieving the NAS objectives as stated earlier.

The main components of EOC-2 will be described in greater detail in the following sections. To set the stage, Figure 4 depicts the equipment used, or planned, for EOC-2. The main changes between EOC and EOC-2 are as follows: the Cray C-90 has been added as the most powerful high speed processor and the Cray 2 has been retired; two Convex 3820s have replaced the Amdahl to handle the mass storage (NAStore) and storage units will be upgraded; next-generation workstations will be procured; and the long haul communications capability will

Figure 4. NPSN EOC-2 Functional Components

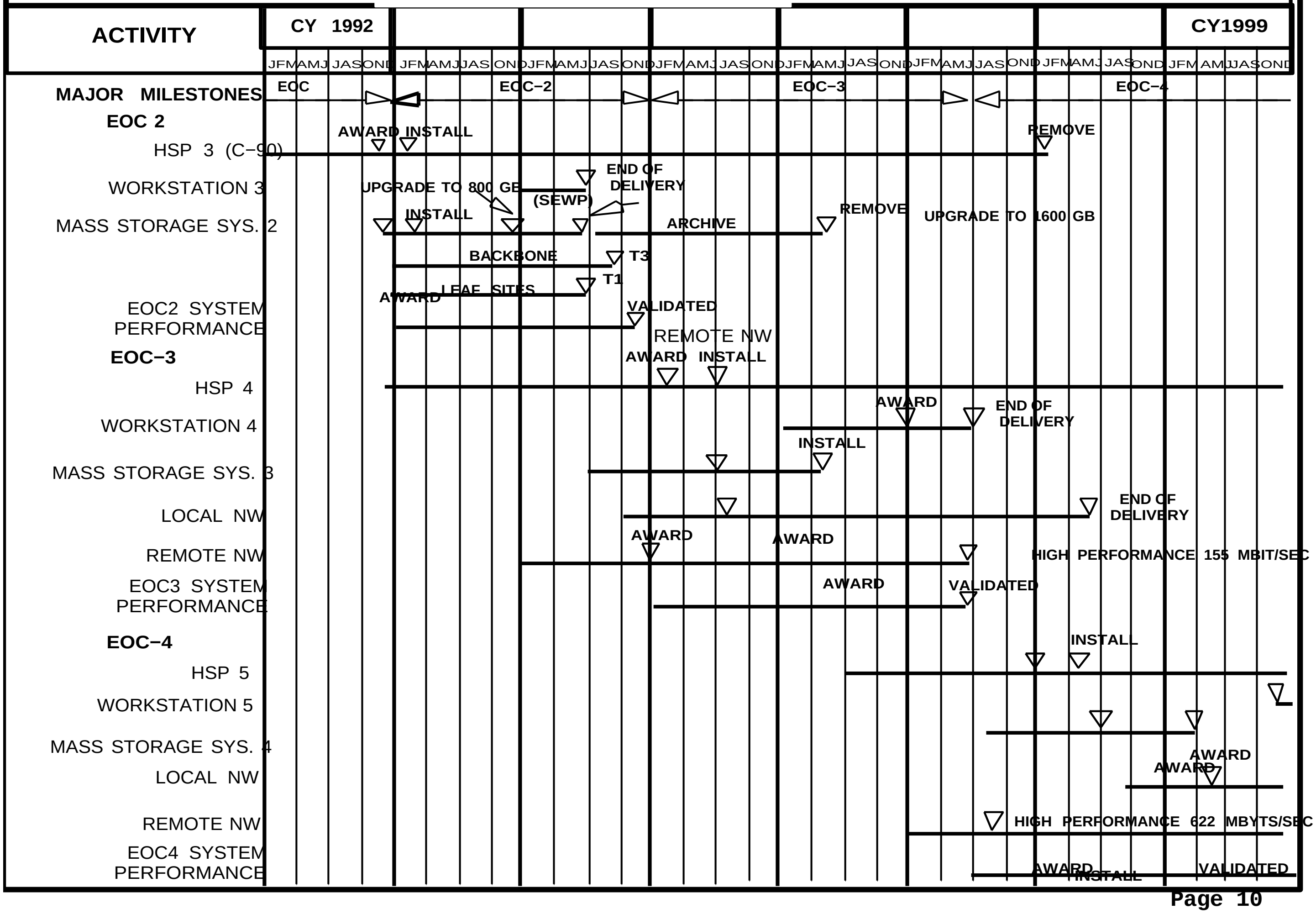


be significantly upgraded by the use of new technologies. There will be no significant changes for ‘support processing.’

In summary, IOC delivered 250 MFlops and EOC-1 had a sustained production performance of 700-800 MFLOPS. The EOC-2 system is expected to deliver 3.5-4 GFLOPS for regular workload.

The EOC-2 phase extends over 1993 and 1994. Its major milestones can be seen in Figure 5, (from Ref. 1—‘NAS Program Plan’). The chart also shows EOC-3, and EOC-4 which ends in 1999. The reader will notice that these phases build on each other and are somewhat intertwined.

Figure 5: NUMERICAL AERODYNAMIC SIMULATION PROGRAM
Program Plan Master Milestone Schedule



2. NAS Workload

2.1 NAS User Community

The NAS facility is a national resource, providing computational resources to researchers through a peer review process. The usage is organized, and approved, by projects. Typically, there are several accounts, or users, for each project. The number of projects has been between 350 and 400 over the last couple of years. The number of users peaked in '89/'90 at over 1,700. Since then, due to a practice of limiting projects to those requiring over 200 hours of machine time, the number of users has fluctuated between 1,000 and 1,400. It is expected to stay at this level. Figure 6 summarizes the level of projects and user accounts at NAS since '87. First, and foremost, the NAS user community comes from NASA centers and facilities. By the end of '92 two-thirds of all the user accounts were from NASA sites (see Fig. 7).

