IBM

Washington
Systems
Center

Technical
Bulletin

Capacity Planning

Implementation

By Dr. LeeRoy Bronner
Washington Systems Center

DAPS Code 0933
GG22-9015-00
January 1979

Washington Systems Center
Gaithersburg, Maryland
Technical Bulletin

Capacity Planning Implementation

by LeeRoy Bronner
Systems Capacity Planning
Gaithersburg, Maryland

This Technical Bulletin is being made available to IBM and customer personnel. It has not been subject to any formal review and may not be a total solution. The exact organization and implementation of the functions described will vary from installation to installation and must be individually evaluated for applicability.

A form is provided in the back for comments, criticisms, new data, and suggestions for future studies, etc. IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation whatever.

DAPS Code 0933
GG22-9015-00
January 1979

ACKNOWLEDGEMENTS

There are many people whose help and lively discussions on the many aspects of capacity planning were instrumental in the preparation of this technical bulletin. Three in particular to whom I must express my special gratitude are August Deckert, James Cooper and Theresa F. Washington. August Deckert, Manager of System Capacity Planning, for the many corrections and helpful suggestions resulting from a painstaking reading of the manuscript. My colleague James Cooper for his work and many discussions on the section entitled "The USAGE Technique". Theresa F. Washington for her speed and efficiency in the clerical aspects of preparing this document. Especially, the preparation of most of the figures via her Advanced Text Management System (ATMS) Terminal.

LeeRoy Bronner

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

LIST OF FIGURES

LIST OF FIGURES (CONTINUED)

LIST OF FIGURES (CONTINUED)

LIST OF FIGURES (CONTINUED)

1.0 Introduction

1.1 What is Data Processing Capacity Planning?

The term "Capacity Planning" as used in the field of data processing has many different definitions. The primary factor that tends to link all capacity planning definitions is that each is concerned with the management of a computer system through some form of performance analysis. Also, the elements of prediction or forecasting of certain performance parameters is usually present in these definitions. Performance is defined as a measure of the effectiveness of the operation or functioning of the various components (people, hardware, software) of the computer installation. For example, the response time at a teleprocessing terminal is a measure of the on-line users perception of the performance of the computer system. In light of all the current controversy on capacity planning, the definition given below is one view of this complex subject.

Capacity planning might best be understood, if it is thought of as a process or methodology. The basic concepts underlying capacity planning are not new. For example, the concept of management by system performance, tracking or monitoring system operation in an ongoing fashion, workload measurement and definition are all ideas implemented in bits and pieces by many data processing installations over the years. Capacity planning, as defined here, is basically a systematic approach of bringing together many of the past performance management ideas and integrating them with current performance management and measurement technology (RMF, SMF, IMS/CICS Performance Data, etc.).

Capacity planning is a methodology developed for the management and control of the complex data processing environment. Capacity planning addresses the problems involved in managing the computer resources, namely

- What parameters to collect to characterize the workload.
- What parameters to collect to characterize the software and hardware components.
- What parameters are required to forecast future workloads and system performance.
- What products are required to collect, analyze and report the data items described above.

How should the DP executive manage his installation on a continuing basis using the data described above and the results of analysis (required reports, reporting formats, report flow, recipients, etc.).

Capacity planning is basically a performance oriented approach to data processing management. By this process, the loading, utilization and response of the various system resources are monitored and analyzed. Also, the flow of current and future work through the system is controlled to provide the best overall user satisfaction. User satisfaction is the most critical factor in the capacity planning process.

1.2 Purpose of Technical Bulletin

Capacity Planning is still very much an "Art", yet significant progress has been made since reference 1 was published in understanding it as a practical process. This bulletin will review some of the new findings and discuss "State of the Art" techniques for implementing a capacity planning program.

1.3 Integrating Capacity Planning Into the Organizational Structure

Current capacity planning techniques are primarily associated with performance measurement data, system modelling, analysis and prediction. However, experience is indicating that another very important consideration is the unique characteristic of the organization in which the process is being implemented. Therefore, a consideration of this technical bulletin is a discussion of capacity planning as it relates to the organizational structure (Figure 1).

In many companies capacity planning is being implemented. Since practical implementation of capacity planning dictates that any current planning process be understood and evaluated for retention of those parts (procedures, measurement tools, reports, people, etc.) deemed necessary for future use, organizations will probably find that the evolution of their capacity planning program is quite different from another installation. In essence, there is no "canned" technique which is right for every organization but there is a methodology to provide guidelines for the continued development and enhancement of any current process.

Another consideration which affects the capacity planning development process is for various data processing functions (operations, systems programming, applications development) and management responsibility to be located at seperate

locations or in different line organizations. A key factor in developing the process is for close coordination to be maintained across operations, systems programming, and application development even though these functions are separated or fragmented across corporate and divisional lines. If this is the situation within a company, then adequate procedures for obtaining the required coordination must be put in place. The management structure as well as separation in distance of these functions may increase the difficulty of implementing a successful capacity planning program.

