General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Produced by the NASA Center for Aerospace Information (CASI)

Interim Report

STUDY OF SYSTEMS AND TECHNIQUES FOR DATA BASE MANAGEMENT

December 1976

▼▼TELEDYNE
BROWN ENGINEERING

Cummings Research Park • Huntsville, Alabama 35807

INTERIM REPORT
SD76-MSFC-2060

STUDY OF SYSTEMS AND TECHNIQUES

FOR

DATA BASE MANAGEMENT

.

December 1976

Prepared For

GEORGE C. MARSHALL SPACE FLIGHT CENTER
MARSHALL SPACE FLIGHT CENTER, ALABAMA

Prepared By

SYSTEMS DIVISION
TELEDYNE BROWN ENGINEERING
HUNTSVILLE, ALABAMA

ABSTRACT

This report presents the results of studies in a number of
data management areas with emphasis on the identification of issues
and problems that NASA data users will encounter and on emerging
technologies that will be available to these users. Specific areas
discussed include the identification of potential NASA data users other
than those normally discussed, consideration affecting the clustering
of minicomputers, low-cost computer system for information retrieval
and analysis, the testing of minicomputer-based data base management
systems, ongoing work related to the use of dedicated systems for
data base management, and the problems of data interchange among a
community of NASA data users. The number of subject areas covered
prevented an in-depth analysis of any one area. Thus an attempt was
made to identify pertinent problems and issues that will affect future
NASA data users in terms of performance and cost. A number of these areas
deserve additional study as the requirements associated with the NASA
Data management program are better defined.

Although interrelated in terms of their application to low-budget
NASA data users, the sections of the report are basically independent
and may be read individually without reference to previous sections.

APPROVED BY: APPROVED BY:

_____   _____
R. R. Parker   H. M. Barnard
Manager, Data Processing   Vice-President

TABLE OF CONTENTS

TABLE OF CONTENTS – Concluded

1. INTRODUCTION

Data base management technology encompasses a variety of disciplines that include the hardware, software, and procedural and/or protocol elements that enable the creation of an integrated data base from logical files and the subsequent retrieval of information from the data base using either specified keys and/or relationships. This report analyzes a number of specific areas of data base management technology that are of interest to NASA as part of the Office of Applications Data Management Program. Topics covered within the report are:

- Potential NASA Data Users

- Clusters of Minicomputers

- Low-Cost Computer Systems for Information Retrieval and Analysis

- Testing Minicomputer-Based Data Base Management Systems (DBMS)

- Use of Dedicated DBMS Processors

- Data Interchange Among a Community of Users.

Related subjects that have been covered in separate reports include:

- Interface Standards for Computer Equipment

- Design Requirements for a Programmable Data Communications Controller

- Investigation of Disk Systems

- Survey and Capabilities Projections for SEASAT User Data Systems

- DMS 1100 Test and Evaluation

- NASA Data Users Requirements for Processing Equipment.

The principal effort to this point in the program has been a survey of the state of the art in data base management technology, an

identification of the key technology areas that require assessment, and, to a limited degree, the identification of the key issues/phenomenology that provide the basis for any future technology assessment efforts. The results of this work will form the basis for technology forecasts and the subsequent analysis of applications and consequences in a number of technology areas associated with data base management systems.

Each section of the report is basically independent of the other sections. Thus the reader may read only those sections of interest.

2. POTENTIAL NASA DATA USERS

The programs that are under the auspices of the NASA Office of Applications appeal to a wide variety of users, ranging from Federal Government departments and agencies (Departments of Interior and Agriculture, National Weather Service, etc.) to state and local governments, local and regional councils of governments, planning commissions, national and international organizations, companies, universities, secondary schools, and individuals. In addition, similar government departments and ministries of foreign countries, foreign universities, foreign organizations, companies, and individuals have need for such data as they relate to their part of the world.

This section of the study examines those situations where there appears to be a use for available NASA data and a capability exists for transfer of the data to the potential user, but a specific requirement has not been identified. The results are presented in terms of the user interface, identification of potential users, and some specific considerations affecting the distribution of NASA data.

2.1   CAPABILITIES FOR TRANSFER OF NASA DATA

A variety of methods exist for providing data to potential NASA data users, including the mail and telecommunications links. Different users will have a diversity of requirements for data types, data processing, data timeliness, etc. Many of these requirements will be a function of the user's capability to handle the data, which is most probably a function of economic considerations. Thus secondary schools, individuals, etc., may require final processed data via the least expensive distribution method, which is probably the mail. Whether this requirement to process data is placed on NASA or on another organization, such as the Department of Agriculture, is not of concern in this study, however.

Data technology is to the point where it is technically feasible
to provide data electronically to virtually any point in the United
States via existing telecommunications systems. Similarly, on a worldwide
basis, practically every major country in the world is interconnected
via the INTELSAT Satellite Network, and at least that portion of the
population located near the INTELSAT Earth Terminals have the technical
potential for acquiring data via this network.

Although the technical feasibility for acquiring and using data
exists, the overriding considerations for many potential users are
economic limitations. Economic factors become even more of a limitation
when using telecommunications facilities to transfer data because
of the recurring costs involved. Thus, although it is technically
feasible to transfer data almost anywhere in the world where there is
a potential user, it is not always economically feasible for low-budget
us...s, especially if they are located outside the United States.

2.2   USER IDENTIFICATION

Investigations were conducted to determine potential users of
NASA data who have both the need and the technical capability of receiving
and processing the data but who have not expressed a requirement for such
data. The classes of users considered were government, government groups
(regional planning councils, etc.), universities, national and international
organizations, companies, and individuals. The conclusion was reached
that users within the first two categories (government and government
groups) have been fairly well identified within the United States. A com-
prehensive study from the standpoint of classifying potential users within
these categories was conducted for the Goddard Space Flight Center by
the Center for Development Technology of Washington Universtity. This
study identified the potential government and government group users
within a five-state area of the Midwest. Although the study was restricted
geographically, the results were comprehensive and representative of the
types of organizations that would have need for NASA data in other parts
of the United States. Also, similar organizations almost surely exist

within friendly foreign countries around the world, and these organizations would have need for NASA data as they relate to their geographic areas of interest.

2.2.1 Universities and Secondary Schools

Large and medium-size universities throughout the Unites States have been using NASA data for years. The data are available as both research tools and teaching aids. Limited types and amounts of data have also been available to secondary schools. With the continuing decreases in the cost of data processing equipment and the continuing improvements in telecommunications capabilities via the Direct Distance Dial (DDD) network, it appears that there will be increased demands for data from universities and secondary schools.

The largest increases in demands by universities will likely come from the research centers within the universities. Practically every large-to-medium size university has a number of research centers in at least a few of the following areas: agriculture, water resources, mining and minerals, environment, energy, population planning, marine life, forestry, natural resources, and pesticides. As the cost of processing data decreases, each of these research centers will acquire their own processing and communications capability and demands for data will increase.

2.2.2 Independent Research Center

In addition to university research centers, a number of company-owned and independent research centers will place new demands on NASA for data. A comprehensive listing of university, company-owned, and independent research centers can be found in a book entitled Research Center Directory by the Gale Research Company.

2.2.3 National and International Associations and Organizations

Potential users within the list of national and international organizations are almost limitless. The libraries are full of listings of such organizations. Potential user categories are:

- Agricultural and Food Associations

- Commodity and Trade Organizations

- Conservation Associations

- Banks and Banking Organizations

  ▲ International Bank for Economic Cooperation

- Forestry Commissions

- Marine Organizations

- United Nations and Associated Organizations

  ▲ Economic and Social Council
  ▲ Economic Commission for Africa
  ▲ Economic Commission for Asia and Far East
  ▲ Economic Commission for Europe
  ▲ Economic Commission for Latin American
  ▲ Food and Agriculture Organization
  ▲ World Bank
  ▲ International Development Association
  ▲ United Nations Children's Fund
  ▲ World Food Program
  ▲ World Meteorological Organization

- International Government Organizations

  ▲ European Common Market
  ▲ Inter-American Municipal Organization
  ▲ Inter-American Planning Society
  ▲ Inter-American Program for Urban and Regional Planning
  ▲ Organization of American States

- International Relief Organizations

  ▲ Red Cross

- Free World Communication Organizations

  ▲ Radio Free Europe.

2.2.4 National and International Companies

National and international companies represent one of the major groups for potential use of NASA data. A survey of the nation's largest

businesses shows that the majority of the largest companies have use for
NASA data. The following list categorizes companies in terms of their
output product and/or service and suggests one or more uses for NASA
data for each category:

- Industrial Companies

    ▲ Oil

        – Exploration

    ▲ Automobile

        – Pollution monitoring

    ▲ Steel

        – Exploration

    ▲ Chemical

        – Pollution monitoring

    ▲ Farm Equipment

        – Agricultural distribution

    ▲ Heavy Machinery

        – Construction and land use

    ▲ Food Production

        – Agriculture

    ▲ Paper

        – Forestry

    ▲ Lumber

        – Forestry

- Nonindustrial Companies

    ▲ Banking

        – Land use/urbanization

▲ Life Insurance

   − Environmental/urbanization

▲ Diversified Financial

   − Land use/urbanization

▲ Retailing

   − Urbanization

▲ Transportation

   − Urbanization and agriculture

▲ Utilities

   − Urbanization and water availability.

2.3   CONSIDERATIONS AFFECTING DATA DISTRIBUTION

The previous section only began to identify potential users of NASA data. A more comprehensive analysis would pinpoint specific users, requirements for specific data, requirements for various levels of processing, and numerous other factors. Also, a detailed analysis is required to determine precisely what would be the most effective mechanism for transferring the data to the users. In particular, where should the data be processed, how timely must the outputs be to be useful, and what level of processing will be provided by the Government.