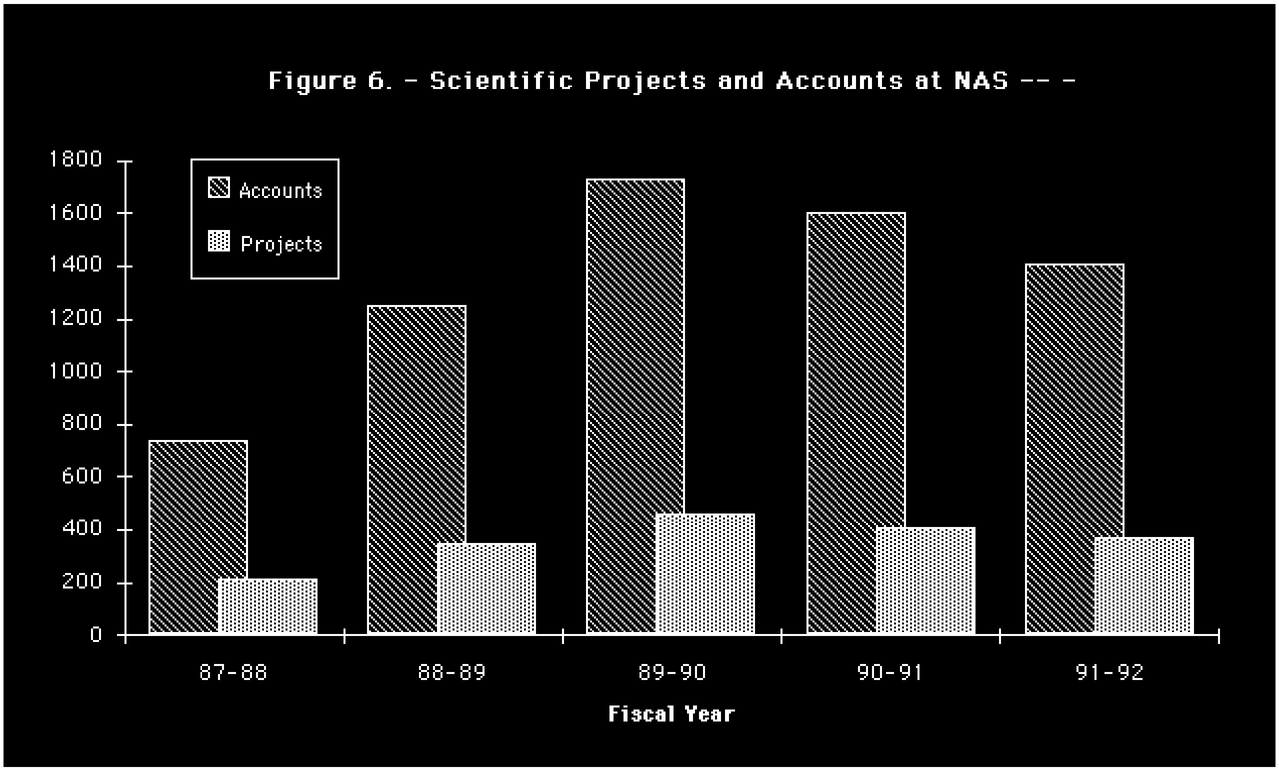


Figure 6. - Scientific Projects and Accounts at NAS -- -

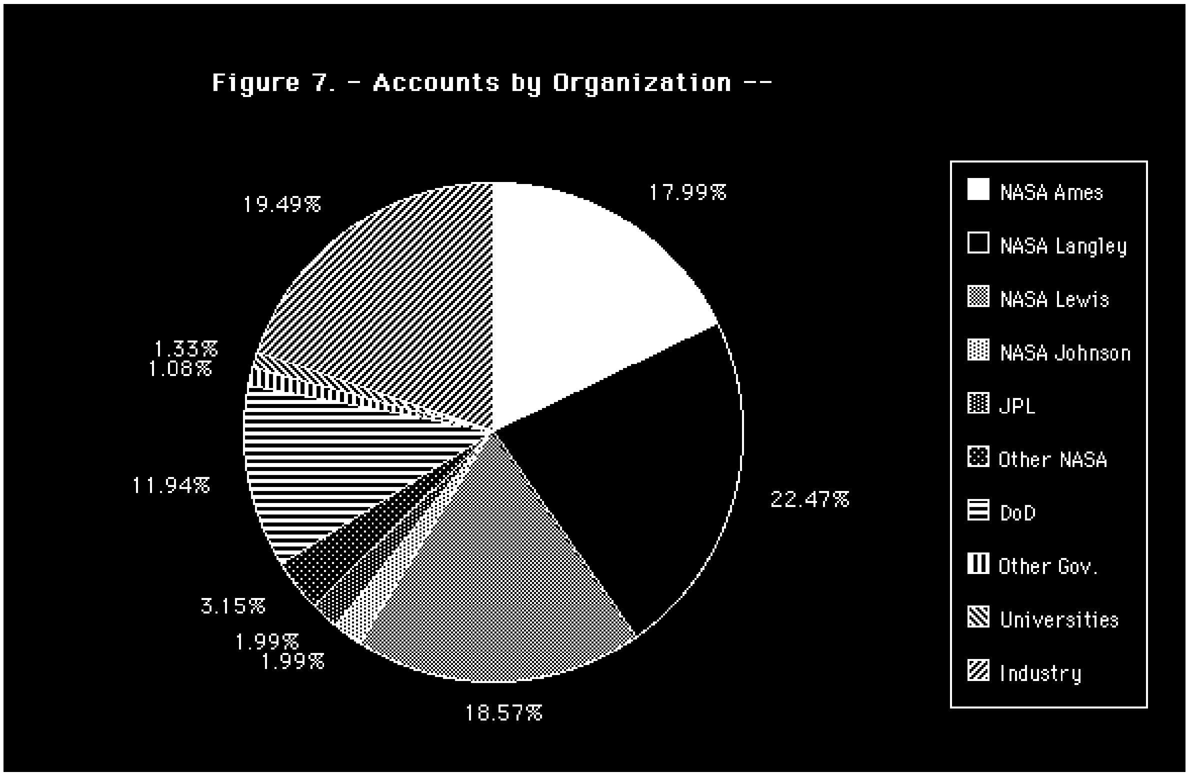
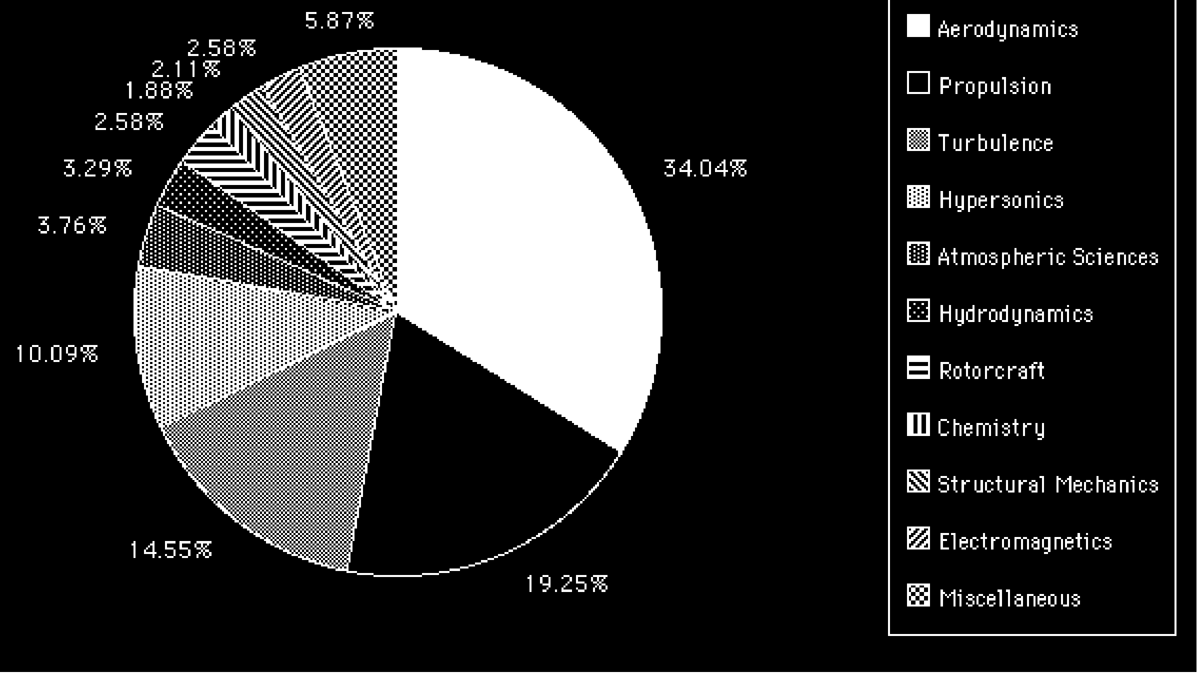


Figure 7. - Accounts by Organization --

Industry, or commercial aircraft and aerospace companies, amounted to nearly 20% of the users; 12% of the users came from DoD, with the remaining 2% from Universities and other Government agencies.

A variety of disciplines are supported by NAS, but the emphasis is overwhelmingly on aeronautics. The list of projects under aerodynamics (Propulsion, Turbulence, Hypersonics, Hydrodynamics, and Rotorcraft) adds up to over 80% of the projects, as can be seen from the breakdown of projects by disciplines taken from the NPSN Monthly Report (Ref. 3) and depicted in Figure 8. When we add related fields in chemistry, structural mechanics, and electromagnetics, we cover over 90% of the projects. The remaining disciplines include atmospheric sciences, astrophysics, and a few miscellaneous projects.

Figure 8. - Projects by Disciplines (Sep. 93) --



It is, arguably, even more significant to look at the amount of utilization of computer resources, since not all projects and users have similar levels of usage. The information below reflects the cumulative data from the first 9 months of '93 taken from Ref. 3. When usage is considered by discipline (Fig. 9), the aeronautics disciplines amount to nearly all the processing time. Apart from Chemistry, all the other disciplines (Atmospheric, Astrophysics, Structural Mechanics, etc.) account for too little to be included. Aerodynamics, Propulsion, and Turbulence use over 80% of the total available resources. Figure 10 shows how the utilization is composed by organization, or 'project class.' Local usage, from NASA Ames, is only 23% of the total. NASA sites use two-thirds of the HSP compute time (interestingly, these sites also account for two-thirds of the number of users). It is expected that the Commercial use, by the private sector, will increase from its current 18%. DoD, with close to 15%, is the other large user sector.

Figure 9. - HSP Utilization by Discipline (1-9/93) --

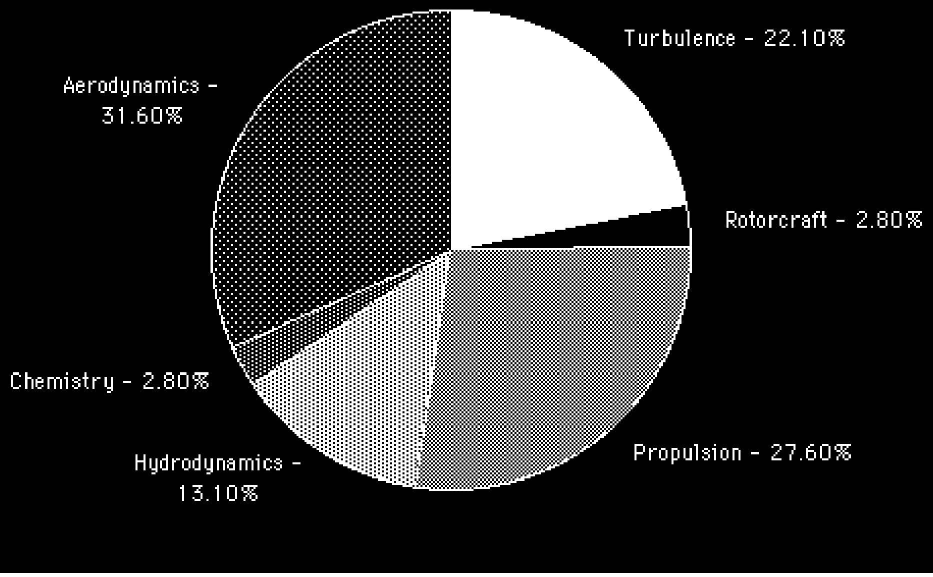
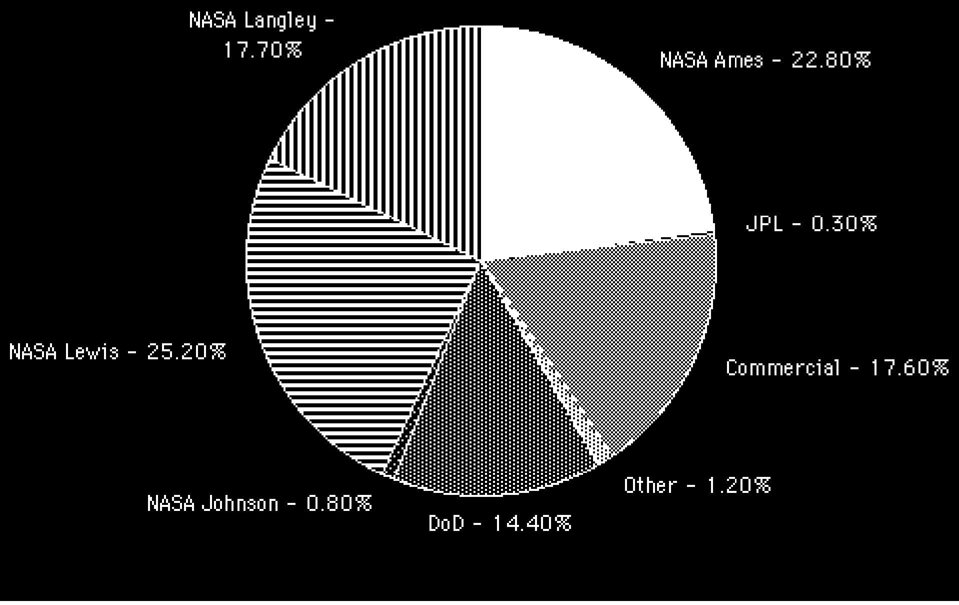


Figure 10. - HSP Utilization by Organization (1-9/93) --



As mentioned above, the NAS users are distributed across many organizations. Less than 20% of the users are local to NASA Ames (Fig. 7). All the others (4 out of every 5 users) are remote users relying on the long-haul communications facilities. This aspect of the users' distribution, as well as collaborative projects with NASA Langley and NASA Lewis, explains the attention given to the remote networking capabilities.

This distribution of projects and usage at NAS, as described above, justifies the characteristics of the computational model described below, which is based primarily on aerodynamics computations.