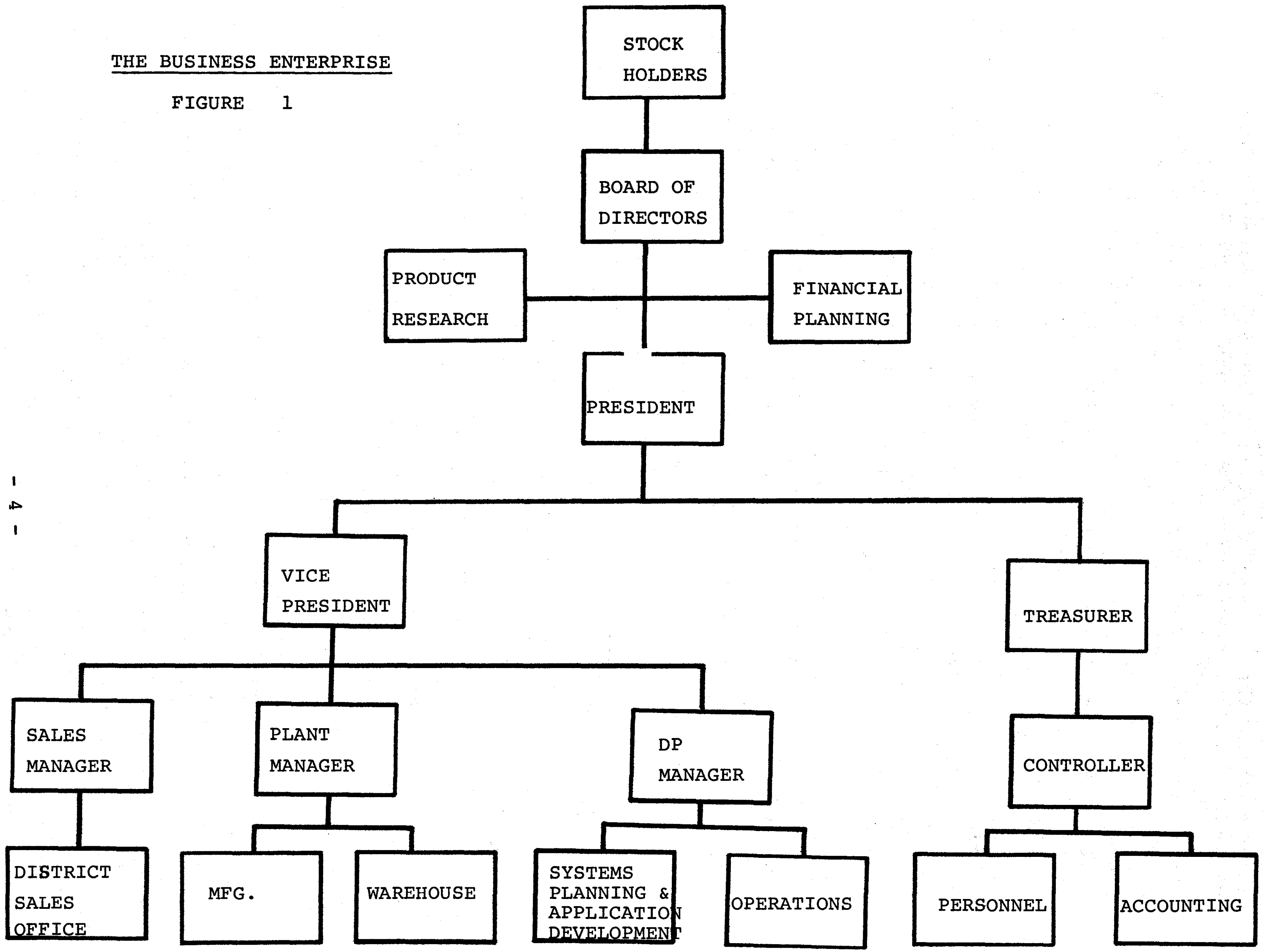
In most organizations, the capacity planning function is initiated or developed under the systems or technical support side of the data processing organization. Then, with various types of performance related inputs (workload forecasts, resource utilizations, CPU times, elapsed times, etc.), and some modelling technique (statistical, queueing, discrete simulation, benchmarking, etc.), the capacity planning effort is begun. This process is initiated in many cases with no regard for the importance of understanding the organizational structure (lines of responsiblity, where various DP functions reside within the structure, level of management committed to the capacity planning effort, etc.), and one year later no real progress is observed.

Capacity planning must involve many other departments outside of the one in which it is developed. Development of a capacity planning program should involve close coordination between certain designated people in the following areas as shown in Figure 1:

- Systems programming
- Operations
- Application design and development
- Users
    - Sales offices
    - Manufacturing
    - Warehouse control
    - Accounting
    - Administration/personnel

THE BUSINESS ENTERPRISE

FIGURE 1



The users are creating the workload and it is unrealistic to think that capacity planning can be adequately initiated without a close working relationship with the user concerning forecasting, workload characterization, and establishing user service objectives. Even in a service bureau type environment, contact must be made with the large critical users. These and many other relationships that the capacity planning process has within the organizational structure will be addressed throughout this technical bulletin.