Many of the users will want only final processed, summary results. Others may want raw data that they will process in-house. Although the latter appears to be the approach to data distribution that NASA favors, it does not appear to be making use of the data in a manner that provides equal access to all potential users. During the research for this study, it became obvious that thousands of potential users have limited capabilities and budgets. On the other hand, a number of larger organizations (e.g., commodity trade organizations) have the capability to acquire NASA data and use it to their own advantage at the expense of other organizations and the general public. Because of the importance of much of the NASA

data and because it is public property, there may be overriding con-
siderations, such as those above, that would encourage the Government
(either NASA or other user organizations) to provide processed data on
an equal basis to all potential users. Such an approach would have a
tremendous impact because of the volume of data involved. The rationale
for such an approach and the effects of the approach require further study.

3. CONSIDERATIONS AFFECTING THE CLUSTERING OF MINICOMPUTERS

Minicomputer clustering is technically feasible, and in fact several minicomputer clusters are currently in operation for a variety of applications, including the replacement of large systems. The appropriateness of such arrangements for particular applications and/or environments, however, is a substantially more complex question, and some of the considerations affecting such a decision are discussed herein.

This section examines the feasibility of using clusters of minicomputers to replace large systems such as IBM 370s for space data processing. Clustering is examined from the standpoint of hardware configurations, resource allocation and scheduling, applications, performance, and cost.

3.1   HARDWARE CONFIGURATIONS FOR CLUSTERING

Interfacing of processors is basically a matter of providing facilities for communication and resource sharing. A computer cluster can include any combination of the interface options illustrated in Figure 3-1 and discussed below.

3.1.1 Shared Memory

Two or more processors can share a single bank of memory. This scheme offers the fastest possible buffered communication between processors since data transfer is limited only by the memory access time. In general, when one processor is accessing a memory bank, that bank is locked out for access by the other processors. This scheme is particularly applicable for common control tables that multiple processors must access frequently. Programs and/or data may also be stored in this common area so that jobs serviced by the various processors all work from the same version of the data and program. A particularly appropriate program for this shared area would be any control logic that all processors need to use. Currently, clusters of Interdata 8/32 minicomputers
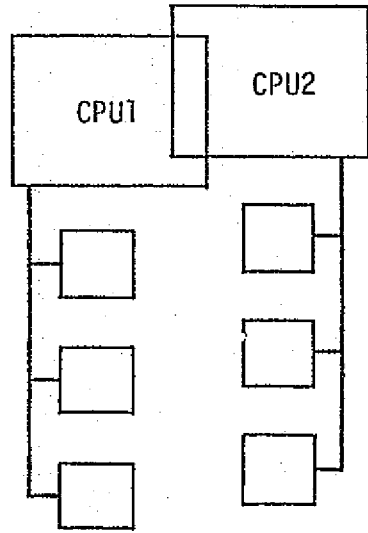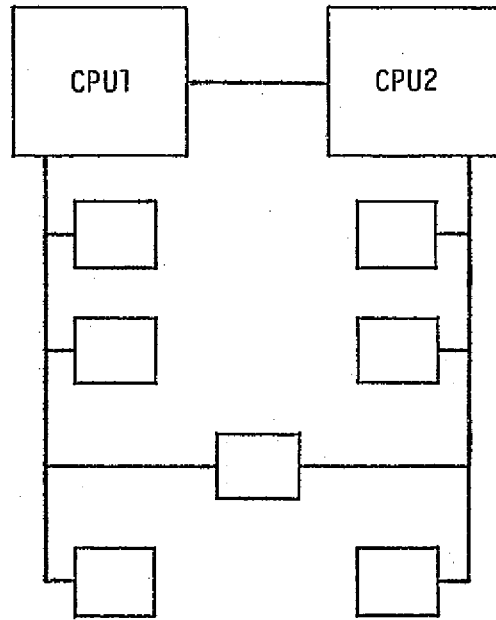
SHARED MEMORY   SHARED PERIPHERALS   COMMUNICATION LINK
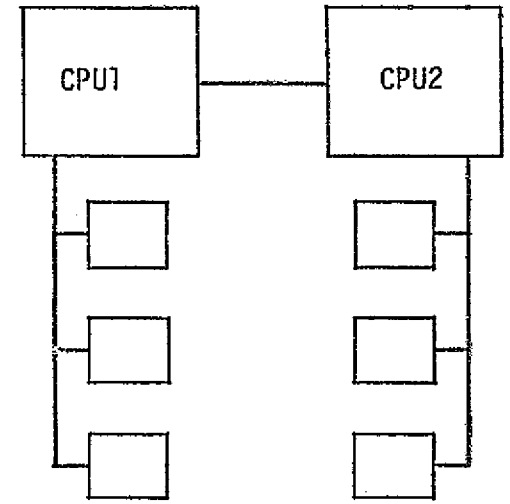
FIGURE 3-1. INTERFACE OPTIONS FOR CLUSTERED COMPUTERS

are in operation with shared memory capabilities. Up to 14 processors may share a single memory bank on the Interdata System. One NASA installation that uses this approach is NASA/JSC, where either four or five Interdata 8/32 minicomputers operate in a shared memory configuration for Space Shuttle simulation.

3.1.2 Shared Peripherals

Clustered processors may share peripherals, either as a communication medium or to increase the utilization of these peripherals. As a rule, however, the only effective peripherals for communication purposes are on-line mass storage devices. Virtually all vendors, including minicomputer vendors, supply disk systems that are at least dual ported. Thus these disks can be used as buffer areas for inter-processor communications, as well as a shared area for common programs, data, and control tables. Logically, this scheme is identical to the use of shared memory. From a processing point of view, use of a shared disk introduces the added software overhead for device handlers to access common storage. The shared use of other peripherals, such as card readers, punches, plotters, and printers, is technically feasible, but in general would require the development of special interfaces to accept multiple I/O lines.

3.1.3 Communication Links

A variety of high-speed processor-to-processor communication interfaces are currently available and in use, including serial and parallel interfaces that operate under program control and in the DMA mode. In fact, the current Atmospheric and Oceanographic Information Processing System (AOIPS) configuration at Goddard Space Flight Center uses a pair of asychronous serial interfaces to interconnect a PDP-11/70 and a PDP-11/45. Using this type of interface, each processor appears as a peripheral to the other processors for purposes of communication. Recently announced minicomputer products provide for DMA/DMA interfaces to achieve maximum transfer speed and minimum CPU overhead.

3.2   RESOURCE ALLOCATION/SCHEDULING

To take advantage of these hardware capabilities for communication and resource sharing, the collective use of resources must be carefully scheduled. Usually this scheduling is by system software, although operator scheduling is sometimes performed. The resources required for a job can either be assigned to the job itself or to a particular processor. When resources are assigned to a particular processor, a job must execute only on the processor to which it is initially assigned. If resources are allocated to a job, various steps of that job can logically be executed by various processors. The four major categories of scheduling algorithms are illustrated in Figure 3-2 and are discussed in subsequent paragraphs.

Current implementations of 32-bit minicomputer operating systems do not address the problem of scheduling processors in multiprocessor configurations. Such processors have been configured into clusters, but system control/scheduling is handled by special-purpose applications programs.

3.2.1 Logically Separate Processors

In this configuration, processors in a cluster are logically subdivided into two or more separate systems, each with one processor, some main memory, and peripheral devices. Logically, it is like having separate systems in close proximity. Although multiple processors may physically be capable of accessing a common resource, concurrent use of resources is not attempted. With such separate systems, there is no communication between systems for job scheduling. All scheduling and system reconfiguration is performed manually by the operator.