2.2 The Computational Model

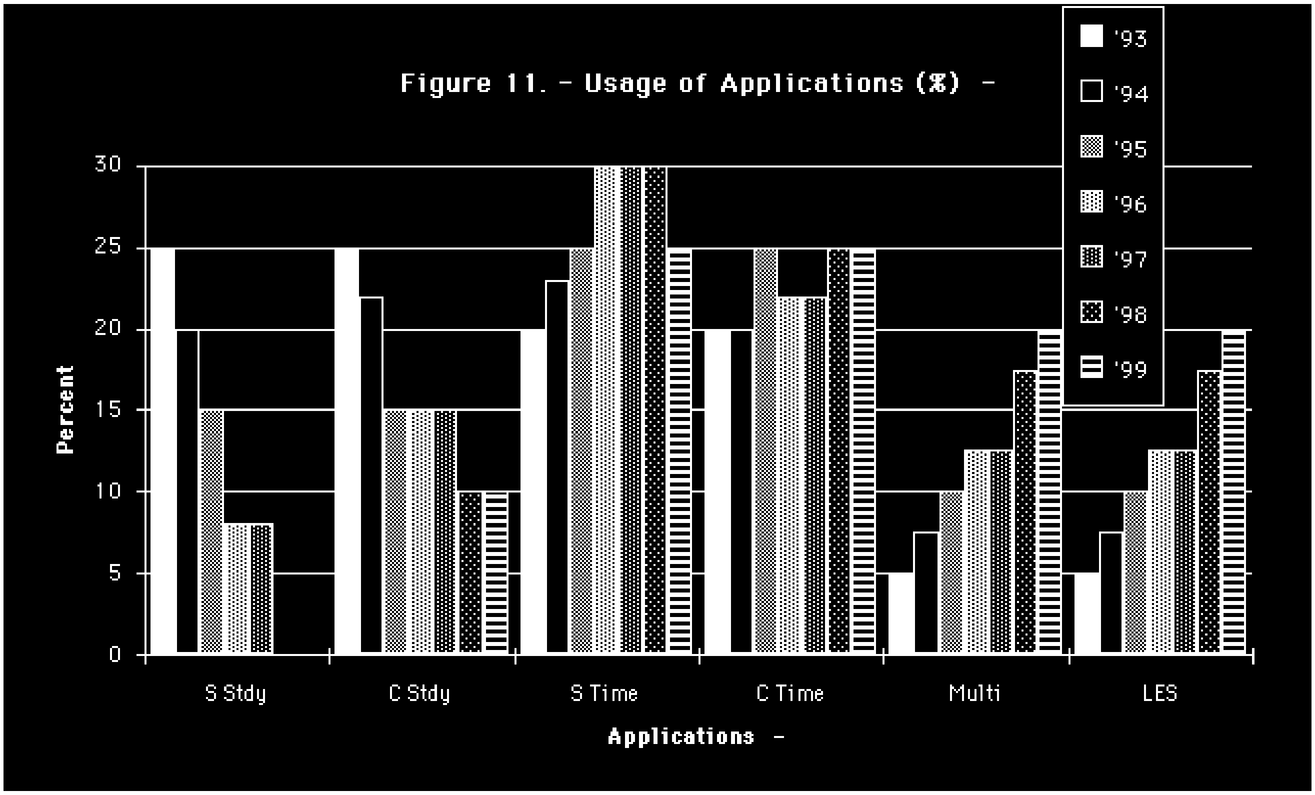
The model which approximates the workload at NAS (Ref.4), is parameterized such that future projections about capacity and computational work can be made. This model enables us to ascertain that the various components are in balance with each other, e.g., that the amount of data produced by the compute engines can indeed be stored on the mass storage devices available, and that the network can support the traffic generated by moving that data.

It should be noted that the model attempts to define the high-end large computational jobs to be executed on the NAS supercomputer system(s). It is known that smaller, shorter job are being submitted to the current systems. These are not taken into account for two reasons: 1. The mission of NAS is to service the very large flow jobs. 2. With the very impressive advances seen with the high-end workstations, more and more of these smaller runs will be migrated away from the supercomputer.

To define the workload, six categories of applications have been identified:

- Steady State—simple
- Steady State—complex
- Time Accurate—simple
- Time Accurate—complex
- Multi-Disciplinary
- Large Eddy Simulations (LES).

The vast majority of processing done at NAS falls into one of these categories. Each of them has a percentage of the total resources allocated to it, as prescribed by representatives of the user community. The allocation of resources to the six applications classes is shown in Fig. 11.



As time goes by, the fraction of the complex problems, multi-disciplinary, and LES increases. To quantify the time needed for solution, the amount of computation is determined by an estimate of the number of grid points for each class of problems, the number of computations per point, and the number of iterations to reach a solution. That is, the number of operations for a problem (NOP) is:

$$NOP = NP * Ops * NIT \qquad \text{(EQ 1)}$$

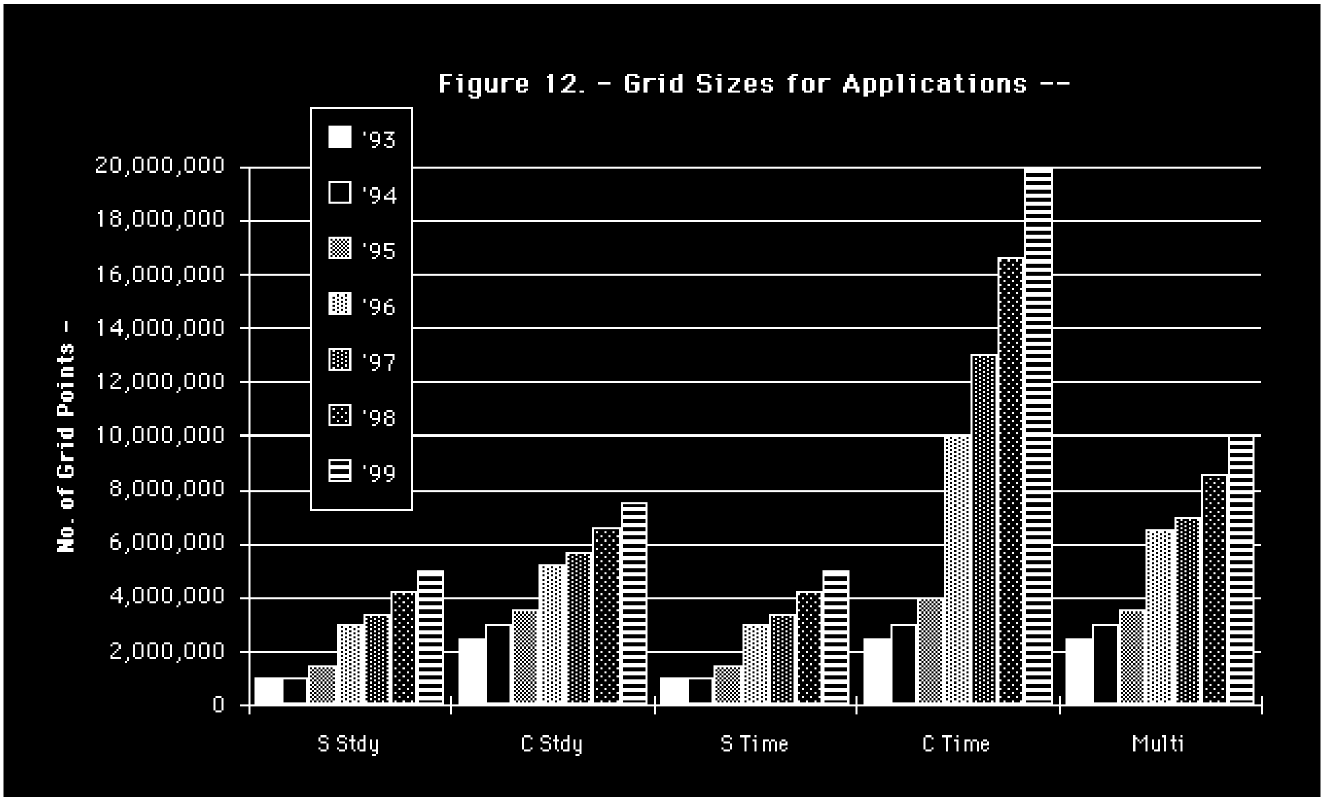
where:	NP = Number of grid Points for the problem
Ops = Number of Operations per point (per iteration), and
NIT = Number of ITerations for the problem.

The size of the problems within each category also increases in time from year to year. Figure 12 shows the grid sizes as they evolve in time for the application types, with the exception of LES (see below). A 4-5 times increase in grid size over the next six years would be typical. A number of values are associated with each grid point of the application. These are the values carrying the physical and dynamical properties of the fluid, as well as those related to the geometry and the programming techniques used. It is assumed that the number of variables per point is the same for all the applications types, and it increases in time linearly from 40 in ‘93 to 100 in ‘99. With this information the memory requirements for the applications (Mem) can be easily computed:

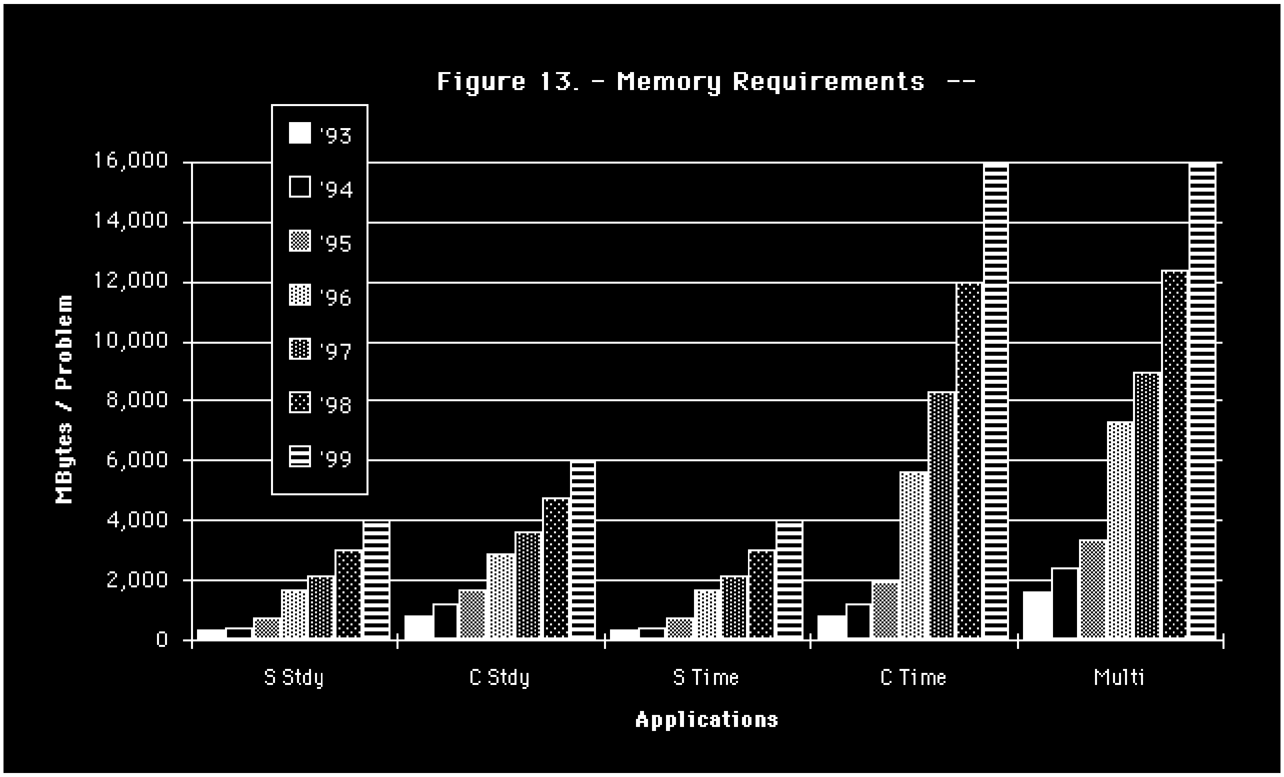
$$Mem = NP * NV \qquad \text{(EQ 2)}$$

where,

NV = Number of Variables per grid point



The projections of the memory requirements are summarized in Fig. 13 (again, without LES). These memory sizes are not very demanding compared to what is available on computer systems already, considering that NAS has 8 GBytes on the C-90 today. For the implications on overall system memory requirements see the discussion below.