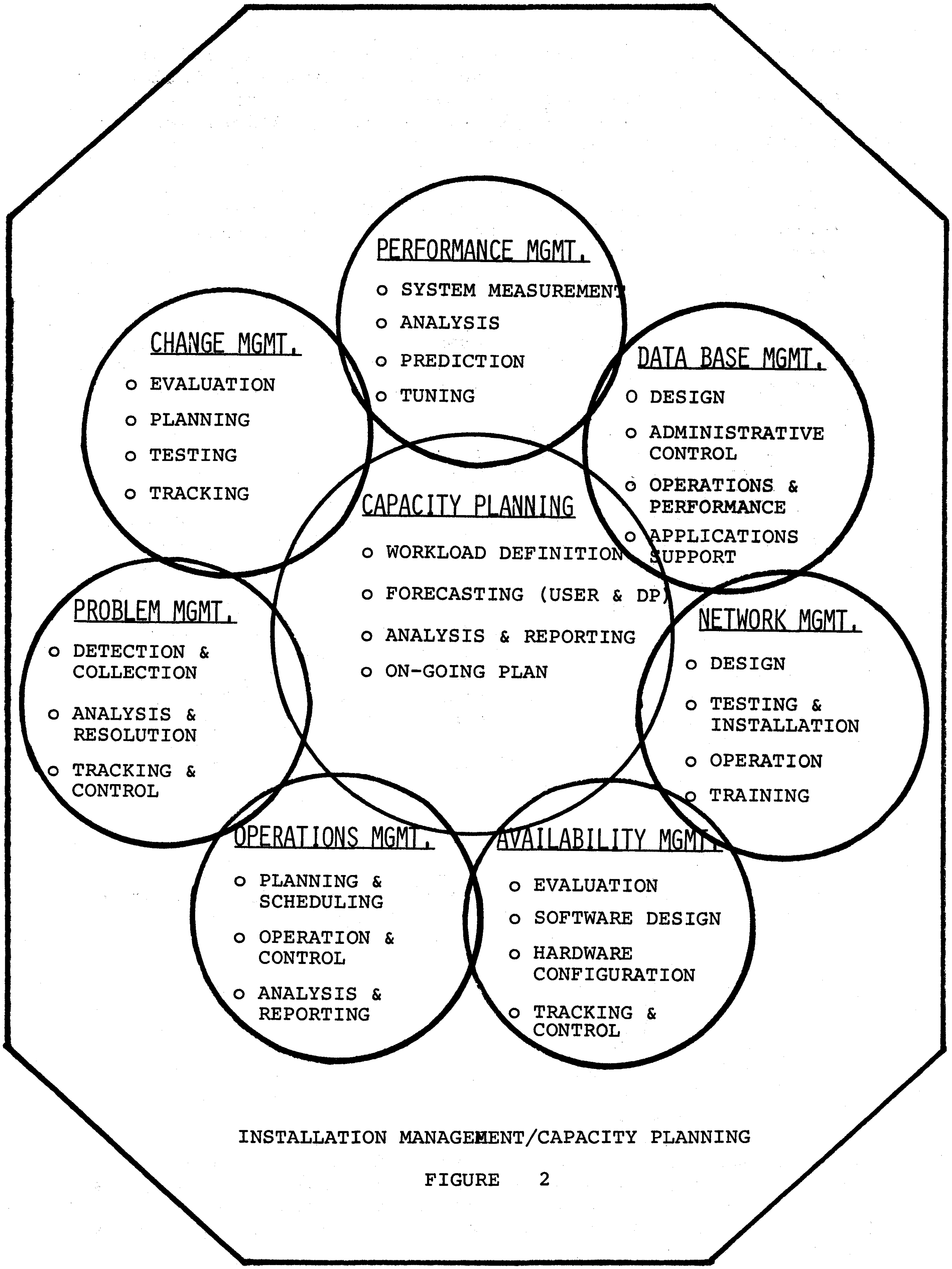
1.4 Capacity Planning In Perspective

The capacity of a computer system is defined in many ways depending on where you are in the organization (e.g., user, operator, system programmer, executive, etc.). Although these definitions will disagree on many points, most practitioners will agree that a critical factor in defining a system's capacity is its perceived availability by the user. Availability perception is closely related to a user's perceived service (response/turnaround time). The most important factor in understanding a system's capacity is the user service objective (clearly defined or implied). This will be discussed in Section 1.5. For example, with regard to availability, the computer hardware may be up and functioning fine but a software problem may cause a major application (e.g., IMS) to be down for its entire normal shift. Then, a week's (5 days) availability for this major application is reduced by one-fifth and user service may be severely degraded. Since Installation Management is a process directed at understanding, correcting and controlling anything that distracts from the normal operation of the computer system, it is very natural that a capacity planning effort overlap many Installation Management functions.

The objective of this section is to show the relationship of capacity planning to the larger subject of Installation Management (Figure 2). Installation Management is concerned with the management of the following areas of a data processing installation:

- Performance
- Changes
- Problems
- Operations
- Availability
- Networks
- Data Bases

Although Figure 2 only depicts the interrelationship of these areas with the capacity planning function, there is a definite overlap among these installation management functions. For example, Network Management would intersect all other areas.



INSTALLATION MANAGEMENT/CAPACITY PLANNING

FIGURE 2

1.4.1 Performance Management

Before discussing Computer Performance Management, it seems appropriate to establish a definition for performance. Webster's New World Dictionary defines performance as follows:

- Performance - "The act of performing"
- Perform - "To meet the requirements of, to fulfill, to achieve, to accomplish, etc."

Hence, it appears that performance is concerned with defining certain objectives, requirements or specifications for operation and measuring or determining in some way how effective operations are in meeting these requirements.

Performance Management of a computer system is primarily concerned with providing adequate service to a given user community at minimum cost. Therefore, the term "adequate user service" must be quantified and then it will become the primary basis for computer performance evaluation. The service a user experiences in a batch or on-line environment is the result of many diverse factors (workload, availability, hardware configuration, software, etc.). Performance Management is concerned with the measurement, analysis and control of these factors affecting user service. Some of the objectives of a Performance Management effort are:

- The ability to monitor the system's operation and determine the level of performance attained (user service objective).
- The ability to analyze measured data and tune the system to improve performance or to increase the workload without degrading performance.
- The ability to predict the impact on systems performance due to changes in the hardware or software.
- The ability to predict the impact on performance due to increased workload on existing applications or the addition of new applications.