The primary advantage of this approach is that processors, memories, and I/O devices can easily be reconfigured to yield particular systems needed for special applications. For example, this configuration might be used to meet infrequent needs for a processor with a particularly large memory. Another advantage of this configuration is the inherent
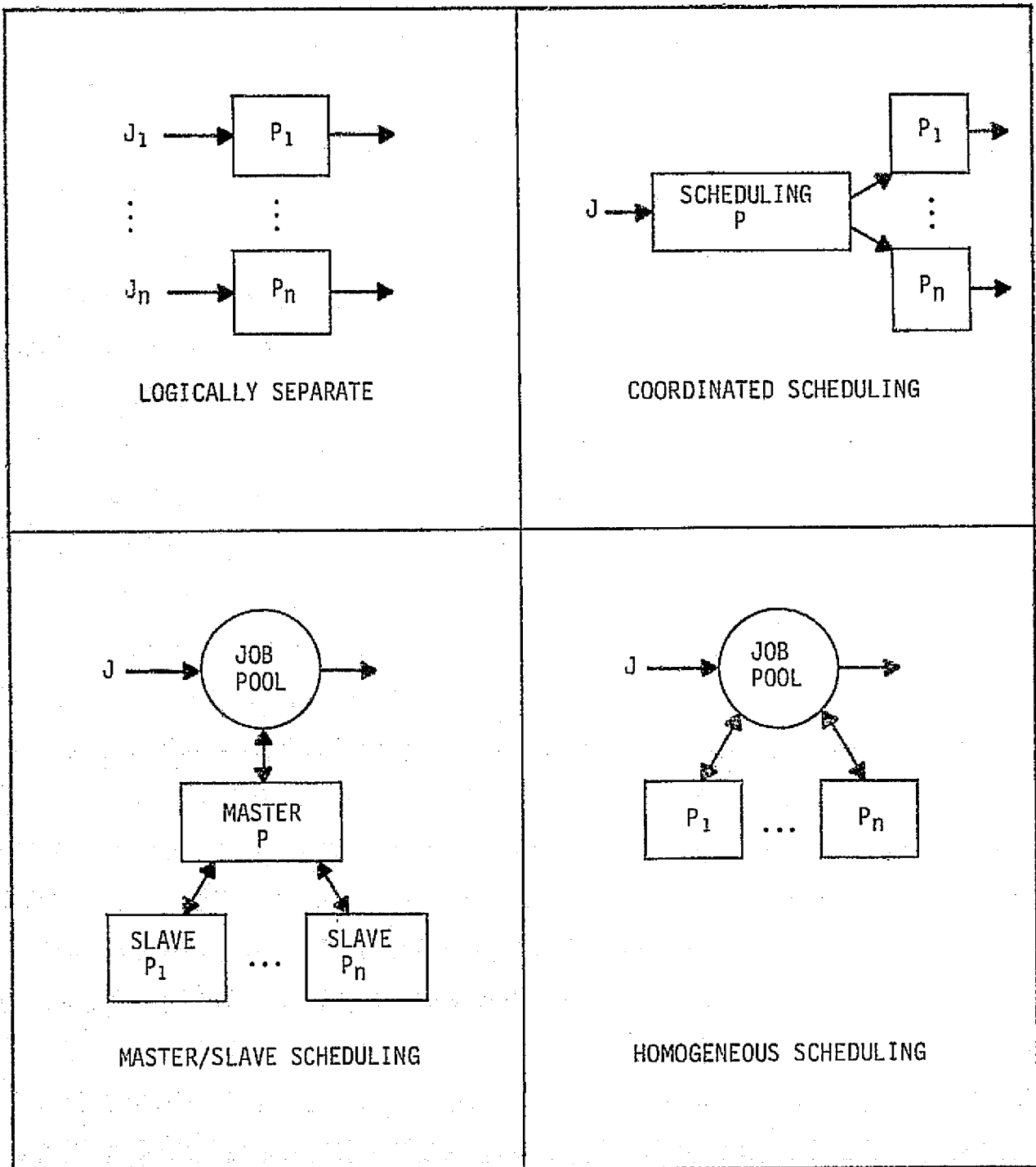
FIGURE 3-2. CENTRAL SCHEDULING OPTIONS FOR CLUSTERED COMPUTERS

equipment back-up capabilities it affords. This scheduling and control scheme has been in use for some time now. The IBM System/360, Model 67, for example, can be logically subdivided into separate systems.

3.2.2 Coordinated Job Scheduling

Coordinated scheduling (also called loosely compled multiprocessing) is similar to the logically separate scheme described above in that each processor is associated with a separate set of resources and peripherals. Similarly, jobs are assigned to a processor and remain with the processor to completion. The distinction between this scheme and the logically separate scheme, however, is that software may be used to assign jobs to a processor based on some priority scheme, such as the lightest load. This scheduling software can be implemented either on a special-purpose processor (e.g., the Octopus System at Lawerence Radiation Laboratory) or by one of the basic system processors (e.g., IBM OS/VS-2 Job Entry System).

3.2.3 Master/Slave Scheduling

Using master/slave scheduling, one processor monitors the status of all jobs and processors in the system and schedules the work of slave processors. Called tightly compled multiprocessing, resources can be assigned to jobs, and once blocked a job can later be resumed by another processor. This scheme is much more effective in the short-term balancing of activity among processors than the schemes described above, where a job must remain with a single processor until completed. The drawback to this scheme is that all scheduling must be handled by a single processor, which under certain circumstances could become a bottleneck in the system.

3.2.4 Homogeneous Scheduling

Homogeneous scheduling, also called floating executive, refers to the scheme in which all processors are capable of scheduling their own activity. All processors have access to job processor status tables and can either use the same or unique algorithms for selecting their next

3-6

activity. Thus a processor is not dedicated to any particular duties
and no processor will become overloaded. A significant implication of
this scheme is the need for providing processor lock-out protection from
the common job-scheduling tables. It is necessary that only one pro-
cessor be capable of accessing and/or updating these tables at any one
time. Otherwise, more than one processor may attempt to process a
single job step or a job may be skipped. Such schemes for software lock-
out exist but do add slightly to system overhead.

3.3 EFFECTS OF APPLICATIONS ON CLUSTERING

Not all problems appropriate for a single large computer are
equally as appropriate for clustered minicomputers. In particular,
applications in which large amounts of data or processing are required
for individual job steps may require more resources than are available
to the individual minicomputers of a cluster. Also, some applications
are not appropriate for breaking into smaller segments.

To take advantage of a clustered configuration, an application
must be partitioned in a manner that allows concurrent activity on as
many processors as possible. This can be accomplished by dividing the
job into independent serial or parallel processing steps or a combination
of such steps.

Partitioning jobs into processing steps is particularly difficult
for a general-purpose computing system in which the mix of application
jobs is unpredictable. Either the jobs must contain flags to indicate
how they can be partitioned or each job must be considered as a single
unit of processing. One method is to consider the job as partitioned
into serial processing steps as a result of the service requests it
issues. Such an approach, however, can lead to an unreasonable amount
of processor switching and in considerable overhead to support the asso-
ciated communication between processors. The most likely environments
for clusters are those where a predictable or a repeating set of jobs
is to be executed.

When using clusters, care should be taken in advance to partition
jobs to promote concurrent activity by all processors in the cluster.
Each step must be within the processing capabilities of at least one
processor in the cluster. Processing steps that can proceed in parallel
lend themselves naturally to concurrent processing. Dividing a single
task into serial processing steps, however, does not necessarily allow
for concurrent activity. Only when multiple tasks are active can these
serial steps proceed concurrently with processing of other tasks. If a
task is continuing or repetitious, such as the processing of a serial
input data stream, and if input data can be divided into independent
portions, the processing of each portion of the data can be considered
as independent tasks. For example, processor A could be performing Step
1 on data portion i while processor B is performing Step 2 on data portion
i + 1, etc.

Space data processing has many functions that lend themselves to
processing with clusters. Examples include the pre-processing functions
of converting the serial data bit stream to parallel measurement data
with calibrations, the stripping of data to provide the required measure-
ments to individual users, data base management functions (as described
in Section 6), specialized processing functions that consume excessive
processing time (such as correlations and power spectral density calcula-
tions), communications functions for data acquisition, user interaction
from multiple sources simultaneously, simulations, and any number of
other activities.

In attempting to partition jobs into serial and/or parallel
processing steps, the following rules should be observed:

- Each step must be self-contained, and all necessary
  data should be available at initiation; and

- Each step must be within the capabilities of one
  of the processors of the cluster.

If all processors of the cluster are to be identical, with no particular
processing step dedicated to an individual processor, all serial steps

should require approximately the same processing time to minimize potential idle time, particularly where only serial activity is proceeding.

3.4   APPROPRIATENESS OF CLUSTERING

Performance and cost should be the deciding factors when considering a minicomputer cluster. Given that available equipment can be clustered as desired and the applications to be processed can be partitioned to take advantage of clustering, it must be established that system performance will be adequate and costs will be acceptable. Unforunately, it is frequently difficult to obtain data pertaining to performance degradation for clusters except through empirical measurements. Since clustered systems are not yet in wide use, such information is difficult to obtain.

3.4.1 Performance

Performance as discussed herein relates to the number of clustered processors required to provide a specific processing capability. For purposes of this discussion, assume that a cluster is being compared with a particular IBM 370 system. For the cluster to be acceptable, it must provide the same level of performance as the IBM 370 system. Computer system performance is a difficult commodity to measure and can be expressed in terms of a variety of factors. The particular factors selected for a specific evaluation will depend on the data available and the performance characteristics considered most significant for that evaluation. For one application, the average number of instructions executed per second might be an adequate performance indicator. For a different application, data throughput might be the best performance indicator. For still a different user, a weighted combination of these two indicators may be needed.

If the average number of instructions per second is the performance indicator used, and if all variables in instruction sets and instruction times have been normalized, the performance can then be expressed in terms of the following variables:

- I_1 – average number of instructions per second for a single processor

- D_n – total degredation in instructions per second by coupling n processors

- I_n – effective number of instructions per second in an n processor configuration.

The relationship of these variables is:

$$I_n = (n \times I_1) - D_n$$

For this example to be acceptable from a performance viewpoint, a value of n must be found such that I_n is equal to or exceeds the instructions per second for the IBM 370. I_1 is fairly simple to obtain or compute for a given set of applications, but D_n is more difficult to determine. First, D_n is a function of the cluster configuration and the resource allocation/scheduling algorithms used. Second, the complexity of the problem does not easily lend itself to a straightforward analytical examination. Thus the analyst is heavily dependent on statistical data from similar configurations, which may or may not exist.

One example for a tightly coupled multiprocessing configuration using a mainframe system demonstrated that performance cannot be improved by adding more processors beyond a certain number (eight to nine in this case). This is a result of the fact that the degradation factor for each new processor in this configuration exceeded the processing power of the additional processor. Similar results could be expected for a minicomputer configuration.

Although it has been stated above, it should be reemphasized that comparisons such as those described in this section are highly applications-dependent, and the conclusions drawn may differ substantially for different applications. Further, the flexibilities that are available in larger systems result in capabilities that should not be overlooked since most applications grow and/or change with time.

3.4.2 Cost

In a manner similar to that described above, it is possible to determine the maximum number of processor that make up a candidate cluster based on cost considerations. Again, assume that the decision is between a cluster and a mainframe system. Both initial and operational costs must be within acceptable bounds for the configuration selected. Operational costs will include maintenance, operator cost, and software development cost associated with the system configuration (not the applications themselves).