The Large Eddy Simulations (LES) application is so much more demanding in resources that it had to be separated for reasons of presentation clarity. In addition, this is the only desired application that falls way above the 'TeraFlops Computing' requirement. Fig. 14 depicts the grid sizes and memory requirements for LES. They are an order of magnitude larger than those for the other five applications.

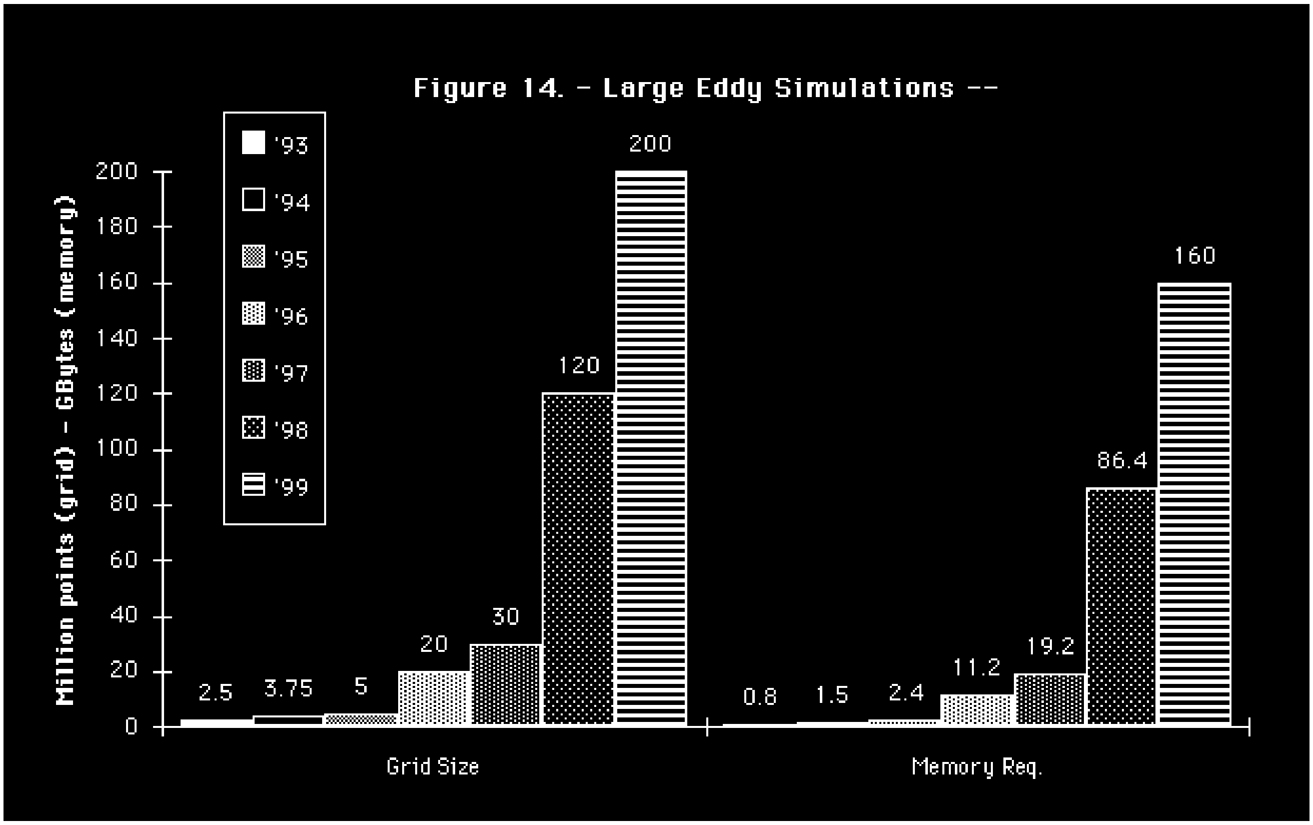
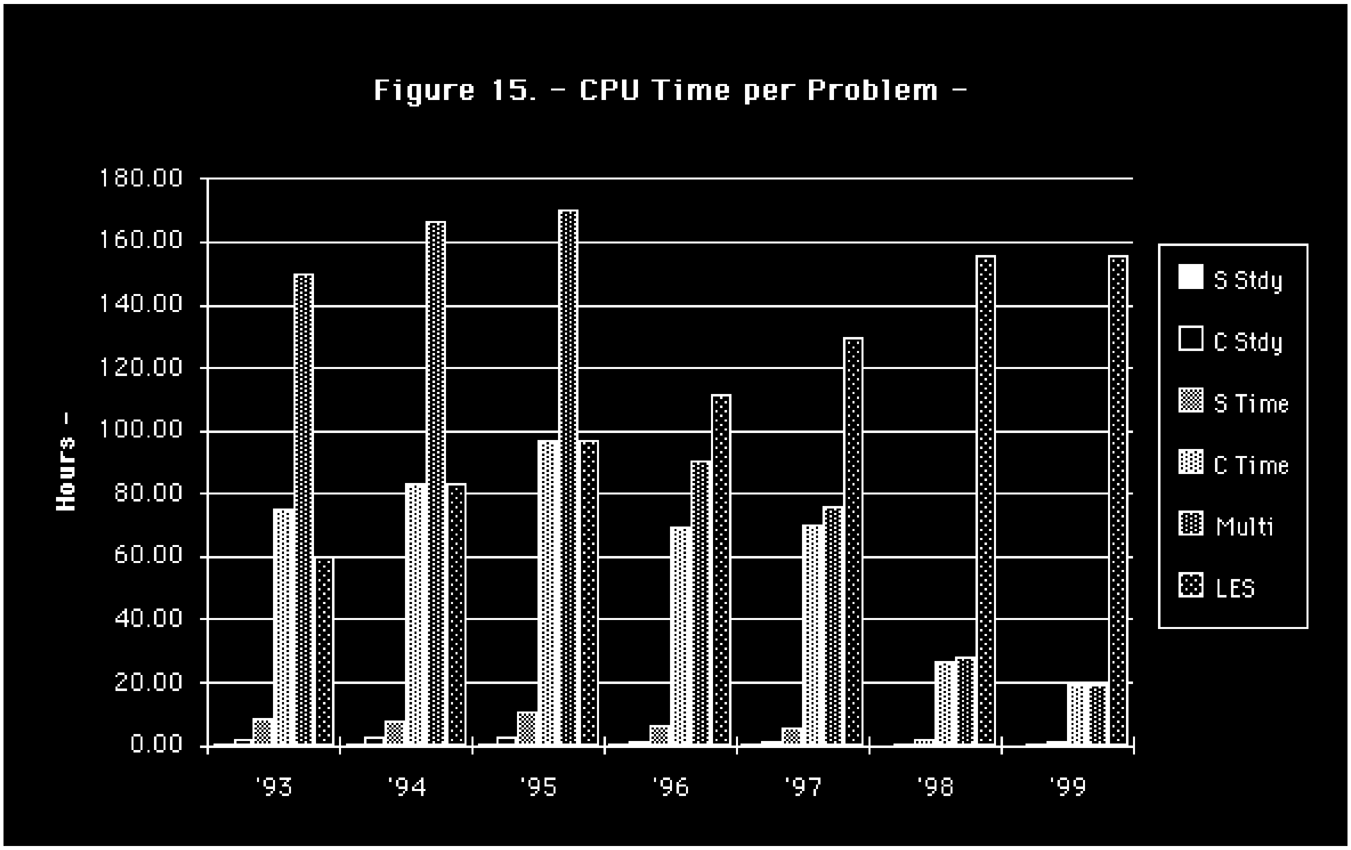


Figure 14. - Large Eddy Simulations --

Next, the projected sustained performance of the computing system is introduced into the model. It is assumed that all six applications will execute at about the same rate. It is now known that the current HSP, a 16-processor Cray C-90, performs at a sustained rate of about 3.5 GFlops (in '93). New systems will be put into production in early '96 and the middle of '98, with approximately 4 times increase in sustained performance in '96, and another factor of 5 in '98 (for a total of 20 times increase between '93 and '98). Once the number of operations for the applications (Eq. 1), and the performance rate, are known, the processing time may be determined. Fig. 15 shows the projected processing time the six applications would have over time if run in a dedicated mode on the system. These numbers range from a fraction of an hour to over 160 hours, or one week.

CPU time = NOP / SusPerfRate **(EQ 3)**



Of course, the HSP will have multiple users on the system at any one time. The number of users on the system was determined by how many jobs, of the mix given by application-class utilization (Fig. 11), will fit in the available memory (with allowance for swapping). Projected technology predicts that memory will increase from 8 GBytes to 256 GBytes between '93 and '98, as shown in Fig. 17. The elapsed time is then given by:

$$\text{Elapsed time} = \text{CPU time} * \text{NoUsers} \quad \textbf{(EQ 4)}$$

This is shown in Fig. 16. Considering that 660 hours represent a calender month, it is apparent that Multi-discipline and LES applications may take several months to complete.