Performance Management is very closely related to capacity planning and in some instances may be taken as being one and the same. It is my contention that it is possible to have performance management without capacity planning but not vice versa. Capacity planning is very heavily dependent upon performance analysis of the computer system as a basis for monitoring and comparison of various system parameters

(e.g., CPU time, response time, elapsed time, etc.). But, capacity planning takes forecasting and prediction a bit further, in that, it is concerned with the end user forecasting as it relates to his natural forecast units (check volumes, printed circuit board volumes, etc.) and the reporting of performance analysis results, including the types of formats and whom should be the recipient of the results, etc. For those that would tend to disagree with the fact that it is possible to have performance management and not have a complete capacity planning solution, the disagreement may only be in semantics. But, if this is not the case, hopefully this technical bulletin is able to clarify the additional tasks covered by the Capacity Planning Process.

1.4.2 Change Management

In today's computing environment, change is synonymous with data processing. Hardware technology advances, new software functions, new releases of the control program, and system modifications for convenience; all force change in the computing system. It is the dynamics of the computing environment that makes installation management such a difficult task. Change management, as a function of installation management, is concerned with the control and scheduling of changes on a regular basis to minimize disruption to the data processing environment [2]. System changes must be evaluated and their impact assessed to determine the reasonableness of implementation. Changes must be planned, scheduled, tested and affected parties notified. Consideration should be given to a change cutover process when changes are incorporated in the production system. Plans for immediate monitoring of these production changes must be developed and implemented. These plans will include procedures for tracking, reporting and backing out changes if necessary. Change history files should be maintained for future reference.

Changes can significantly affect the capacity of a system. The capacity may be increased or decreased depending upon the change. Hardware as well as software performance enhancements can increase system capacity. However, implementation of these changes may cause serious degradation of user service or system capacity over extended periods. This may be caused by improper planning for changes or changes which were not adequately tested, etc. During these periods, the system experiences, among other things, excessive loads, degraded system response and increased levels of user dissatisfaction. Hence, change management is very crucial to the adequate management of system capacity. A gradual decline in system capacity may be traced to some system change and certain capacity forecasts seriously impacted. For example, a 12 month

forecast of adequate system performance may degrade to only 9 months because a change was not incorporated that was scheduled to improve system capacity. Therefore, changes must be tracked and evaluated to determine their actual system impact.

1.4.3 Problem Management

In most data processing installations today, there are three different types of environments, batch, on-line or a combination of both. In most cases each environment will have different objectives, different functions, different organizations. Also, they will have different problems requiring different solutions. These problems may be experienced within or outside the computer room. Within the computer room, most problems will be recorded by operations for immediate or future resolution. Those problems occurring outside the computer room may be identified by a programmer or an end user.

Within the two areas outlined above, there are basically three categories of problems [2] which are listed below.

- Problems which have been in the system for some time but have only recently occurred or been detected.
- Problems which have developed in the system through "natural causes" (e.g., hardware component failure).
- Problems introduced in the system through changes.

Certain problems can be quickly isolated to a particular area of the system and then resolved. But some problems cannot be readily isolated, indeed sometimes one cannot even identify who should be working on the problem. If there are no specific procedures for assigning responsiblilities for resolving each problem, then the problem will linger.

The goal of the problem management process is to identify hardware, software, and operational problems in the system and to provide effective means to track these problems and ensure their resolution [3]. Reference 3 is an actual user account of the effectiveness of a problem and change management system in their installation.

From the previous section on change management and again under problem management, it is clear that those things (hardware and software) which tend to reduce the time the computer system is available to the user reduces the capacity of the system. Where system capacity is directly

related to the system's ability to satisfy user service objectives.

1.4.4 Operations Management

The data processing operations department functions as a service organization carrying out the instructions of various user departments [4]. In this environment, operations must prioritize and schedule for processing a diverse and conflicting user workload. Greatly increased capacity of modern computer equipment, complexity of application systems, the growing number of user departments served, the extensive use of multiprogramming systems, the development of online applications and remote job entry systems, all have greatly increased the amount and complexity of work processed through a single computer. Complexity is increased when computers are interconnected tightly through shared memory, loosely through a common job queue (e.g., JES3) or through communication networking. From a control point of view, it is the objective of operations management to organize and supervise this complex operations environment in a manner that will insure satisfactory accomplishment of user service objectives. This includes a reasonable understanding of the user workload and service objectives to effectively schedule the system's resources (man and machines) and to insure that proper measures are taken for data security. Through the operation management function, the appropriate operations data for analysis and reporting should be made available. For example, data describing the mixture of jobs (batch/online/both) during various periods of the day is required to assess the peak periods. Accurate and timely analysis and reporting of data is crucial to improving or maintaining the required control over the operations environment.

The critical part that operations management plays in the capacity planning effort is not always perceived by many people working in the capacity planning area. A major part of understanding how a computer system actually operates (workload characterization, job scheduling, resource consumption, etc.) will be found in the operations area. It is obvious that operation management can be a source of lost processing hours and reduced resource consumption. This will manifest itself in a degradation of user service or a loss in system capacity. But, from a scheduling point of view, which is discussed in greater detail in section 1.5, the system's capacity can not truly be understood until it is clear what user work is accomplished during various periods of the day, the resources required and their consumption (utilization). Also, it should be understood what time periods of the day can be made available for growth in existing applications, workloads shifted from

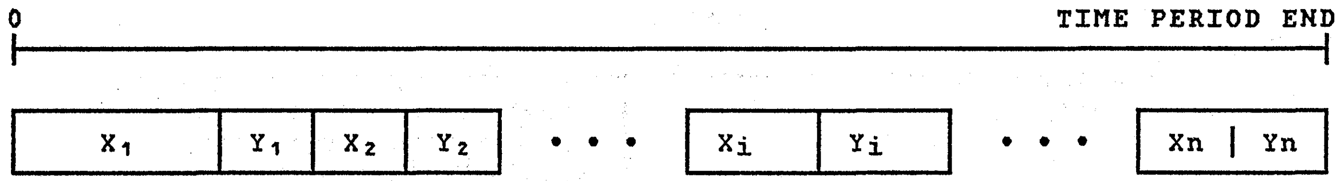
other computing system or planned new applications. Obviously, many other aspects of operation management are key to understanding a system's capacity which are not specifically brought out in this section. In the section on the implementation of the capacity planning process other aspects will be discussed.