Unfortunately, clustered configurations have little existing control software. Special control programs will likely be required for new applications if they are to take advantage of the cluster's capability. Therefore, in a volatile environment, the cost for system software for a cluster may be significantly higher than the software required by an established larger computer.

Given the conditions under which the clustered configuration will be operating, it is possible to determine the maximum number of clustered computers that can be purchased at a cost that is equal to or less than the cost of the mainframe system by adding all of the cost factors for each system over the expected lifetime of the systems. For the clustered configuration to be competitive, the minimum number of computers required to achieve the desired performance should be less than the maximum number that can be purchased as described above.

4. LOW-COST COMPUTING SYSTEMS FOR NASA DATA USERS

This section of the report addresses a particular class of low-cost computing system that is likely to be of interest to low-budget NASA data users. The class considered is made up of the new low-cost microcomputer systems that are evolving out of the "computer hobbiest" marketplace and making their way into minimal-cost industrial and commercial applications. A particular set of capabilities required by selected classes of NASA data users is identified, and systems responsive to these needs are discussed. Previous reports, under separate cover, have addressed different aspects of low-cost systems, including mini-computers, microcomputers, and terminal devices.

Although limited in processing power and I/O capabilities, the low price of microprocessor-based systems has made them important, both as dedicated or independent processors for selected applications and as integral parts (e.g., controllers and interface modules) of large and medium-size computer systems. Microprocessor systems of the class considered herein are distinguished by their low prices and recent entry into the marketplace and by the growing, although still limited, hardware and software support that is available for them.

Although most of the companies producing these systems are still rather small and new to the industry, the capability and credibility of their products are growing rapidly. Currently, a strong movement is underway by government and business to evaluate these systems in terms of their capability and applicability to particular problems. The systems are inexpensive enough that such experimentation is cost-effective. As more groups and individuals recognize the capabilities of such low-cost systems, and as these systems are incorporated into more products, increased competition can be expected, thus causing vendors to offer more support and broader product lines in an effort to maintain their share of the market.

The low cost of these data processing systems makes them attractive to a particularly wide range of NASA data users. In fact, such systems

have the potential to make NASA data available to a new class of data
users that were previously unable to afford the processing costs associated
with the access and use of the data.

4.1   USES AND USERS FOR LOW-COST COMPUTING SYSTEMS

This study addresses the application of low-cost computing systems
to classes or categories of users in terms of processing needs rather
than addressing a particular application (e.g., land use and crop fore-
casting) or type of user (e.g., agricultural agent and hydrologist).
It is envisioned that each application will have users that fall into
several classes of processing needs, and further it is likely that many
new and as yet undefined classes will evolve as additional NASA data
become available and as the capability for handling the data becomes
economically feasible to a wider class of users.

The particular equipment considered in this survey is that which
a typical NASA data user requiring the capabilities of an intelligent
terminal, some local storage, and a limited degree of local processing
would desire. Such equipment should offer a capability for a higher
level language, such as BASIC, local storage either by floppy disk or
magnetic tape, a keyboard for input, an output device, and a communications
capability.

The low-cost computer systems surveyed in this section have the
capacity to perform complex processing tasks but are limited by their
relatively low processing speed and lack of extensive peripherals.
Although these limitations are gradually becoming less of a consideration
as the result of technology advancements, requirements exist for low-
cost systems with capabilities such as those currently available. In
particular, the existing systems are practical for those applications
in which a user needs limited on-site processing capabilities but in
which extensive scientific analysis or large-scale data base manipulation
is not required.

Potential NASA data users in this class are typified by those individuals that use public data that have already been processed and are ready for public dissemination. Any end-user processing consists primarily of reformatting and combining the data to produce results that are compatible with the needs of the user. The fact that the user is remote from the main data base also requires a communications and display capability as well as a limited capability to store data. Typical user needs and system capabilities to respond to these needs are included in Table 4-1.

The majority of the data being collected or planned for the Office of Applications programs is either geographic in nature or is at least related to geographic coordinates. If a user needs to take the individual measurement data and convert them to high-resolution geographically related information, then the processing power needed for that purpose far exceeds that capability discussed herein. However, if the data are already processed at the national or regional level and are available in a data base that is accessible to remote locations, then the end-user at these locations can extract, format, and present that part of the processed data that applies to the local level. In addition to the communications, formatting, and display functions, the local processing capability might include certain editing and decoding functions, and it would most likely involve coordinate conversions, as well as various forms of encoding to present the data in an easily recognizable format.

The geographical data collected by NASA falls into essentially one of two categories. The first category is represented by dynamic data that have a limited lifetime and that must be updated periodically (e.g., weather or water availability data). The second category is represented by those data that are more static and less subject to change over a relatively long period (a few months); e.g., land use data. A remote user has the choice of either storing the data or calling them up from the remote data base and processing them each time they are needed. Both categories of data have certain common parameters that can be stored

TABLE 4-1. USER NEEDS AND MICROCOMPUTER SYSTEM
CAPABILITIES

USER NEEDS	SYSTEM CAPABILITY AVAILABLE
Interrogate Remote Data Bases or Enter Local Data into a Central Data Base	• Communication capability • Edit communications before transmission • Accumulate requests/data for block transmission • Code communications into standard format expected by a central site • Decode replies or compressed data from a central size • Communication error detection and correction
Create and Process Data Files	• Create floppy disk files • Create tape files • Accept floppy disk inputs from other installations • Accept tape inputs from other installations
Update or Add to Existing Files with NASA Data	• Update floppy disk files • Update tape files
Summarize Data from Local Data Base	• Search files • Sort Data • Extract specific data • Perform arithmetic operation • Save summary results • Output summary results
Display Data	• Decode format information encoded into NASA-supplied data • Present data in user format • Display via an output device (CRT terminal or keyboard/ hardcopy device)

(e.g., grid information). Of course the static data can be stored such that an update would involve only changes to those data points that differ from the last time the data were stored. The local storage of data is prudent from a cost standpoint to both the end data user and the facility that maintains the data base. Such storage reduces the communications cost to the end-users and reduces the search time required at the main facility. Thus it can likely be concluded that users in the class considered herein will all have some storage capability. The class of user who would not have a storage capability includes users who depend wholly on the data base facility for all processing and are willing to accept public data in a standard format that is available to everyone. Such users would require only a display terminal of some type, either hardcopy or CRT, and would not need the processing capabilities discussed in subsequent paragraphs.

4.2   LOW-COST MICROCOMPUTER SYSTEMS

The microcomputer systems discussed herein are often referred to as hobbiest systems. One reason for this is that most of these systems are offered both in kit and assembled form. Also, the market that brought most of them into existence was created by hobbiests interested in building and owning their own computers. This heritage has tended to keep the prices of these systems low. As competition in this market intensified, manufacturers have looked to the methodology and equipment used on larger systems for ways to improve their products. Peripheral devices, software, and vendor support are rapidly improving in an attempt to broaden the initial market base. Table 4-2 summarizes the components and attributes of low-cost microcomputer systems that are currently available.

Low-cost microcomputer equipment that is responsive to the potential NASA data user needs discussed herein is available from a variety of vendors. Although the number of vendors is still relatively small, the demand for this type of equipment is expected to bring other vendors into the field. A number of other vendors are already offering basic systems; however, these vendors do not offer sufficient capability to respond to the potential needs outlined herein.

TABLE 4-2. LOW-COST MICROCOMPUTER SYSTEM COMPONENTS AND ATTRIBUTES

COMPONENT	ATTRIBUTES
Processor	Most major microprocessors are available in a computing kit or low-cost hobbiest system, particularly the models that use 8-bit words.
Processor Memory	Most systems can support up to 64k bytes of memory, using any combinations of: • RAM (Random Access Memory) • ROM (Read Only Memory) • PROM (Programmable Read Only Memory)
Mass Storage	The most common media are floppy disks, digital cassettes, and audio cassettes. One vendor offers a 200M byte disk and interface
I/O	Teletype and CRT/keyboard terminals are interfaced to most systems. Higher-speed paper tape readers and punches and a variety of high-speed printers are also available. Many systems support graphic (some offer color graphic) output. Most have both parallel and serial I/O ports available.
Software	Most systems come with a system monitor in PROM. Cassette tape and disk-based operating systems are generally available. Along with assembly language, most systems offer BASIC as an optional higher-level language.

The system configuration chosen as appropriate for the user considered herein includes:

• CPU with system monitor
• Minimum of 16k bytes of memory with at least 4k user space
• Floppy disk or magnetic tape mass storage
• A higher-level programming language such as BASIC
• Keyboard/display terminal device
• Remote communications capability (I/O port for modem).

Table 4-3 presents the system characteristics for some selected micro-computer systems capable of satisfying this configuration requirement.

TABLE 4-3. CHARACTERISTICS AND COSTS OF SELECTED MICROCOMPUTER SYSTEMS

MANUFACTURER	MITS	IMSAI	SPHERE CORPORATION	SOUTHWEST TECHNICAL PRODUCTS CORPORATION
Model	Altair 8800B	8080	System 330	6800 Computing System
Microprocessor (word size)	INTEL 8080B (8-bit)	INTEL 8080 (8-bit)	Motorola 6800 (8-bit)	Motorola 6800 (8-bit)
Memory	24k bytes	24k bytes	20k bytes	16k bytes
Mass Storage	Dual floppy disk	Dual floppy disk	Dual floppy disk	Digital cartridge tape
I/O Terminal	Lear Siegler ADM-3 CRT	Lear Siegler ADM-3 CRT	Keyboard/CRT integral part of system	Lear Siegler ADM-3 CRT
Higher-Level Language	BASIC	BASIC	BASIC	BASIC
System Cost*	$7,970	$6,570	$5,790	$2,800

*Assembled system cost includes serial I/O port for phone communications
with central site but does not include cost of modem or acoustic coupler.