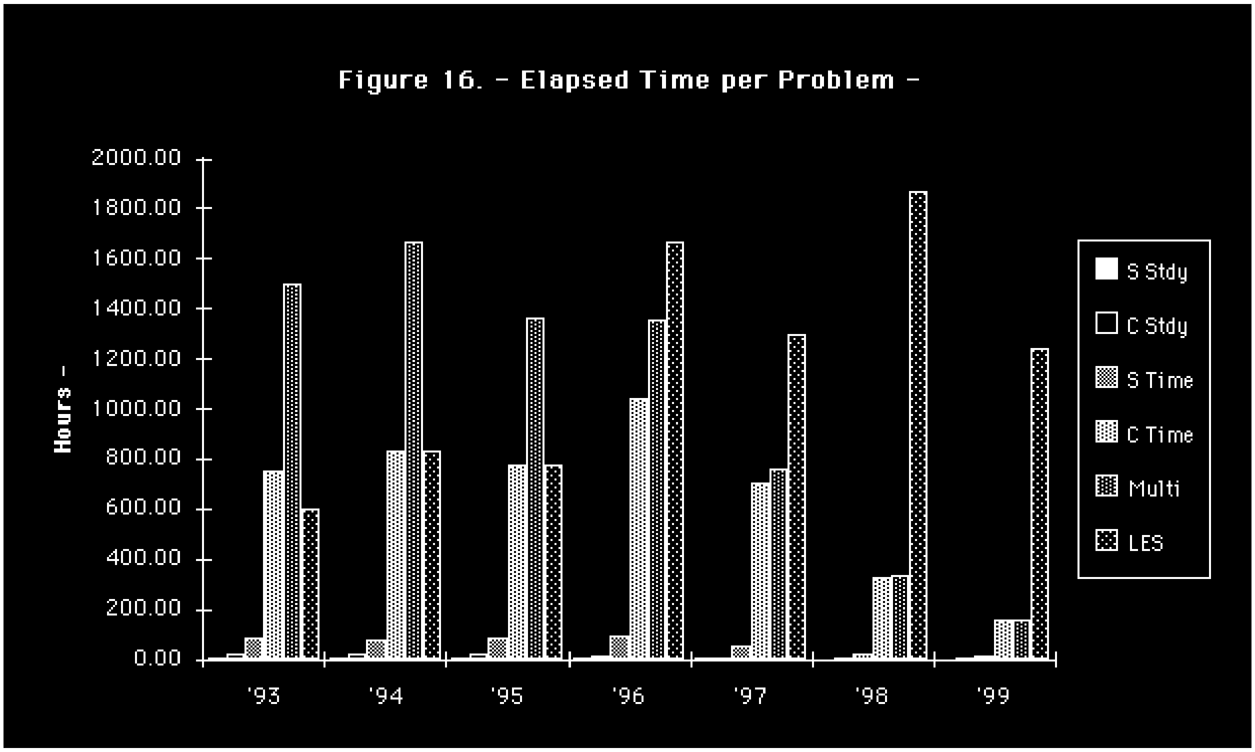
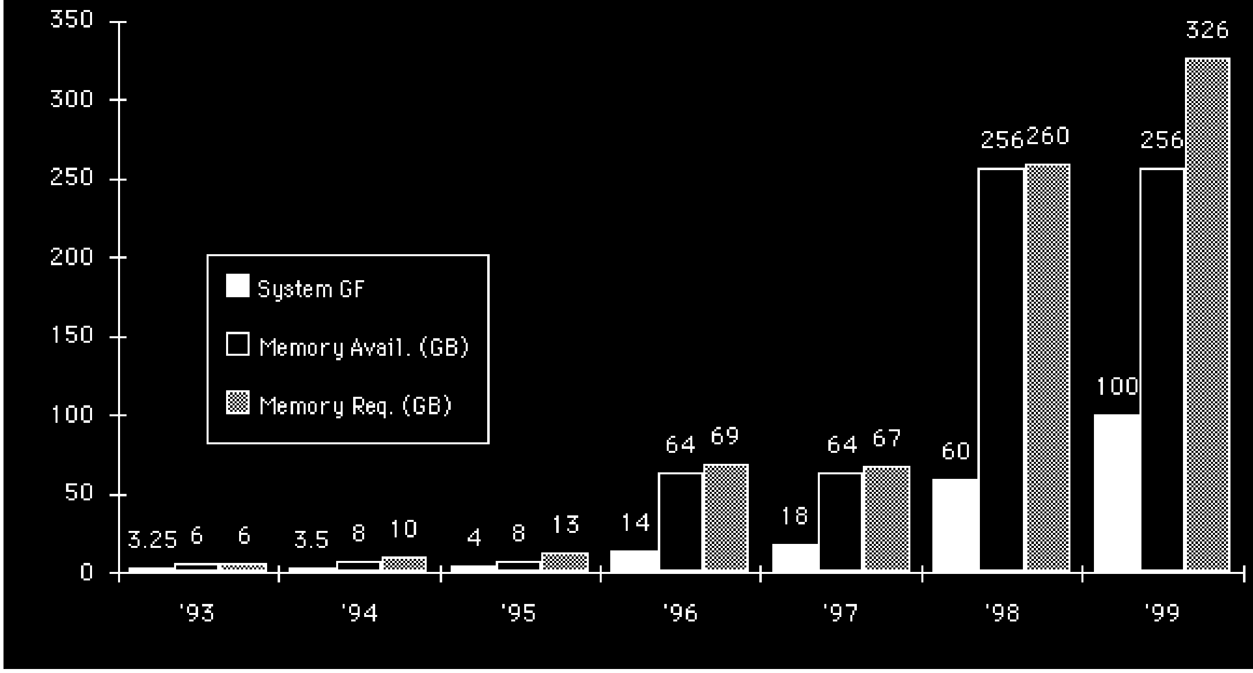


Figure 16. - Elapsed Time per Problem -

Fig. 17 presents the input to the model of the processing power and memory sizes available to NAS over time. It takes into account the introduction of a new system in the middle of a year, and a slight increase in effectiveness of a given system as time goes by. The projection is that a sustained performance of about 100 GFlops may be achieved before the turn of the century. This will happen through a much higher level of parallelism in the NAS production supercomputer (i.e., a system with more processors).

Figure 17. - System Performance and Memory -



The model also projects the disk storage requirements for the future. Of all the variables associated with each grid point, about 8 need to be stored for analysis and post processing. In addition, not all the iterations need to be saved. In fact, only between 0.2% and 2.5% of the thousands of iterations will be saved. The data itself will be required to be on-line for no more than 30 days, according to NAS users. The model computes the number of problems of each type that will be computed in a year, and the amount of data to be saved. With this information, Fig. 18 plots the projected disk storage required. As can be expected, the storage capacity needed increases roughly at the same rate as the computing power does: The projected storage requirements increase 35 fold from '93 to '99, while the computations rate multiplies by a factor of over 30 (from over 3 GFlops to about 100 GFlops). It should be noted that the number of iterations saved may turn out to be larger than stipulated in the model. The figures there were given by users at Ames, but there is some indication that the number of saved iterations is higher in some cases. The model's figures may still be close to average, though. Of course, the disk storage required is directly proportional to the number of iterations saved.

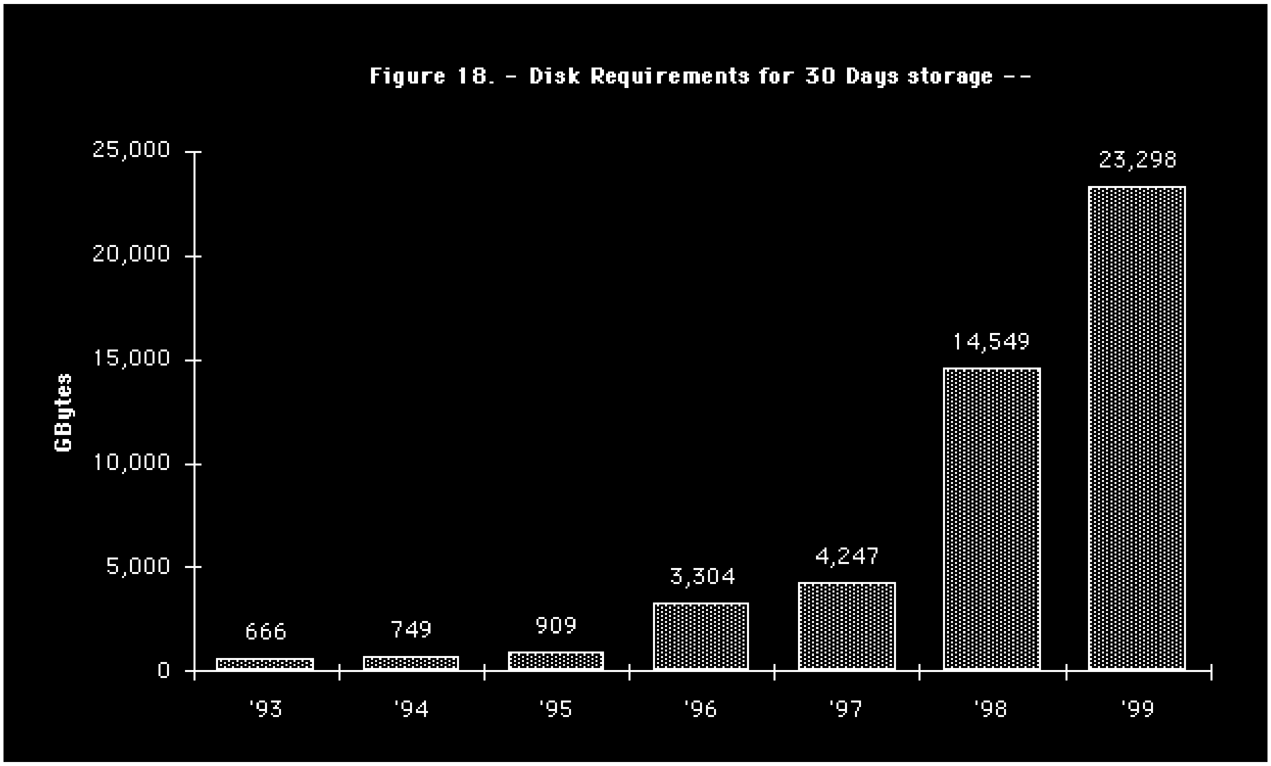


Figure 18. - Disk Requirements for 30 Days storage --

Ref. 4 contains more details about the model and its assumptions. Again, it models the workload which represents the NAS mission, not necessarily the cross section of jobs on the NAS systems at present.

3. Target System

3.1 Summary of Requirements

EOC-2's main driving force for the users is the introduction of the HSP-3 system, namely the C-90. The requirements for this system are detailed in the RFP for the HSP-3 (Ref. 6). The RFP called for a multi-processor system with at least 800 MFlops per processor and a total of no less than 12 GFlops (peak) for the system. The memory requirements were for 2 GBytes (extended to 4) of fast, common memory, and 50 GBytes (expandable to 450 GB) of directly accessible mass storage. Input/Output capability was required to be over 80 MBytes/second with interfaces to HiPPI, FDDI, and Ethernet; and with a capacity of 4-fold increase. The workload for this phase is simulated by a suite of benchmarks—the HSP-3 Application Programs, a set of 32 programs. They are required to perform at an aggregate level of no less than 3 GFlops in a multi-user environment.