1.4.5 Availability Management

Today's computer systems include many inter-related hardware components, a large number of software packages (system and application), complex teleprocessing networks and various levels of skilled personnel. In the event of disruptions (hardware, software, people) to the normal system operation, it becomes necessary to establish switchover and recovery methods, backup and failure analysis procedures. The complex environment created by data processing installations today makes it more difficult to meet stringent system availability requirements. In this environment, system or resource availability has many implications, whereas, the real "bottom line" of availability management relates to the end user. The primary question to be addressed is, what is the application (batch, on-line, interactive) availability observed by the user? There are many factors that affect the availability of a given end user's application, several are listed below and in Part 2 of Figure 3.

- system hardware
- system terminals
- system software
- program products
- application programs
- system operation
- communication facilities
- support facilities (e.g., power, air conditioning, etc.)
- etc.

As pointed out in section 1.4.1., system capacity is directly related to end user service. Application availability is also very crucial to an understanding of system capacity. As a user, it is immaterial that a CPU is providing 99 per cent availability if his on-line software (e.g., IMS, CICS, etc.) is providing less than adequate availability.



X_i - one time increment (i) of operation before application failure
Y_i - one time increment (i) to repair application failure

n - Total number of "Xi & Yi" time increments within time period of scheduled operation

$X = \sum_{i=1}^{n} X_i$ = Total time application available

$Y = \sum_{i=1}^{n} Y_i$ = Total time application not available

$$MTBF = \frac{X}{n} \quad , MTTR = \frac{Y}{n}$$

MTBF - Mean time between failures

MTTR - Mean time to repair

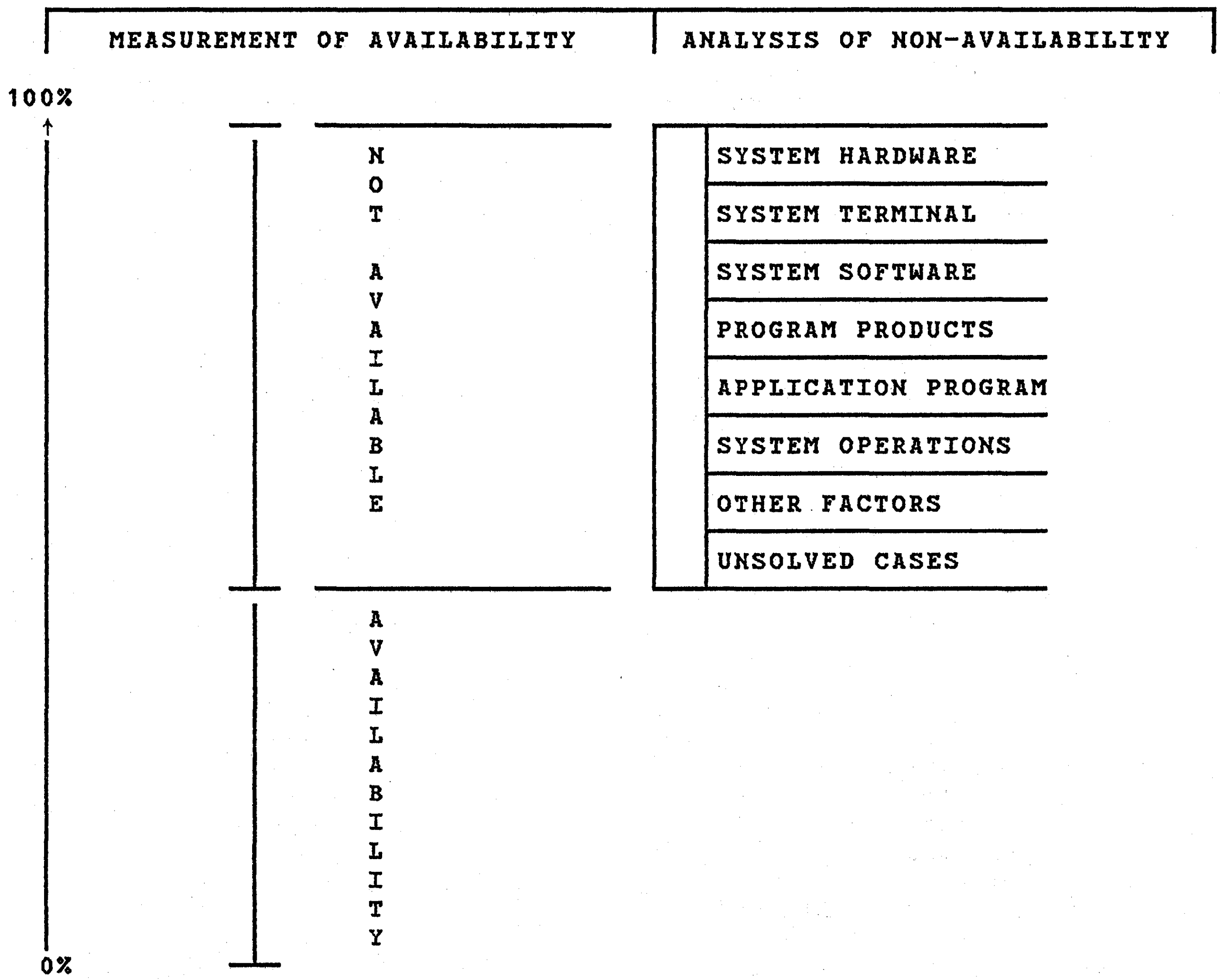
$$\text{Availability (\%)} = \frac{MTBF}{MTBF + MTTR} \times 100$$

$$= \frac{X}{X + Y} \times 100$$

Not available (%) = 100 – Availability (%)

APPLICATION AVAILABILITY (PART 1)

FIGURE 3



APPLICATION AVAILABILITY (PART 2)

FIGURE - 3

Obviously, user service will suffer just as much as if it were a hardware availability problem. The same would be true for time lost because of poor operation practices by the computer operators. In all cases, the user's scheduled work is not being accomplished.