Obviously, the major advantage of these systems is their low cost. For less than the cost of renting a larger computer, a user can purchase a very respectable system. This cost advantage is achievable because few unused capabilities are engineered into these systems and, so far, a low level of software and hardware support is provided. Currently the only type of system expansion available is the addition of more sophisticated peripherals. One vendor (IMSAI) offers options to construct multiprocessor systems with shared memory, but no one offers a line of more powerful and faster central processor.

Those microcomputer systems available from the mini-vendors such as DEC and Data General, feature instruction sets that are subsets of their mini line, ensuring upward compatibility for their software. The cost of such upward compatibility and extensive support, however, is included in the purchase price of the equipment, making it nearly twice as expensive as the low-cost systems included in this survey.

4.3   LOW-COST COMPUTING SYSTEM PROJECTIONS

The equipment discussed in this report has emerged and developed so rapidly that projecting its future is difficult. Emerging from a technically oriented marketplace, most of these systems still require technical personnel to configure and operate them. Their low cost, however, makes them attractive to people with a wide variety of backgrounds. Such a demand is already being felt and responded to by the vendors. During the next few years, a wide variety of new products in this area will emerge. Increased availability of support for nontechnical users, along with the development of dedicated application and turn-key systems, will result. Costs for dedicated systems will drop as component prices drop, but it is expected that general-purpose systems will experience capability increases rather than drastic price reductions.

4.3.1 New Products

New products in this market area are expected to appear in two categories. First, the success of current products will lure more vendors into competition for this business, resulting in a wider selection

and added features in the type of products now available. This market could currently be called a sellers market. The entry of more vendors should correct this situation. In fact, the success of these systems will probably prompt some of the larger minicomputer vendors to place their bottom-line computers into the competition for this business.

In conjunction with the added selection for existing products, an industry of new peripheral and support products is bound to develop. There are currently very few peripherals commensurate with the low cost of these basic computers. One of the main problems is that existing peripherals are designed to tolerance levels and with speed capabilities that are unnecessary for most of these systems. This degree of sophistication in design and production necessarily impacts the cost of these devices, thus pricing them out of the range of many potential users.

This void for peripherals will be filled with products that are either scaled-down versions of existing equipment or imaginative new designs. These new designs will be the result of new technology developments, as well as the implementation of techniques that are impractical for larger computing systems but are adequate for these low-cost systems. To keep prices down, mechanical parts will be held to a minimum with an accompanying emphasis on electronic devices. Areas where the need for products is acute include hardcopy output devices, on-line mass storage, and packaged software.

4.3.2 System Support

As less-technically oriented users become interested in these low-cost microcomputer systems, the availability of system support will increase. Approximately one year ago, the first computer store that retails computers and system development support opened. Now there are approximately 100 such stores. To recruit customers, these stores are offering local assistance to users rather than requiring direct contact with the manufacturer. In addition, many engineers who started as computer hobbiests are now offering contract support for these systems. This growing local availability of assistance, along with the expected increase in vendor support, should greatly benefit the market for low cost microcomputer systems.

Another aspect of increased user support will be the emergence of standard packaged and turnkey systems. Local government agencies and small businesses that are not able to support a technical staff to tailor a system to their needs are candidates for these types of systems. Also, the number of potential customers with specific needs in certain areas will prompt the development of systems designed for specific applications. Already many ambitious new companies are attempting to respond to the need for low-cost scientific, business, and process control computer systems. Potential customers, however, will not invest in such systems without proof of the reliability of the product and assurance of its continuing support. Since these products are relatively new, such evidence is not available, and the initial response is not as great as can be expected later on.

4.3.3 Price Trends

Price trends for the electronic components in low-cost micro-computers are in the midst of a steep decline. Technology and manufacturing advances are expected to maintain this steep decline in component prices. Thus, for those products whose price is primarily based on component costs, substantial price reductions should continue. On the other hand, the continued increase in support services, as desirable as it is, will necessarily increase product costs. Already many of the current prices offered by vendors are well above their initial levels. These increases are partially the result of added services being provided and partially because of the added levels of management and marketing that have been introduced in order to handle the large demand for this equipment. As the result of these factors, users should not expect major near-term price decreases in the low-cost microcomputer field but rather should expect increases in system quality, capability, and support. Of course, stripped-down systems and individual component prices should decrease.

4.3.4 NASA Applications

Advances in the technology discussed herein will result in significant improvements in the public's potential access to NASA data.

Advances in storage devices make it more practical for users to have relatively large local data bases, and improvements in system support will provide definite advantages to the nontechnically oriented users such as land-use planners and agricultural users. Just as turnkey systems are being developed for business, they can be developed for NASA users. Such an undertaking will require the definition of a set of specifications that are appropriate for a large group of users. For example, an agricultural reporting network could be developed in which local groups or agencies have microcomputer-based systems also serving as intelligent terminals to central data sources. Local data could be entered and made available to other users, while NASA data could be summarized and made available at the local level through these systems.

In summary, the decreasing cost of computer power will allow current NASA data users to perform more extensive analysis and will enable new users with limited budgets to become NASA data users.

5. MINICOMPUTER-BASED DBMS ACQUISITION AND TESTING

Current interest in data management is beginning to manifest itself on today's minicomputers in the form of sophisticated information retrieval and data base management systems packages. Taking advantage of the wealth of experience amassed by users of similar systems on large computers, these emerging minicomputer systems appear to rival some of the larger, more established systems in terms of user features. Such systems, however, are very complex and conceivably could be lacking in efficiency and/or reliability. Only a few individuals are acquainted with the range of data management systems available on minicomputers, and many of these systems are so new that only a very few installations are using them. They do, however, represent a major influence on the development and use of applications programs. Once a system is installed, users naturally become dependent on that system by virtue of the programs they develop. Choice of such a system is therefore particularly significant and must be based on experience with such systems rather than "vendor-claimed" capabilities or features.

Subsequent sections discuss problems of acquiring data base management systems for test purposes and identify some of the tests that should be performed to compare one system with another.

5.1   PURCHASING DATA BASE MANAGEMENT SYSTEMS FOR TESTING

Teledyne Brown Engineering (TBE) investigated the availability and cost of various low-cost minicomputer-based information retrieval and data base management systems to establish the feasibility of purchasing and testing these systems. Although lower in cost than similar systems for large computers, their cost is still substantial. The costs for Varian's TOTAL is approximately $10,000, and DEC's DBMS-11 is approximately $15,000. Equally important, these systems have been adapted and tuned to take advantage of the architecture and instruction sets of specific lines of computers. Therefore, they cannot be tested on a fixed

set of equipment. Thus, each DBMS requires its own unique set of equipment to obtain a fair evaluation of its performance. Designed for large data base applications, these systems generally require extended instruction sets, large memories (64K to 128K words as a minimum), and large disk storage units for most efficient operation. Such minicomputer systems alone cost more than $75,000. An investment of $100,000 to acquire a data base management system and the appropriate computer equipment would not be unusual.

In light of these figures, it is not generally feasible to purchase minicomputer-based information retrieval or data base management systems simply for test purposes. Instead, an attempt was made to identify less costly ways that would enable NASA to test such systems. The alternatives identified were:

- Ask vendors to provide access to their demonstration systems

- Arrange to use time on an existing government or industrial installation that has the required configuration

- Purchase time on a commercial service network having the desired system

- Acquire the DBMS on a loan basis from the vendor and secure equipment time independently

- Acquire the DBMS on a short-term lease basis and secure equipment time independently.

Tests that are performed at installations where the DBMS is already installed and operating are more desirable than similar tests that require the installation of a system. The first advantage is that of cost, since installation of such systems is generally a major and expensive undertaking. Second, such sophisticated systems must go through an initial shakedown period before they are operating smoothly. To test a system during this period would unfairly bias the test results.

In general, the necessary tests could be conducted at remote sites using telecommunication equipment. Test data bases could be transmitted to the site using either magnetic tape or communication lines. A major drawback to such remote testing schemes, however, is the lack of direct contact with experienced users. It would be appropriate to arrange for such assistance regardless of where or how the test is conducted.

The use of vendor demonstration systems has a number of advantages, including the availability of trained personnel and minimal cost requirements. Most major vendors have such systems. Unfortunately, vendor cooperation is likely to be proportional to the chances they perceive for selling their products. Such tests may be difficult to arrange under the auspices of a technology study where immediate purchase is not an objective. Initial contacts with Varian representatives indicate that a reasonable amount (undefined) of test time for TOTAL could be provided to the government as a service and that extensive tests could be arranged on a leased-time basis. DEC also supports demonstration systems and indicated that similar arrangements were possible.

The use of existing government installations is also attractive from a cost viewpoint if the DBMS and the computational facilities are available on a noninterference basis. Also, use of government-owned equipment, in conjunction with either the loan or lease of the DBMS, is attractive if an installation with the DBMS already installed is not available.

5.2   TESTING MINICOMPUTER-BASED DATA BASE MANAGEMENT SYSTEMS

Testing data base management and information retrieval system is complicated both by the intricacy of such systems and by virtue of the wide range of capabilities and performance they offer. Tests must be designed that completely exercise each system to determine what

features are available and how well these features are supported. Some aspects of how well a particular feature is supported include such factors as:

- Quality of support documentation

- Use orientation requirements (how easy is it to take advantage of the feature)

- Processing efficiency for providing this feature.

To ensure comparability of individual features among various systems, care must be taken to identify a basic set of tests for each feature and to apply these tests as identically as possible among the systems. This will entail the definition of representative data management applications and a basic set of data to use in each test. Plans for evaluating the features discussed in the following paragraphs must be developed in preparation for these tests.