The computation power of the system is closely related to the amount of data generated, which needs to be stored. A study based on past trends and growth, done for the NAStore project (Ref. 8), estimates that about 10 TeraBytes of data will be generated during '93 and '94. The amount of data generated about doubles every year. Most of the data needs to be stored. The storage media is partly on-line rotating disks, and partly a tape archival system. The disks provide fast access, but are less economical in terms of floor space and cost per unit of data. The way to determine the disk requirements is to relate it to a 'level of service,' i.e., specify how long a data file should reside on-line. The plan is for a level of about 30 days residence on rotating disks. With this requirement there is a need for about 600 GBytes of disk space in both '93 and '94. The remainder will be moved to a back-up system utilizing silos of magnetic tapes. For further discussion see Section 5.2.

The other main component necessary for a balanced system is the ability to move the data, or 'networking.' The network is divided into local and remote ('data network' and 'long haul communications,' respectively, in Figures 3 and 4). The local network uses several technologies, from ethernet to HiPPI, which provide different capabilities and features. They range from 10Mbits/second to 800 MBits/second (or, 100 MBytes/second), and are deemed sufficient for the EOC-2 period. The requirements for the Long Haul Communications (LHC) are driven by the need to transport data (rather than control or mail messages). The need has increased due to the increase in complexity and size of computations performed, and the increase in large simulations performed by remote users. For more detailed analysis see Ref. 5. New technologies will be added with enhanced bandwidth and reduced latency.

3.2 Technology Forecast

The details of the technology forecasts are given in the next sections for the respective products. The findings are summarized here.

For the HSP component, it is still not clear when the highly parallel systems will be suitable for production work. On the other hand, it is clear that those architectures have a better chance of eventually reaching multi TeraFlops performance levels. Given the maturity and the efficiency of the vector processors and the highly parallel systems, it is projected that either architecture will provide NAS with about 100 GigaFlops by the turn of the century. It is anticipated that, by that time, the HSP will be a system with several hundreds to thousands of processors, with hundreds of GigaBytes of memory.

Disk storage devices will continue with rapid improvements, such that their density will continue to about double every year. Similar advances will occur for magnetic tape devices.

Workstations enjoy huge improvements in speed and memory sizes. Some workstations have the capabilities of supercomputer processors of only two years ago. Significant work will soon be done by such desktops. This will impact the workload of the HSP (towards larger jobs), and the ability to use high performance visualization for the analysis phase of the computation cycle.

One of the areas advancing the fastest is long haul communications (LHC). National attention is focused on the ‘data super highway.’ This fits very nicely into NAS’ needs to support remote users.

All these projected developments, as described in each of the following sections, promise to provide the technologies required by NAS for a high performance and balanced computing system.

3.3 EOC-2 Components

Figure 4 depicts the high level components of EOC-2. The details are given in the following sections. Here is a summary of the components that have changed from EOC to EOC-2:

- Cray C-90 for HSP-3.
- Two Convex 3820s to support mass storage—NAStore II.
- Workstations (WKS III)—double the performance over the previous generation.
- Three new communications technologies—ATM, SMDS, and Frame Relay; Upgrade in LHC bandwidth—2 to 8 times.

A major software effort is associated with the switch over from the Amdahl to the Convex as the NAStore hardware platform.

3.4 Summary of EOC-2 capabilities

The processing power of the production supercomputer will have increased 4 fold, to close to 4 GFlops on daily average workload. With this, the memory available on this system will have quadrupled to 1 GWords. Basically, this system, the Cray C-90, will provide the user community with the ability to compute 4 times more than with EOC resources.

To keep the system balanced, the storage capacity will be upgraded accordingly. In line with the expected increase in production of data, the storage under NAStore will be 3-4 times larger than during the EOC period, increasing to about 1.6 TBytes. At the same time, the bandwidth of the major AEROnet paths will be increased by 3-6 times, with new technologies being introduced to guarantee fast and reliable data transfer to remote users.

4. High Speed Processor (HSP-3)

4.1 Introduction

The high speed processor is at the center of the NAS computing facility. Its capability provides the motivation for the remote NASA and Commercial users to continue as NAS users. It has traditionally been a vector supercomputer, and more recently, a multi-vector-processor shared memory system. This component, and the support functions around it, is what made NAS the unique facility it is. And, of course, NAS will remain so only if the most recent and biggest configurations will be acquired and made to provide a reliable production environment.

4.2 Technology

There have always been only a few viable vendors in the very high end computing system market. This is especially true if we consider only the U.S. market. In the past, during the early days of NAS, these were companies like Control Data and Cray Research. Later Cray Computer Corp. and Supercomputer Systems, Inc. became potential providers. The latter is already defunct, and the former has yet to complete a full-size system. The scope of providers for this type of architecture is shrinking, as is the size of the market. This is mainly due to the advances in PCs, workstation and graphics products, and also due to the introduction of highly, or massively, parallel systems. Our technology charts (see Ref. 7) show that for the next 5-6 years the vector shared memory systems will continue to increase in power at the rate of 4 times every 4 years or so. For example, the EOC machine, first shipped in '88 peaks at under 3 GFlops; for EOC-2 (1993) we expected over 15 GFlops (though only 12 GFlops was specified). Technology progress in memory parts offer comparable improvement factors in density; not so much in the speed of access which stays fairly flat.

In conjunction with the technology analysis it should be noted that the rate of improvement for highly parallel systems is higher, though their current performance, reliability, and stability are not up to production quality yet. However, for the next generation of high speed processors it is expected that those parallel systems may very well compete with the vector shared memory architectures.

4.3 Requirements

As was discussed before, there are CFD problems which can use any amount of processing power available (after a short period of algorithmic adjustments to new architectures, but without the need for 'breakthroughs'). So, in this instance we attempt to anticipate what can be expected from the vendors, and choose the point on the 'applications curve' which can be tackled. Once this is translated to GFlops and GBytes, the requirements for the other components of the system fall out fairly naturally.

The performance requirements for HSP-3 were defined in terms of computing times for a suite of application programs representing the workload at NAS. The summary of the detailed requirements was a system operating at a minimum of 3 GFlops, with 2 GBytes of memory (expandable to 4 GBytes), and 50 GBytes of high speed storage. In addition, it was required that each processor has a peak performance of at least 800 MFlops. Other requirements tested for features and functions of the Operating System, compilers, and utilities. These requirements were such that migration from the EOC machine was smooth, and advances in software tools are available to the NAS user community.

4.4 Plan and Implementation

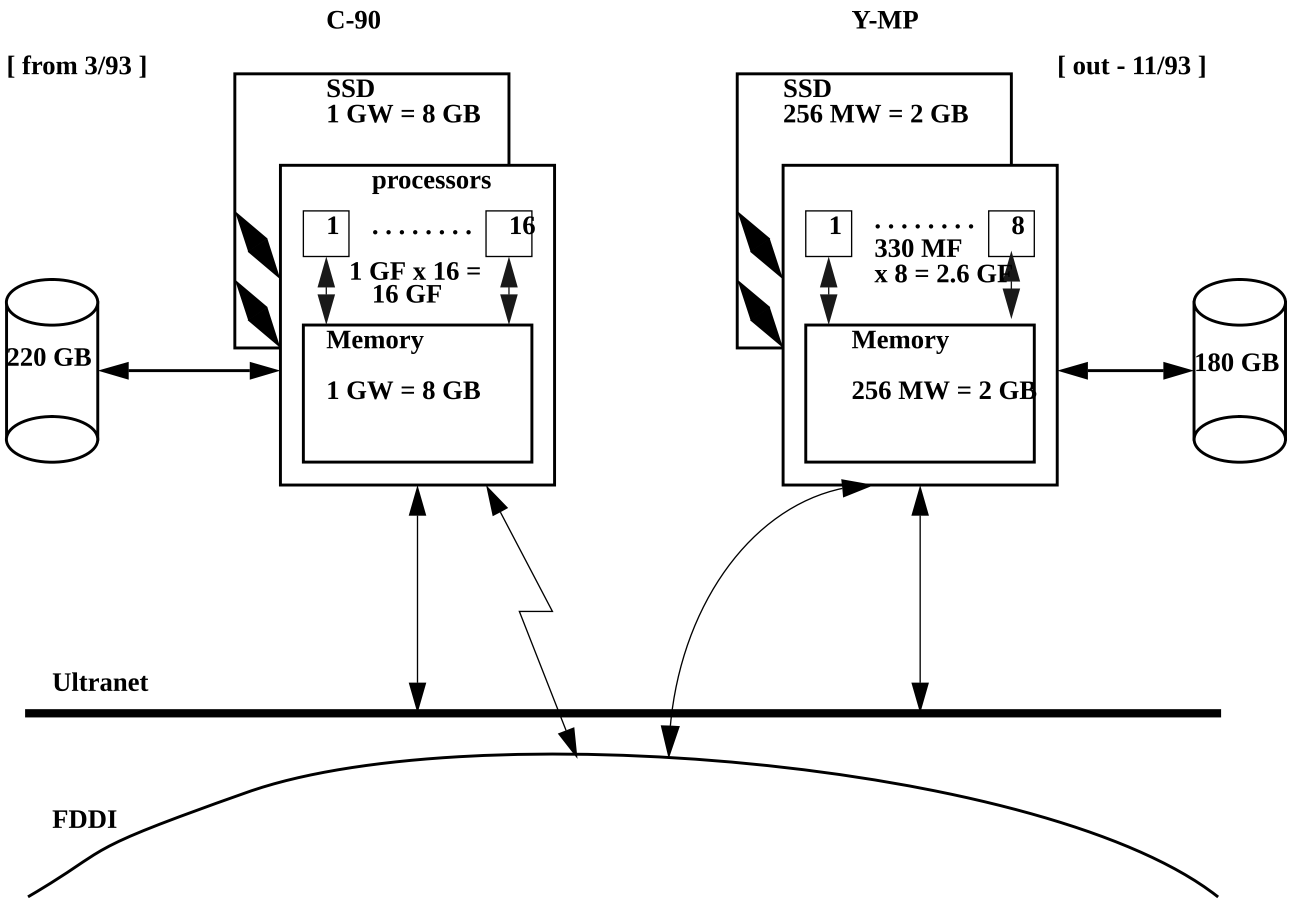
The definition of the requirements and the selection of the HSP-3 component has taken place prior to the EOC-2 period. The machine selected, the Cray C-90, was installed in March '93. The system has 16 processors running with a clock of 4.1 nanoseconds, and producing up to 4 results each clock period per processor. Thus, the peak of each processor is close to 1 GFlops, with the system's peak just under 16 GFlops. The C-90 was originally delivered with 256 MWords (2 GBytes) of memory.