An understanding of the quantification of application availability might be best understood by referring to Figure 3. The applications being addressed might be batch, on-line or interactive. As viewed by a user across some time period (day, week, month, etc.), his application is available (X) or not available (Y). Then, as shown in Figure 3, it is possible to determine the per cent of application availability across this period. The periods of non-availability would then be analyzed to improve overall system availability.

For the purpose of improving or maintaining a satisfactory level of application availability, availability management includes two distinct areas,

- Systems designed for availability

- Availability management practices

System availability is not a question to be addressed only when it is found that application failures is affecting user service. It should be addressed in the design and configuration of new hardware and software systems as well as the design and implementation of operations practices. Many current products (hardware and software) are now offering designed-in facilities to be used by the system designer to improve system availability, (e.g., multiprocessing capabilities as well as other redundant hardware features). The major concern in the design of systems for availability is the trade-off between costs and the required availability goals [14].

With regard to availability practices for systems already in operation, one is concerned with studying existing practices in the installation, performing outage analysis and recommending techniques or actions to improve availability. These practices are well known throughout the computer industry and much has been written on the subject. [5,6,7,8].

1.4.6 Network Management

Recent technological advances are allowing the data processing industry to move into an age of remote information processing. An industry heavily oriented to local batch processing and manual transport of data to the central processing facility has allowed information

processing and data storage capabilities to be distributed to remote user locations. This requires the design and implementation of telecommunication networks. This technology has paved the way for many new applications as well as the redesign of batch applications for on-line use. In many installations, the fastest growing portion of the workload is the teleprocessing (TP) work. Large complex telecommunication networks have developed to handle this TP workload and this has presented DP management with many new and difficult problems. Within this environment, a process termed "Network Management" has emerged to aid the DP manager in the management of these complex networks.

The network management process is primarily concerned with the following functional areas,

- Project planning and system definition
- Network design
- Network testing and installation
- Network operation and monitoring
- Problem detection, tracking and resolution

This management process provides for the structure, integration and direction of these functions. Obviously, there is a great deal of overlap among these functional areas.

The planning and definition phase is a very critical function in that it provides the foundation or base for all subsequent efforts (design, testing, installation, etc.). Under this function, the goals of the project are determined, the organization's potential to accomplish these goals is established, the problem is defined and possible solutions considered. Obviously, many other factors must be considered in the planning process, this is only intended as a brief introduction to the subject. For a much more detailed discussion on this functional area and the others briefly discussed below, please refer to reference 9.

In the network design phase, the hardware configuration and supporting software is determined. The various network components must be functionally evaluated and selected. Performance and availability criteria will be established. The required training programs must be clearly defined. After design specifications are determined, simulation techniques may be used to verify certain general design requirements. It should be understood that the critical factor in the network management process is that the proper plans and controls (reports and reporting procedures) are in

place to integrate the design process with all the processes preceding and following it. For example, it is critical that class outlines and documentation for the training process reflect system design that is ultimately implemented. Hence, it is very reasonable to develop training materials during the design process.

In network testing and installation, planning and controls are very critical for success. Testing procedures must be established outlining order of test execution, test stressing techniques, etc. Test data and terminal scripts must be created. A procedure for test evaluation and analysis must be defined. Within this evaluation, criteria for test success or failure must be determined. Also, provisions for test plan modification should be provided as needed. Network installation is concerned with conversion from an existing system or implementation of a totally new system. This process should be accomplished slowly and in a segmented fashion by implementing the simple portions of the system first and progressing to the more complex phases. For example, if many terminals are to be installed, the process of installing one or two terminals first, then a gradual build-up over time would be preferred to going on-line with all terminals simultaneously. For an existing application, a period of parallel operation must be planned. Also, maintenance procedures must be clearly outlined. During this phase and all others, the proper documentation must be developed and updated as required.

When the network is placed in operation (production), the measurement tools required for monitoring and control of the network must be installed and should have previously been checked out. Tools are needed for monitoring performance (line loads, response times, etc.) and monitoring data errors and equipment malfunction. This is the phase that integrates so closely with an installations on-going capacity planning efforts. Capacity planning is very concerned with the workload processed by the network. The kinds of service end users are being provided. If the service is not adequate, "Where is performance being degraded, at the host, lines, terminals, etc.?"

It is proper monitoring and tracking of the system's operation that will provide the needed feedback for design enhancements, understanding of current workload and its growth, understanding of new workloads and how to size them and how to improve performance predictions.

One of the key elements to the successful operation of a communication network is the logging, detection, tracking and resolution of problems encountered by the end users. The problems are encountered by the end user but the cause

and resolution is not always under the control of the Communication Manager.