5.2.1 Ease of Implementation and Use

Ease of implementation and use is particularly difficult to qualitatively assess. It is fairly simple to investigate the existence of various features such as particular types of user documentation, supported languages, and user aids, but their quality is as important as their existence. This qualitative judgement will develop during the course of the test.

Documentation should be complete and easy to understand. Detailed documentation to assist the data base administrator in designing the system should include detailed information on system operation as well as tips and guidelines for optimizing system usage. Higher-level documentation should also be available to assist users who are not concerned with the data's internal organization but only with the use of the system.

Data Definition Languages (DDLs), Data Manipulation Languages (DMLs), and/or query languages have to be evaluated. If the DML uses host languages, its interface with each host language should be considered. Such factors as language completeness and flexibility will have an important effect on the user's opinion of the system.

Systems may or may not provide effective user aids and prompts during use. In particular, system-generated error and warning messages should be investigated for accuracy as well as clarity.

5.2.2 Data Independence

Data independence is a primary concern of data base management systems. It implies that the applications software is unaffected by the logical data base description and/or physical storage of the data. The degree of independence provided is to a large extent a function of the schema data description language and the physical data description language provided by the data base management system. Each DBMS must be tested to determine the degree of independence provided when the data base is updated and/or modified and when the physical file structure is modified. The data, the structure, and the storage media of the data base should be varied while the user's view of the data is monitored for inconsistencies.

5.2.3 Processing Efficiency

Processing efficiency can generally be measured in terms of CPU time required to perform particular operations. Such measurements take into account the software efficiency of the data base management system as well as processor efficiency. Of course the same set of operations will not be available on all systems. Operations representative of all capabilities of a system should be tested, with care taken to ensure that those operations available on different systems are tested under as nearly identical circumstances as possible. The following list of operations should be considered when evaluating this aspect of the DBMS:

- Data base initialization

- Data access, update, delete, and add operations
  with a constant environment

- Data access, update, delete, and add operations
  as the data base grows in size.

5.2.4 File Structure

Data base management and information retreival systems will support a wide variety of physical and logical (as viewed by the user) file structures. These will range from simple sequential files through networks, trees, and relational structures. Ideally, all of the system aspects presented in Section 5.2 should be considered in relation to all data structures available. This, however, would present an unreasonably formidable task, especially in light of the possible combinations and various configurations of these file structures in a complex data base. Therefore, it will only be practical to test representative combinations of available structures, with emphasis on the structures that are most representative of NASA applications.

5.2.5 Other Features

A number of other features will be available on various of the systems to be considered. If present, they too should be examined. Such features include error recovery, data reorganization, and data security.

Most DBMSs provide capabilities to maintain records of data updates performed and, if necessary, to "roll back" data files to some previous state. This feature is valuable for file back-up purposes (a record of updates can also be used to "roll forward" an old copy of the files), as well as for removing errors introduced by erroneous data. If available, this feature should be exercised to establish whether it is indeed easy to use and is reliable.

As physical files are modified, they may require added storage space or they may free storage space. How these two cases are handled can affect system performance and resource requirements. The degree of required user intervention, as well as the efficiency of the techniques used, should be examined by forcing their initiation.

Data security is available in many forms on the various data base management systems. It ranges from file read/write protection to

cases in which individual data fields can be assigned a number of access characteristics. Whatever type of data security is available should be tested in an attempt to violate that security if data base security is of concern to NASA.

6. USE OF DEDICATED DBMS PROCESSORS

This section presents a brief synopsis of the ongoing work by various vendors and universities in the area of dedicated processors and microprogrammed special-purpose processors for data base management. The information presented is survey in nature and covers some of the most pertinent work ongoing as of November 1976.

As the cost of microprocessor hardware decreases, it is increasingly more appropriate to assign individual processors to perform specific functions. Prime candidates for such assignments are commonly used functions, such as data base management, that can be handled independently of the main processing functions. This scheme permits the off-loaded processing to proceed concurrently with activity in the main system processor and in other system processors. The scheme is particularly effective when the dedicated processor is able to perform its functions faster than the main processor, either because the dedicated processor is faster or because its lower software overhead results in more efficient execution of instructions. Another important factor is that the high overhead functions associated with a DBMS use extensive processing resources, which can be provided at less expense on a small dedicated system than on a larger system. The improvement would be even more dramatic for microprogrammed processors that could be configured to handle lower-level functions within a data base management system.

The most publicized commercial venture that uses dedicated processors for data base management has been the work by the Cullinane Corporation to implement their Integrated Data Management System (IDMS) on a "back-end" processor. The system being developed is capable of using any IBM 370 processor as the host computer and a DEC PDP-11/70 as the back-end processor. The PDP-11/70 responds to the data access requirements of the host computer, thus leaving the host free to proceed with other tasks.

The results of this initial effort is that DEC is now offering a version of IDMS (called DBMS-11) as a supported software package.

Also, work is proceeding on the development of high-speed communication interfaces to increase the communication capability between the PDP-11/70 and the host IBM 370 system.

The initial effort, which has been funded by a number of Government agencies, with primary funding provided by the National Security Agency, is nearing completion, and a prototype system is expected to be available for demonstration and test early in 1977. This prototype implementation will use an IBM 370 Model 158 as the host processor. The PDP-11/70 requires 128K words of memory and operates under the IAS Operating System. Communications will be bi-sync, using an IBM 3705 communications controller and the DEC DQ-11 Bi-Sync communication adaptor. Work is expected to continue for some time in the area of communications and tuning operations on both computer systems to derive maximum benefits from this configuration.

Other prominent companies working on dedicated DBMS processors include IBM, which is reported to be very active in the field. Although they have not announced a product, there are rumors that their next major line of computers will be oriented toward data base applications and that dedicated or distributed processors will play a major role. Currently, a team at their San Jose Research Lab is investigating the feasibility of developing a data base management machine. Indications are that a final decision has not yet been reached.

The academic community is also very active in the development of dedicated DBMS processors. Investigations indicate that academic institutions are researching advanced concepts and designing original data base management machines rather than adapting existing equipment to their applications. Some of the well-known universities active in this field and their special projects include:

- University of Florida - Developed one of the earliest data base machines. This machine was designed primarily for hierarchically structured applications. The individual in charge is Professor S. Su.

- University of Toronto - Developing a data base management processor designated as the Relational Associative Processor (RAP). The individual in charge is Professor Schuster.

- University of Utah - Developing a relational data base machine. The individuals in charge are Dr. and Mrs. Smith.

- University of Illinois - Developing an information storage and retrieval processor that is designed to optimize the use of inverted lists. The individual in charge is Professor Hollar.

- Ohio State University - Currently working on the development of a data base machine with emphasis on a general-attribute-based model that is intended to be appropriate for a wide variety of data management techniques. The individual in charge is Professor Hsiao.

- Kansas State University - Participating in Cullinane's project to develop a "back-end" DBMS processor, with primary activity in the area of communication capabilities for dedicated DBMS processors.

The majority of these projects are oriented toward the use of a single processor to handle all data base management functions.

Another scheme that appears to hold promise is to use multiple smaller processors, with each processor being responsible for specific data access or management functions, such as index file maintenance, data compression/expansion, and binary search. Although the current work in this area appears to be limited, an increase can be expected in the near future.

7. DATA INTERCHANGE AMONG A COMMUNITY OF NASA DATA USERS

Previous sections demonstrated that NASA has the potential of providing data to a widely diverse and growing community of users who spend valuable resources processing raw data and creating data bases that satisfy their current and future needs. Much of the data within these data bases is common to the needs of other users. These other users may either have data bases of their own or simply be recipients of data with neither the need nor the capability to maintain a data base. This section addresses the issues and problems that affect data interchange among a community of users, with the intent of providing early identification of several areas that require future attention to facilitate data interchange as the data volume and the user groups continue to increase.

A number of important issues arise when examining the feasibility of interchanging data among users with different requirements, different facilities, and widely varying capabilities. It is safe to assume that data base owners will use dissimilar computers, different languages, and a variety of methods for storing and retrieving data. As an example, some data base owners may use the file management system provided with their computer; others may use slightly more sophisticated information retrieval systems; and still others will use data base management systems with varying levels of sophistication and capability. This section of the report looks at the problem from a general point of view where possible but gives more coverage to on-line data bases that use data base management systems than to the other approaches. It is realized that there are many applications within the NASA data user community where other approaches will be used. The effects of specific types of data bases on interchange and the justification for one approach over another deserve more attention than could be provided in this study.

Data base information interchange among a community of users can take on a variety of forms ranging from completely manual approaches

(including the transfer of magnetic tapes, printouts, maps, etc., via the mail) to completely automated computer-to-computer and terminal-to-computer inquiries and transfers of data. All appraoches used have common generic requirements for data base definition, user languages, formats, data representations, communication paths, and procedures to facilitate the transfer of data. Some of the basic issues are listed below and are discussed in the following subsections. Issues that are discussed include:

- Identification of data base content in a manner that enables a potential user to determine compatibility with requirements

- Definition of a communication language and data base format that facilitates data interchange

- Implementation of a data base organization that permits response to specifically defined require-ments in an accurate, timely, and efficient manner

- Selection and/or design of a communication medium that provides for an interchange that is compatible with the requirements of the users

- Use of control mechanisms that structure and main-tain the interchange system in an orderly and cost-effective manner.

The major emphasis throughout this section is on a general class of user who has requirements for data via electronic means from both a NASA-maintained data base and user-maintained data bases.

7.1   DATA BASE CONTENT

Data base content involves a number of basic considerations. The three considerations that affect users the most are:

- What data items make up the data base?