The performance achieved was higher than required, or expected. The benchmarks perform with timings equivalent to an average speed of 5-6 GFlops; considerably higher than the 3 GFlops specified in the RFP. In fact, even the daily workload, composed of a general mix and often not well optimized, executes at a rate of 3.5-4 GFlops. This is a reliable system with stable, mature software delivering about 25% of its peak on a general workload. This should be contrasted to the less than 5% efficiency achieved by the highly parallel systems.

The memory was upgraded to 1 GWords (8 GBytes) in Oct. '93. Together with large memory there is also 1 GWords of SSD used for swapping and for staging files, and close to 220 GBytes of high speed disks. The added memory has improved the turnaround time for jobs in the system. The disk system will be augmented by a RAID disk array.

The Cray-2, an EOC machine, was removed in Feb. ‘93. The HSP-2 system, the Cray YMP-8, was in service until the end of Oct. ‘93. Fig. 19 depicts the high level configuration diagram for these systems.

Fig. 19 - HSP-3 / Vector Systems



5. Mass Storage (NAStore)

5.1 Introduction

This section will cover the developments in storage devices in the NAS computing configuration. The main media are disks and tapes systems. The software involved is NAStore. The major projects for EOC-2 are:

1. The transition from an Amdahl processor to Convex as the processor managing NAStore.
2. Testbeds and development towards a distributed mass storage system.
3. Upgrade of the tape system from the 3480 tapes technology to the 3490E tapes, and investigation of the new D2 technology.

5.2 Requirements

The storage capacity has to increase to accommodate the amount of data produced by the more powerful computational engines. This has to be modified by the gradual changes in the applications characteristics— for more complex applications there will be more operations performed relative to the data stored. Another consideration is the length of time the data has to be on-line. This, in turn, is determined by the time it takes to perform the analysis. As the tools become more sophisticated and the visualization engines more robust and powerful, this time decreases. It has been determined that between 15 to 30 days are sufficient for the data to be on-line for analysis. After that it can be archived on a tape system.

A combination of analysis of the computational model and a study of data storage use at NAS (Refs. 4 and 8, respectively) leads us to expect about 10 TeraBytes of data will be produced during 1993-94. Of that, about 600 GigaBytes need to be on-line in order to provide the 30 days disk residency for the data for the purpose of analysis. This last figure needs to be padded for uneven data production, for system overhead, and for other system uses of storage devices (software development, testing etc.). It should be noted here that the computational model (see Section 2.2 and Fig. 18) predicts a slightly higher level (by 10-20% for ‘93 and ‘94) of required on-line disk storage then the study. This has to do with the model’s assumptions about the nature of the workload. The study based on historical data is expected to be more accurate for the next couple of years. The model, with its implicit assumptions on future changes in the computing environment, may well be more accurate for later years. Given that, the requirement for on-line disk storage has to be at the level of about 1 TeraBytes. The tape systems will need to accommodate the other 90% or so of the data. So far, it has not been determined what criteria to use for discarding archived data. The options are: (i) Keep all data forever. (ii) Discard data which has not been used for some ‘x’ years. (iii) Create categories such that the users can specify the period a given file needs to be kept.

5.3 Available Technologies

The disk storage technology still advances at a fast pace. The most significant aspect is that the density of the medium about doubles every year. This development trend has two benefits: The amount of data stored may be doubled without increasing the floor space occupied by the devices; and storing twice the data will cost roughly the same, as the price per unit tends to remain approximately constant. Present technology allows to store about 12.5 GBytes per square foot (50GBytes in a unit with dimensions of 2x2 feet). The 1994 devices are expected to provide about 25GBytes/sq.ft. (100 GBytes per unit).

Together with the increased density, the form factor of the basic unit will shrink from 8” to 5” by 1994, and to 3.5” shortly thereafter. A by-product, and a benefit, of the smaller form factor is that the latency of reading from, or writing to, a disk will be reduced from 20-25 msec at present, to 8-10 msec within the next couple of years. The inherent bandwidth of disk devices is not expected to change before the end of ‘94. IPI drives provide up to 5 MBytes/sec, and SCSI drives provide a (nominal) 20 MBytes/sec. Bandwidth can be any multiple of these figures through the application of striping, whereby files are spread across multiple devices and each can be accessed simultaneously via multiple ports. This technique will become more popular since it is a natural implementation for highly parallel systems when each processor can access a portion of the file containing the data relevant to it.

Magnetic tape technologies are expected to exhibit substantial advances for the next few years. Present tapes, 3480s technology, offer up to 200MBytes per unit. In ‘93 we already had the successor product, known as the 3490 (or, 3490E, a longer tape version), with much higher data densities. Depending on the ability to compress the data, one can store from 800 to 1600 MBytes on each unit. The follow on to the 3490s will be available around ‘95 with the very impressive capacity of up to 15GBytes.

In addition to the evolutionary progression in tape products, there is a new technology emerging. The implementation is known as ‘D2’, to be available early in ‘94 (for testing only, initially). Its density is much greater than that of the 3490E. A unit can store about 25GBytes. The drawback is that the technology allows for no more than about 500 passes on the data. The access is intrusive, and if the data is not temporary, it will need to be copied to new units before the medium deteriorates. The technology is still promising, however, and the next generation, ‘D3’, with the same density will allow for up to 5,000 passes on the data. With D3 coming out in ‘95, it will be a competitive technology to the ‘3490E’ successor products.

5.4 Plan

The mass storage plan can be described by grouping activities into three major projects:

1. **NAStore engine:**
The Amdahl 5880, which controlled NAStore, had to be replaced as a result of loss of data integrity. A Convex 3820 system was selected to replace the Amdahl. Two systems were delivered in the first quarter of '93. The software was converted and tested. A third Convex was installed later in '93. In addition, the storage capacity was increased in response to the data storage requirements. The plan was to have 800 GBytes of storage capacity before the end of '93. A second capacity upgrade is planned for mid '94—doubling the capacity to 1600 GBytes. This disk storage is additional to that attached directly to the Cray C-90. The requirements analysis shows this amount will be sufficient for on-line retention on disk of more than the 30 days required by the users. The bandwidth of each Convex is about 6 times higher than the 5 MBytes/sec achieved on the Amdahl.
2. **Next generation storage systems:**
Start the development of a distributed storage system. The first testbed will be a HiPPI attached RAID disk array to NPSN processors. A study of the Northwestern Adaptive RAID started in the second half of '93 and will end in '96. A second testbed will employ Maximum Strategy equipment to create a network file system. The first prototype arrived at the end of '93; the second prototype is due 9 months later. The results of these studies will be used to prepare for Storage Systems III. The RFP is scheduled for mid '94, with an award a year later, and installation in the second half of the EOC-3 phase.
3. **Tape System:**
The 3480 tapes will be replaced by 3490Es. Some units were delivered by StorageTek before the end of '93, and a second delivery is due in Q2 '94. When all is here and functional, the tape storage capacity of the eight StorageTek 4400 NearLine robots will be increased to 40-50 TBytes. Another activity will be the introduction of the D2 technology. An initial device arrived before the end of '93, and a second delivery is planned for after mid '94, when new robots will be included.

6. Workstations

6.1 Introduction

The main function of the workstations in the NAS context is to provide the required hardware and software interfaces for the users, which will enable them to make effective use of the compute-intensive engines at NAS. This will be achieved by staying current with the latest generation of workstations. The graphics capability of workstations is important in many instances, and is being measured in addition to other performance metrics.

Workstations enjoy the most impressive advances in capabilities among the different levels of computer systems—from PCs to vector supercomputers. This, in turn, affects the distribution of workload between the different platforms. In particular, it will free the supercomputers for, essentially, only large scale and 'grand challenge' levels of computational tasks.

6.2 Requirements

The next generation of workstations (referred to as 'WKS III') will be available for delivery in early '94. The requirement was defined for machines that will deliver over 20 MFlops for a computational benchmark —the LINPACK 100x100 test, and will have graphics capability of performing at the rate of 300,000 3D independent z-buffered Gouraud-shaded polygons per second. They would have 512 MBytes of memory, and capability for supporting 8 GBytes of local disk. The changeover will be over time. The specified requirements were for 10 new workstations each for LeRC and LaRC, and 20 at ARC.

6.3 Technology and Available Products

The capabilities of near term new workstations have improved even faster than anticipated. This is due mainly to advances in the power of microprocessors, but is also due to increased density of memory chips, which enable much larger memories and cashes. In addition, there are architectural features such as increased bandwidth to memory and cache, more registers, and multiple functional units. And, of course, more mature software to take advantage of the hardware capabilities.

The other refreshing aspect is the number of vendors offering competitive products. There are at least 5 solid vendors—SGI, HP, DEC, IBM, Sun—who offer high end workstations, with their new generation shipping in '94.

Early results and estimates indicate we may expect about double the performance specified by the requirements listed above for these next generation machines. Their peak performance is much higher, and will reach the 200-300 MFlops range. This is for one processor. In addition, there will be products offered with multiple processors. Whereas today there are desktop workstations with two processors, the new products will offer as many as 6 processors. Server machines may have over 20 processors (at least from one vendor).

Another future development is the workstations support of HiPPI capability. This will allow for direct attachment to high performance networks, which will provide for high bandwidth I/O, and enable utilization of large file systems.

6.4 Plan

The SEWP (Scientific-Engineering Workstation Procurement) mechanism will be used in place of issuing a single RFP. Most users will continue to use their current workstations (many of which with graphics capability are SGI's 4D 320 VGX with 2 processors at 30MHz). It is now expected that about 10 new workstations will be acquired in '94 (less than were originally requested and hoped for).

The main activity will be in investigating and benchmarking the new workstations. The benchmarking is divided between two areas—one is a 'cpu' benchmark— a flow solver, a typical NAS application; the other tests the graphics performance capability of the workstation— corresponding to the requirement mentioned earlier. It should be noted that a performance of about 6 MFlops on the flow solver is approximately equivalent to the required 20 MFlops on the LINPACK test. It is expected that benchmark result would be 10-12 MFlops on the new workstations. Hence the expectation of availability of products with double the required performance.

It is expected that all the results will be in for early '94 procurements. The results will be evaluated and verified. Some of the evaluations will be done with early models installed at NAS for short periods for this purpose.

In addition to continuation work, code builds will be performed on new software releases—including libraries, Explorer, Inventor, etc. Other work is tied to support of visualization—for example, video editing, high resolution printers.

Applications-related software development includes a GL interpreter (libglto), a desktop video capability, a user interface to the network routing database (HNMS), a graphical interface for PLOT3D, and an accounting package.