1.4.7 Data Base Management

The data base management function is concerned with the management of all the automated data made available to an organization for application program processing. There is no distinction made as to the medium of data storage (card, tape, disk, etc.). Whereas, there is a definite distinction made beteen an integrated and non-integrated data base management system. Briefly, an integrated data base management system [10, 11] provides a collection of interrelated data stored together to optimally serve one or more user applications. The system objectives are to reduce physical data storage requirements (file integration), eliminate processing redundancy, provide application program independence, etc. A non-integrated data base, which has normally evolved with the application growth over the years, provides, in many instances, separate data files for each application. This normally means that an enormous amount of data redundancy will build up in each data base. Such a system requires synchronization of files for maintenance as well as large amounts of physical storage space. The following discussion is concerned with data bases in either of the formats discussed above, however, many organizations where large volumes of data are becoming unmanageable are moving toward an integrated data base management systems as a solution.

Some of the basic functions required in managing a data base are listed below. Provisions will be made for:

1. The accuracy, security and auditability of the data.

2. The design of the data base organization scheme.

3. Data base measurement and improvement.

4. Definition of data recovery strategies.

5. Data base maintenance.

6. Technical support for applications development.

Many of the functions outlined above are self explanatory but others require some amplification. The auditability of the data base is concerned with providing as a part of the data base design easy access to date required by the auditors. Audit requirements are a very necessary part of data base requirements but have not received adequate DP attention. With respect to data base design, the concern is

for proper space management and access method, various views of the data, logical content, etc. Measurement tool technology is one of the primary concerns in trying to assess performance in any part of the computer system. Hence, data base measurements are a very critical function and current tool technology will affect one's ability to measure certain performance parameters effectively.

From a capacity planning perspective, the data base management process will provide an input to I/O performance analysis and prediction. Monitoring and tracking information will be used to verify predicted hardware (channels, control units, tapes, DASD, etc.) and software (VSAM, ISAM, etc.) performance requirements. Also, the interface between capacity planning and data base management may indicate specific design alterations (data set placement, data set consolidation, etc.) are necessary to improve current performance.

1.4.8 Summary

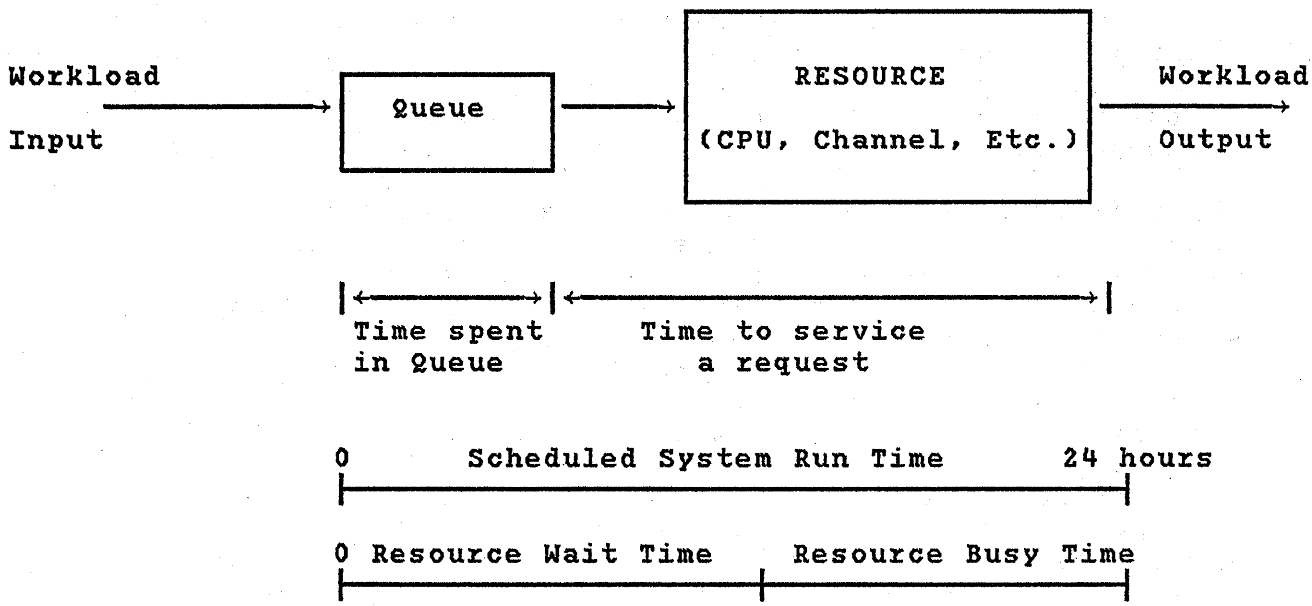
Section 1.4 contains a brief introduction to the installation management process and some of the primary functions required. It has shown how these functions interface with the capacity planning process. As this bulletin is written to cover the implementation of the capacity planning process, the following sections will, in greater depth, outline how closely capacity planning interfaces with the installation management process in other areas.

1.5 What is The Capacity of a Computer System?

In discussing the capacity of a computer system, it is necessary to differentiate between the capacity of a resource (Figure 4) and the capacity of the computer system (Figure 5). Although many installations indicate they are unable to set or establish specific user service objectives or that such requirements are not practical in their environment, I submit that it will be very difficult to establish or understand the capacity of a computer system until user service objectives are clearly defined. As part of the computer capacity question, the criticality of user satisfaction will be discussed in the following paragraphs.

The primary indication of the capacity of a resource is the time it requires to complete a request for service (Figure 4). Hence, in the scheduled operation of a resource over some period of time, the summation of all the completed service requests is equated to the resource busy time. With respect to the measurement tools being used today, it is the resource busy time or resource utilization that is an indicative of a resource's overall capacity. When the busy time of a resource is equal to its scheduled run time, the total capacity of the resource has been consumed (no wait time component). In essence, everytime a request for service has been completed there is another request to be serviced. In queueing theory, it would be said that the resource is 100 per cent utilized. This state of a resource gives rise to very large queue sizes (i.e., large numbers of requests waiting to be serviced).