- Are the contents maintained current and accurate?

- How is the potential user made aware of the contents of the data base?

Each data base owner has different requirements and different data base content and operates within a different environment from any other data base owner. Therefore, the data bases will vary widely in type and in content. As an example of the many options, consider the potential content of a NASA data base for pre-processed and semi-processed data. The volume of data that will be accured over extended periods of time on numerous NASA missions necessitates detailed planning as to how the data bases should be implemented. The implementation will, of course, be a function of the requirements, which are in turn affected by the current funding and technology for handling the data. For example, it is highly improbable that sufficient funds will exist to maintain on-line data bases for all data that are collected on future NASA missions. Therefore, decisions must be made as to what constitutes an on-line data base and what means will be used to honor requests for data from off-line users. A NASA on-line data base might contain only a data directory that identifies the contents of an off-line data base; it might contain a directory plus current (i.e., less than 24 hours old) data; or it might contain all data related to a given mission or program.

There may be good reason (e.g., magnitude of data or lack of redundancy in data) for maintaining several data bases within both the NASA system and the user system rather than large integrated data bases. Specifically, data would probably be divided into separate data bases by program, and there is also good rationale for dividing data by mission. Further, the data would likely be stored in different data bases depending on whether it was raw data, pre-processed data, or final processed data.

Regardless of the content, the data within the data base must be maintained current and accurate. A user that interrogates the data base should be able to obtain the most current data; and most importantly, the data obtained must be accurate and unambiguous. The ability to achieve these goals is a function of the update philosophy and of the data base design philosophy.

Finally, for the data base to be effective, the user community must be aware of its contents. The vehicle for informing users of

data base content is the data index or the data directory. The directory lists all data in the data base and provides information required to use the data, including such information as the time the data were gathered, position with respect to some geographical reference, and data volume. In addition, the directory identifies the processed status of the data (e.g., raw data, preprocessed data, engineering units, combined, smoothed, or compressed) so that the potential user knows what data are being obtained and what processing functions need to be performed. Basically, the same information is required in the directory regardless of whether the information is available in an on-line or an off-line mode.

For discussion, it is assumed that NASA will maintain a centralized data base consisting primarily of pre-processed and semi-processed data that is available to all users. Similarily, some NASA data users will maintain data bases, consisting of various levels of processed data, that will be available to the community of users that are working in related areas. A user network could take on the appearance of the diagram presented in Figure 7-1. The data paths within such a network may be either electronic communication paths or some less sophisticated path such as the mail.

7.2 LANGUAGE AND FORMAT

The language referred to in this discussion is that language used to manipulate the data base. (It is sometimes referred to as the query language and/or the data manipulation language.) For this study, the use of the language is limited to that interaction required to interrogate the data base in order to request data.

The query language basically consists of a set of macroinstructions that are recognized as requests for data by the data base management system (if one is used). The language is also used by the data base management system to issue status messages to the applications program. These macroinstructions may be an extension of the applications programming
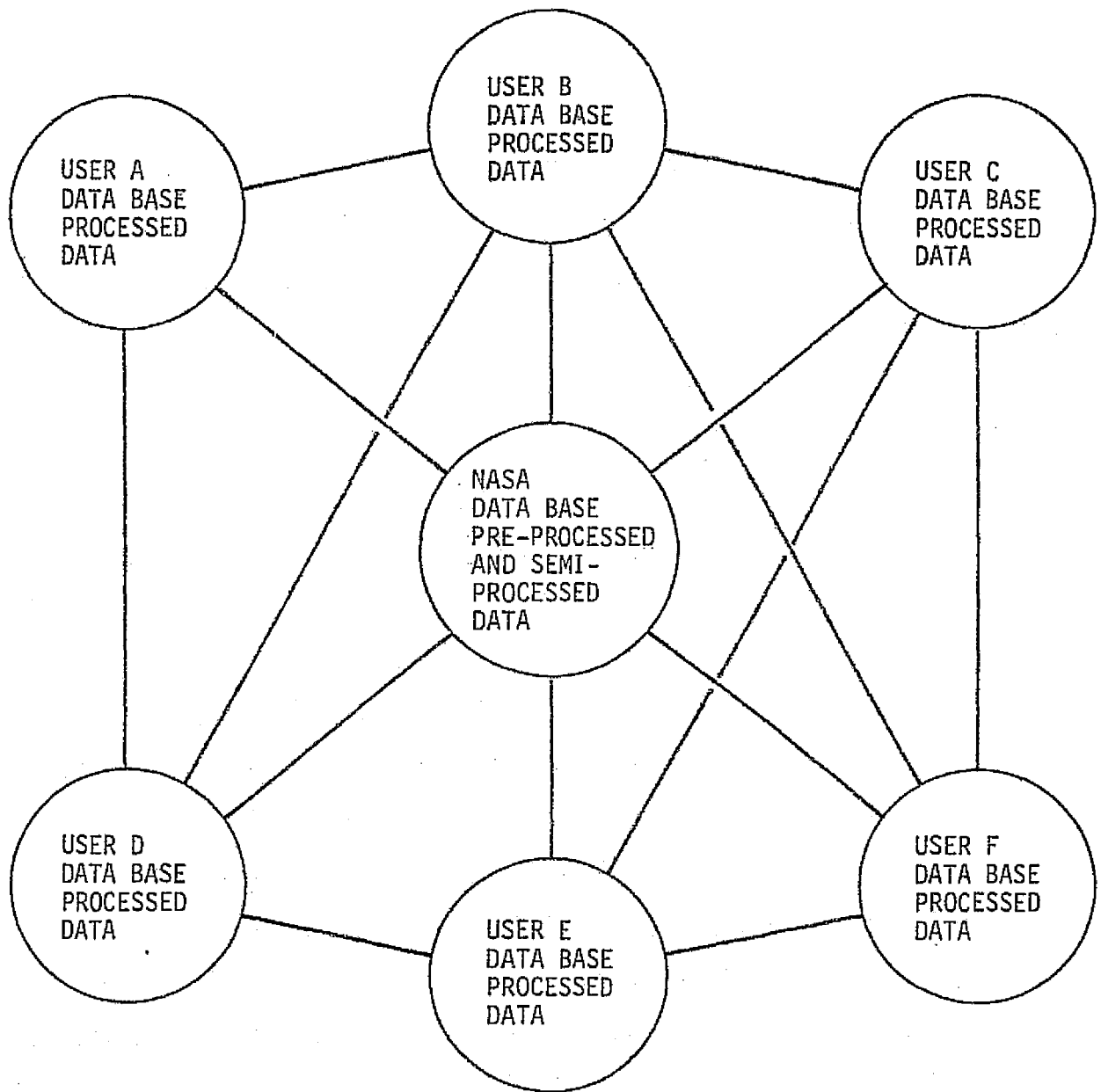
FIGURE 7-1. DATA BASE INFORMATION INTERCHANGE NETWORK

language, a separate sublanguage, or merely a set of call statements provided by a particular data base management system.

Each of the above approaches offers advantages and disadvantages. If the macroinstructions are an extension of the programming language, they can be independent of any particular data base management system. On the other hand, if they belong to the data base management system, they are independent of the programming languages. Finally, if a separate sublanguage is used, the language could in principle be independent of both the programming language and the data base management system.

The ideal situation would be if every user of NASA data that maintains a data base used the same language as well as the same data base management system. Such an approach is, of course, unfeasible in view of the fact that many data base owners will be adding to existing data bases that already have a defined language and data base management system. Possibly many others will depend on a file management system or some other approach to access their data base. Therefore, the problem of language for interrogation of data bases has the potential for being significant. One solution is to require users to interrogate with a standard language that each data base owner accepts and transforms or maps into a language that is compatible with the data base management system being used.

One language that has gained rather wide acceptance is the CODASYL Data Manipulation Language (DML) proposed by the CODASYL Database Task Group. This DML is said to be application-program-independent and capable of being implemented on a wide variety of data base management systems. However, the DML has been criticized rather widely for not providing full data independence, which is highly important in data base systems.

A data base owner who is not using a DBMS would surely find it impractical to respond to certain of the DML commands that are used by data base management systems. It may prove appropriate, therefore, to have various levels of data retrieval available for different data bases, depending on what each data base owner is willing to support.

Finally, the question of data format arises when requesting data. A data index or data directory was defined in Section 7.1 for determining what data items reside in a data base. One additional level of definition is provided by the data base dictionary. This dictionary, which may be either on on-line or an off-line (hardcopy) entity, describes the detailed format of the data base for the systems that do not employ a DBMS and describes the schema for data bases that employ a DBMS. In addition, it specifies the primary and secondary keys, plus any other information required to select those elements from the data base that are of interest to the user. Thus this document provides the primary mechanism for accomplishing format compatibility. If a standard query language is used, and if the communication protocol is compatible between users, the desired results will be produced by correctly specifying and formatting the request.

7.3   DATA BASE ORGANIZATION

Data base organization has been the subject of many books and articles, and it is not the intent here to repeat the results of these previous dissertations on the subject. This section addresses the effects of data base organization on data interchange among a community of NASA data users, without regard to such factors as efficiency, performance, security, and tunability. Although these features are important to the data base owner, they are not the primary concern to the occasional user who wants to access the data base. Factors that are important when a wide variety of users access a data base are:

- Applicability of the data for a variety of uses such that different users can perceive and use the data differently

- Simplicity of access

- Responsiveness that produces current, accurate, and complete results

- Immunity of applications software to changes in the data base.