The accounting package and HNMS involve overall system administration, and need to be separated from other systems accessed by users. To this end it was decided to acquire a Database Server. The processing requirements will be satisfied by a Sun SPARCcenter 1000 with its 4 processors, 128 MBytes of memory, and a large amount of disk space.

7. Networks

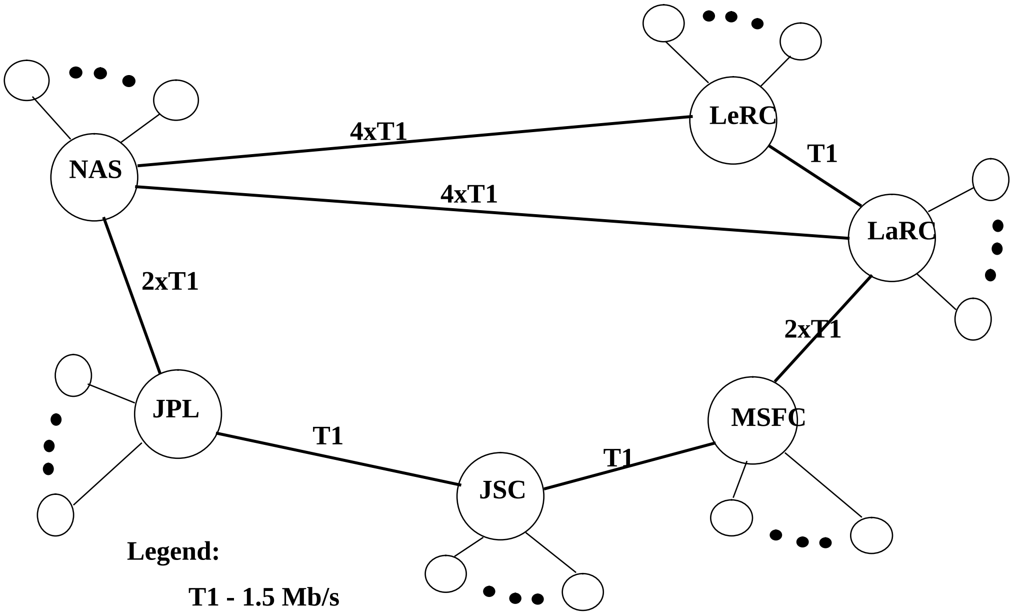
7.1 Introduction

The networks supported by NAS may be broadly divided into two major components—the local network, referred to as Data Network Services (DNS), and remote networks, or Long Haul Communications (LHC). It was initially planned that only the LHC would be changed in migrating from EOC to EOC-2. The upgrading of capability reflects the availability of new technologies, and the increased emphasis on large computations initiated from remote sites—both from NASA facilities and private aerospace industry users.

A need has arisen, however, for a modification in part of the local network. The local network is, in fact, composed of two networks—the high speed LAN composed of UltraNet, and the support network with FDDI ring and Ethernet connected to all the desktop workstations, PCs, and the AEROnet. These networks range in transfer rates from 800 MBits/sec to the Ethernet's 10 Mbits/sec. The UltraNet product for the high speed data transfer was found not to be sufficiently reliable for the NAS environment. Therefore, other technologies (e.g., HiPPI switches) are being evaluated, and one of them will replace the UltraNet. This change is scheduled to take place before the end of 1994.

The rest of this section will be concerned with Long Haul Communications. For details of present capacity see Fig. 20.

Fig. 20 - EOC -1 AEROnet



7.2 Requirements

The requirements for enhancing LHC are derived from the need to support Computational Aeroscience (CAS) simulations from remote sites. And in particular, the need to address the data transfer required to process the numerical output from such simulations. For analysis of these considerations see Ref. 5.

The resulting requirement must deal with the size of the files and the delay time that can be tolerated. It is convenient to think of three types of files which the remote user may be interested in:

1. the solution files—the numerical output;
2. the geometry files—needed for visualization and analysis;
3. the images themselves.

The latter two require delays of under 1 second, and for the file sizes required, this is beyond present technology (see Ref. 5 and below). Therefore, the plan calls for support of transporting the solution files, and relying on the remote sites to do the analysis and visualization at their sites. Though there may be exceptions to this procedure, this will be the norm.

With this model in mind, T3 capacity of 45Mbits/sec (to the main nodes) will accommodate the requirements for transmitting only the solution files (but not the geometry files or the images).

7.3 Technologies Considered

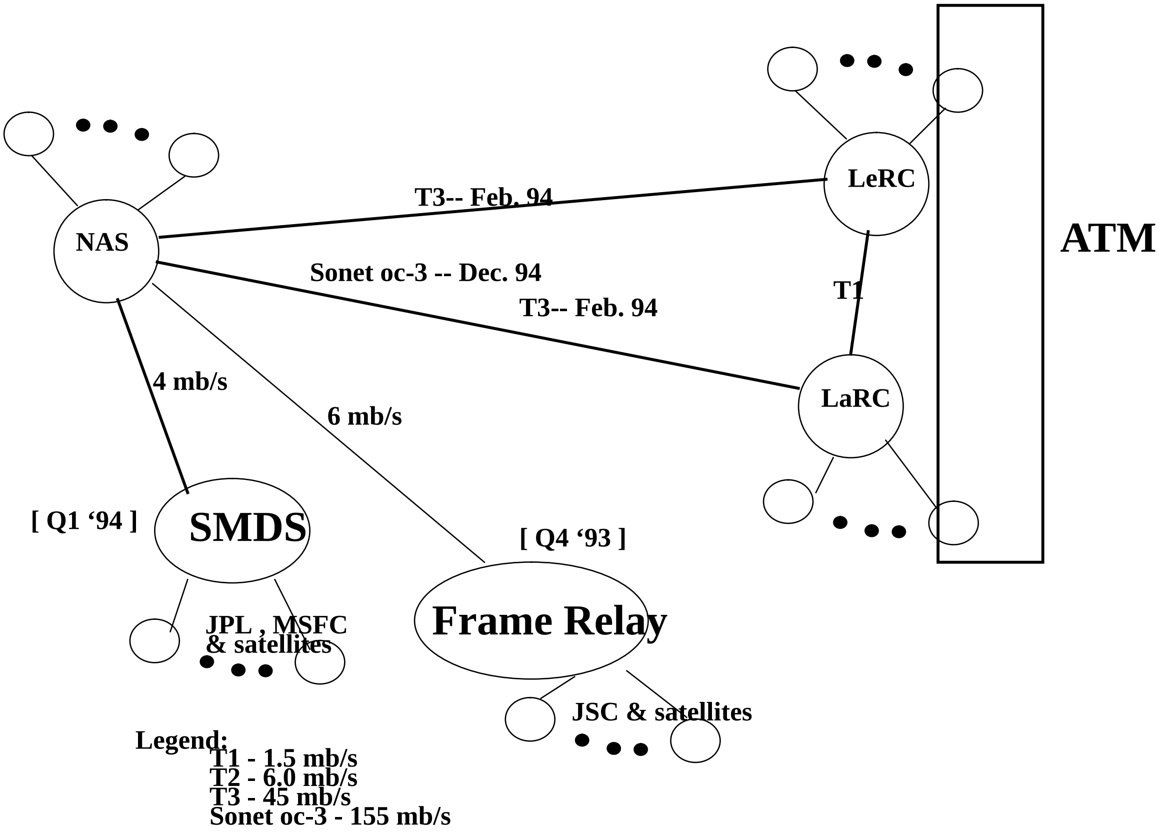
There are three technologies which are candidates for the next generation LHC. They are: (1) Asynchronous Transfer Mode (ATM), (2) Switched Multimegabit Data Service (SMDS), and (3) Frame Relay. They differ in features, capabilities, and protocols, and, therefore, have their own strengths and weaknesses.

It is not clear whether one or more of these three emerging technologies will dominate the market. Therefore, it has been decided to use all three technologies (each in a separate part of the network), gain experience, and perform detailed investigations on their use. We expect all three to have a use in EOC-2 AEROnet.

7.4 Plan

The three technologies mentioned above will be installed during the EOC-2 phase. Fig. 21 depicts the choices made for linking sites using these technologies.

Fig. 21 - EOC -2 AEROnet



The schedule for implementation is as follows:

1. ATM will be used for the connections to LaRC and LeRC. The link to NASA Langley and NASA Lewis will be in place with a T3 by Feb. '94. It will be upgraded to a Sonet OC-3 link by the end of '94.
2. SMDS is assigned for JPL and MSFC, and some of their satellite sites. The links to NAS will be a multiple T1 links, and are scheduled for Q1 '94.
3. Frame Relay to JSC was put in place in Q4 '93.

The capacities on these links represent the relative projected data traffic between the sites. The T3 links to LaRC and LeRC support up to 45 Mbits/sec rates; whereas the rates to JPL (SMDS), and JSC and MSC (Frame Relay) are 4 Mbits/sec and 6 Mbits/sec, respectively. It should be noted, however, that the requirements justify Sonet OC-3 level of service, but budget constraints allow only T3 levels until the end of '94.

8. Conclusions

This document describes the requirements and the design for EOC-2. But it also lays the foundation for future phases towards TeraFlops computing at NAS. It does so through connections between historical data, a computational model for future projections, and technology trends.

It has been shown that EOC-2 computing capacity about quadrupled compared with the previous stage (EOC). The increments in capabilities were designed so that the system would be balanced. There is sufficient memory to take advantage of the additional processing power of the processors. The mass storage has been upgraded to accommodate the increased amount of data generated. And the long haul communications is being enhanced to carry more data at faster rates. The process of remote execution of the CFD applications has been adapted to the available technologies; in particular, those of communications.

The computational model used here is consistent with the NAS mission, and with computing trends of shifting smaller-size applications to workstations. The plan presented here provides a solid basis for the design of EOC-3.

References:

[1] "Numerical Aerodynamic Simulation Program Plan", September 17, 1993.

[2] "NAS Systems Division Extended Operating Configuration (EOC) Design and Development Plan", PP-1124-00-N00, June 30 1987.

[3] "NPSN Monthly report", compiled by W. Kramer, September 1993 issue.

[4] "The NAS Computational Model", spreadsheets form based on previous work by D. Pase, revised and to be described in a technical report by D. Barkai.

[5] "Data Communications Requirements for CAS Applications", J. McCabe, RND-93-012, Aug. '93.

[6] RFP2-33570(RCB), "High Speed Processor 3 (HSP-3) Computer System", November 20, 1991.

[7] "Technology Charts" are maintained by RND (D. Barkai) for five product families—vector/shared memory systems, parallel systems, storage, communications, and workstations. They are changing frequently.

[8] Storage Requirements study, M. Tangney, 1992, unpublished.