The ability to accomplish the above objectives is not only a function of the data base organization but is a function of the architecture of the data base system. Because of the diversity of users and their different levels of sophistication, it is difficult to talk in general terms about the architecture of a data base system. In fact, it is highly probable that many of the data base owners will not use data base management systems, in which case portions of the discussion herein would not be applicable

Prior to proceeding, it appears that it would be appropriate to define a number of terms, some of which have been used previously:

- <u>Schema</u> – The overall logical data base description

- <u>Subschema</u> – A subset of the schema, which may be one user's view of the logical data base description

- <u>Data Independence</u> – The immunity of applications software to changes in the storage structure and/or the access strategy. This definition implies two distinct levels of independence as follows:

  - <u>Logical Independence</u> – Permits the overall logical structure of the data, as defined by the schema, to change without affecting the applications software

  - <u>Physical Independence</u> – Permits changes to the physical layout and organization of the data without affecting either the overall logical structure of the data or the applications programs

- <u>Schema Data Description Language</u> – Used to define the logical data description, including all internal relationships within the schema

- <u>Physical Data Description Language</u> – Provides the mapping between the logical data description and the actual physical placement of the data on the storage media.

In order to be accessible to a variety of users and uses, the data base should be organized to support the different logical files required by the different users These files will be derived from the

physical files of the data base. The organization method must accommodate changes to both the logical data base description and the physical files without affecting the user applications software. The Schema Data Description Language and the Physical Data Description Language are the vehicles for defining the logical and the physical files and the relationships that can be supported. By providing users with their own logical view of the physical data, and by supporting multiple data access methods, a data base management system can present a simple view of a complex data base.

While multiple access capability and simplicity are important features that result from the data base organization, the single most important feature of data base organization is data independence. Such independence implies that a request will result in a current and unambiguous response. Further, it implies that additions, deletions, and/or modification to the data base will not affect the applications software. In turn, independence helps provide the capability to support multiple users, since various users are able to view the data in terms of their own needs as defined by their subschema.

7.4   DATA COMMUNICATION MECHANISMS

A variety of communications mechanisms are available to accomplish the interchange of data among users. Included within the alternative communications options are:

- Formally documented requests that place orders for data within an official requirements document (e.g., the Program Support Requirements Document)

- Informally documented letters, mailgrams, teletypes, etc., that specify data in accordance with the format and definitions provided in the data directory and the data dictionary

- Interactive systems that permit the user to interrogate the data base to determine its latest contents.

Within the latter category of interactive systems, a number of options exist for the actual transmission of data, including the mail, messenger services, and electronic communications. The remainder of this section is primarily concerned with data communications via either commercial or dedicated data networks.

Data communications does not present any unsurmountable problems for data interchange among a community of users other than bandwidth limitations resulting from excessive channel costs. Data users generally have a choice of either the public dial-up lines or leased lines. Of the public dial-up lines available, the most inexpensive method for limited data transfer is the direct distance dial (DDD) network, which is available on a demand basis. In addition to being available on a demand basis, the DDD network provides a capability for communicating with every computer center within the United States simply because it is available over such a wide area.

The disadvantages of the DDD network are twofold. First, the maximum data transmission rate achievable via this network is approximately 4,800 bps when using sophisticated modems that provide adaptive equalization during transmission. It is usually restricted to 2,400 bps for less-sophisticated modems. Second, the cost of the DDD network becomes prohibitive when use exceeds a certain point, and it then becomes less expensive to use other types of communications links such as WATS, leased lines, or value-added networks.

Leased lines are capable of being conditioned to carry up to 9,600 bps via one voice grade line. Wideband switched lines are available to 50 kbs and wideband leased lines are available to 230 kbs; however, the costs of these lines are quite high.

The type of line (switched, point-to-point, or multipoint), the network used (ATT, specialized common carrier, or private) the bandwidth, and other factors are functions of the different data bases to be accessed, the volume of data to be transferred, the bit-error rates permissible, and a number of other factors. Important points to remember are that

point-to-point and multipoint lines restrict access to a very few other users. Specialized common carriers in general restrict the user to those other users on the same network. One form of specialized common carrier, the packet switched network (sometimes referred to as a value-added network), provides access to everyone via the public switched network, but the cost for using this network is greater than the cost for other networks below a certain data volume. Once this threshold data volume is crossed, the packet-switched networks offer an economic and a performance advantage.

The various communications aspects that have been considered above are primarily nontechnical. A number of technical aspects of data communications require compatibility among users, including:

- Data rate

- Modem modulation and signaling

- Protocol/line control procedures

- Synchronous/Asynchronous Control

- Transmission Mode

  ▲ Simplex
  ▲ Half-Duplex
  ▲ Full-Duplex

- Character Codes

- Error Control Techniques.

Certain of the above areas -- namely, modulation and signaling, protocol/ line control procedures, and error control techniques -- are highly complex subjects within themselves, and they deserve further consideration before the implementation of a network for data interchange.

One approach to networking that has the potential for reducing the technical problems, as well as reducing overall communications cost, is for NASA or another interested Government agency to act as the central facility for accepting and routing all communications between users.

Although the central facility would be relatively complex and expensive to implement, it does away with a number of problems. In the first place, all users with a communications capability will probably require a line to NASA. This same line could be used for data interchange with other users when it is not busy and thus eliminate the need for direct lines between users. Second, the central facility could be used to provide those interfacing functions necessary to achieve compatibility between users. A more detailed study is needed to assess the implications of this central facility. The study would evaluate the alternative approaches to implementation in view of the cost and performance capabilities that could be achieved for specific classes of users.

7.5   CONTROL MECHANISMS

Any time two or more end-users need to communicate, controls of some type are required to enable an exchange of information. When two individuals are exchanging data verbally, the control is exercised either as a result of courtesy or the desire to hear what the other individual has to say. When the requirements for data interchange become more specific and more demanding, the control mechanisms have to become more formal. The formal nature of the control takes on the form of standards, procedures and protocols for specifying requirements, formatting or structuring the data and communications. It may also take on the form of an organization to define and implement the control mechanisms; and if it is complete, it may include a feedback mechanism to assure compliance with requirements.

Control mechanisms that apply to the interchange of data among NASA data users will include the data base administrator (DBA), standards, the data dictionary, the data directory, communications protocol and procedures for electronic communications, and interactive feedback provided in the form of error messages, prompts, and the like. All the above mechanisms with the exception of the data base administrator and standards have been discussed previously. Subsequent paragraphs

present recommendations for the establishment of a data base administrator
organization and the implementation of standards that will be encouraged
and enforced by the DBA.

7.5.1 Standards

All users that communicate with the NASA data base will have to
be compatible in terms of query language, line protocol, modem modula-
tion characteristics, data rates, etc. To respond properly to data
requests and to assure minimal impact on the applications software, the
data base should be organized and implemented to provide both logical
and physical data independence. In addition, the data base management
system (if used) should contribute to the isolation of the applications
software from the logical and physical data base organization.

The above compatibility and data independence will be achieved
via detailed analysis and design specifications or standards that apply
to the data base and the interface it presents to users. The data
base design standards will apply to the initial design phase of the data
base and to subsequent modifications to it. The interface standards will
have to be implemented and continually monitored by each user on a
continuing basis to use the data base.

The same standards that apply to the NASA data base and its
user-interface could be applied to user-owned data bases. It is realized
that many of the user data bases are already in existence and changes
will be difficult. However, if some standards are not introduced and
enforced as feasible, data interchange among users will be extremely
expensive, if not impossible.

7.5.2 Data Base Administrator (DBA)

The implementation and enforcement of standards within NASA and
among the users would be the responsibility of a Data Base Administrator
organization that NASA establishes and controls. The DBA responsibility
would encompass all areas of data base design and data interchange,
including:

- The responsibility for structuring NASA's logical and physical data base to make it independent and at the same time efficient in terms of access times and computer utilization

- The responsibility for selecting and implementing data base management systems (either generalized or specific systems), as appropriate, that satisfy the requirements of data independence, speed, accessibility to a community of users, and other special requirements that may be unique to NASA

- The responsibility for keeping the data base current and for assuring that all data users have access to current listings in the form of a data directory, which may be either a hardcopy or an interactive presentation made available via terminals

- The responsibility for developing and distributing a data dictionary that defines all data formats, languages, commands, and other information required for data interchange with the NASA data base and among a community of users

- The responsibility for developing and assisting with the implementation of a set of standards for data interchange among a community of users.

If the DBA is able to satisfy the responsibility outlined above, the savings in time and data reduction cost will be significant. An effective system will serve many functions, including but not limited to:

- Reduction in storage requirements for data at the NASA processing facility and at user facilities

- Reductions in processing requirements among the user community

- Reduction in the time required to update and/or modify data bases at both the NASA facilities and the user facilities

- Reduction in software development costs (applications software and logical and physical data description software) at both NASA and user facilities

- Reductions in the time lag between data
  acquisition and data availability

- Reduction of errors in results produced
  because of inaccurate data.

Implementation of a successful DBA organization that will properly manage NASA's exceptionally large volumes of data and facilitate the interchange of data among a community of users is indeed a major undertaking. The task requires an organization of experts who understand not only data bases, data base management systems, and data applications but the unique requirements of NASA and the user community. In particular, a data base system for NASA data users requires a blending of the requirements for a commercial data base system with the unique requirements of scientific and engineering applications. At the same time, the requirements for privacy and security may be less in the NASA data base than in a commercial data base. Therefore, the requirements associated with this function can be relaxed in the NASA system and the associated improvements in processing time can be realized.

The most crucial phase of the program is the initial analysis and system definition phase. The conclusions drawn and the results produced will have a profound effect on everything that follows. Thus the initial data base structuring and organizing, the selection and implementation of a DBMS, and the development of the guidelines or standards will set the pace for the future ability to communicate among a community of users.