The capacity of a resource as outlined in the previous paragraph might also be thought of as the independent or stand alone capacity. Looking at the resources purely on a box by box basis, each can be monitored with 100 per cent utilization as a capacity constraint. But, to understand the capacity of a resource, as part of a computer system, it is necessary to understand capacity other than as an independent concept. Because, in most instances, a resource does not sustain a continuous busy state over the scheduled period of operation. Normally, user service degrades to such a level before saturation (100 per cent utilization) of the various resources that other alternatives (new hardware, off load work, tuning, etc.) must be taken to improve system response.



$$\text{Resource Utilization} = \frac{\text{Resource Busy Time}}{\text{Scheduled System Run Time}}$$

RESOURCE CAPACITY

FIGURE 4

In attempting to understand resource capacity, you need to forecast or predict the amount of additional work a resource might perform with respect to the amount of wait time experienced during its scheduled time of operation. Obviously, from the previous discussion, the key to this understanding is not the fact that a resource is capable of 100 per cent operation. This knowledge must come from an understanding of the user service requirements and a knowledge of work to be performed.

The capacity of a resource may be viewed as having a potential of 100 per cent utilization. But, in most practical instances, a resource will not realize its full potential which is constrained by user service satisfaction. Hence, the capacity of a resource will vary among installations as well as within an installation depending on the time of day. The upper limit on resource capacity is the utilization (busy time over scheduled run time) above which the given resource becomes a bottleneck and degrades the response/turnaround time so that the user service objective can no longer be met. A channel within a computer system with an average utilization above 35 per cent will elongate response time in an interactive system. TSO users may find their response time degrades to unacceptable limits. Reference 1 provides a more detailed discussion on resource capacity.

For capacity planning, a computer installation should be viewed as a system of resources. In other words, the capacity to be analyzed must be that of the total computer system. It is germane to the subject to know that a given CPU can execute "X" million of instructions per second (MIPS) or that a channel is capable of transferring "Y" bytes per second, but the critical issue is, will the combined performance of my resources provide satisfactory user service in terms of response/turnaround time. Therefore, from a capacity planning point of view, the capacity of a computer system is determined principally by four factors (Figure 5), where user service is the most critical indicator. This is not to minimize the importance of characterizing the workload to be processed or understanding the independent capacities of the various resources.

USER SERVICE REQUIREMENTS *

- RESPONSE TIME
- TURNAROUND TIME
- EARLIEST START TIME
- LATEST END TIME

AVAILABILITY

- AVAILABLE SYSTEM HOURS
  - HARDWARE
  - SOFTWARE
  - USER PERCEPTION
- UNAVAILABLE SYSTEM HOURS
  - MAINTENANCE
  - UNSCHEDULED IPL'S
  - RERUNS
  - ETC.

WORKLOAD

- TRANSACTION LOAD
  - TRIVIAL
  - MEDIUM
  - COMPLEX
- REQUIRED RUN TIME
  - 10 am - 2 pm
- NUMBER OF TRIVIAL BATCH
- NUMBER NON-TRIVIAL BATCH
- ETC.

RESOURCE CAPACITY

- % BUSY
- AVG. QUEUE SIZE
- % AVAILABLE
- ETC.

* SYSTEM CAPACITY IS DETERMINED BY CLEARLY SPECIFIED USER SERVICE REQUIREMENTS BASED ON WORKLOAD

SYSTEM CAPACITY

FIGURE 5

As pointed out in Figure 5, the principle factors to consider in developing an understanding of the capacity of a computer system are:

- User service requirements
- Available system hours
- Workload characterization
- Resource Capacity.

In most computer modelling work done today, the data requirements (top of Figure 6) for transactions processed (i.e., paging rates, resource utilizatons, etc). are normally not enough to adequately access the capacity of a computer system. This is not to say that these are not important parameters but there are other considerations (bottom of Figure 6) which in many cases have been overlooked in system capacity modelling effort. For example, understanding a system's capacity during specific time windows (e.g., 8:00 AM - 11:00 AM or 1:00 PM - 4:00 PM) versus using average daily parameters of transactions per second. CPU utilizations can be a very critical aspect of the system capacity analysis. As shown in Figure 7, the CPU utilization of a computer installation is plotted over a 24 hour period. There is obvious computer resource capacity available as indicated by the many "valleys" on the graph. Assume that this installation's resources are relatively well tuned and no "bottlenecking" of resources is restricting the performance of the CPU. Also, assume that all user service requirements are being met during the peak periods from 8:00 AM to 11:00 AM and 2:00 PM to 7:00 PM where the CPU is sustaining 100 per cent utilization. If it is known that the work being accomplished during the peak period cannot be shifted to other machines or different times of the day, then for all practical purposes the computer is out of capacity during these time windows regardless of what average values or modelling will indicate. Reason would indicate that capacity is a function of the time of the day, week or month and that scheduling of the workload bears heavily upon understanding a system's capacity. It is these kinds of considerations that begins to truly address the critical problem of system capacity and workload characterization. Until you understand how your installation's workload is characterized, capacity planning will be a very difficult task. Briefly, workload characterization is understanding the DP environment (i.e., the frequency of requests for computer service, who is making the requests, the amount of resource service required and when). These and other factors will be discussed in more detail in Section 2.0 (Capacity Planning Implementation).

CURRENT COMPUTER MODELLING

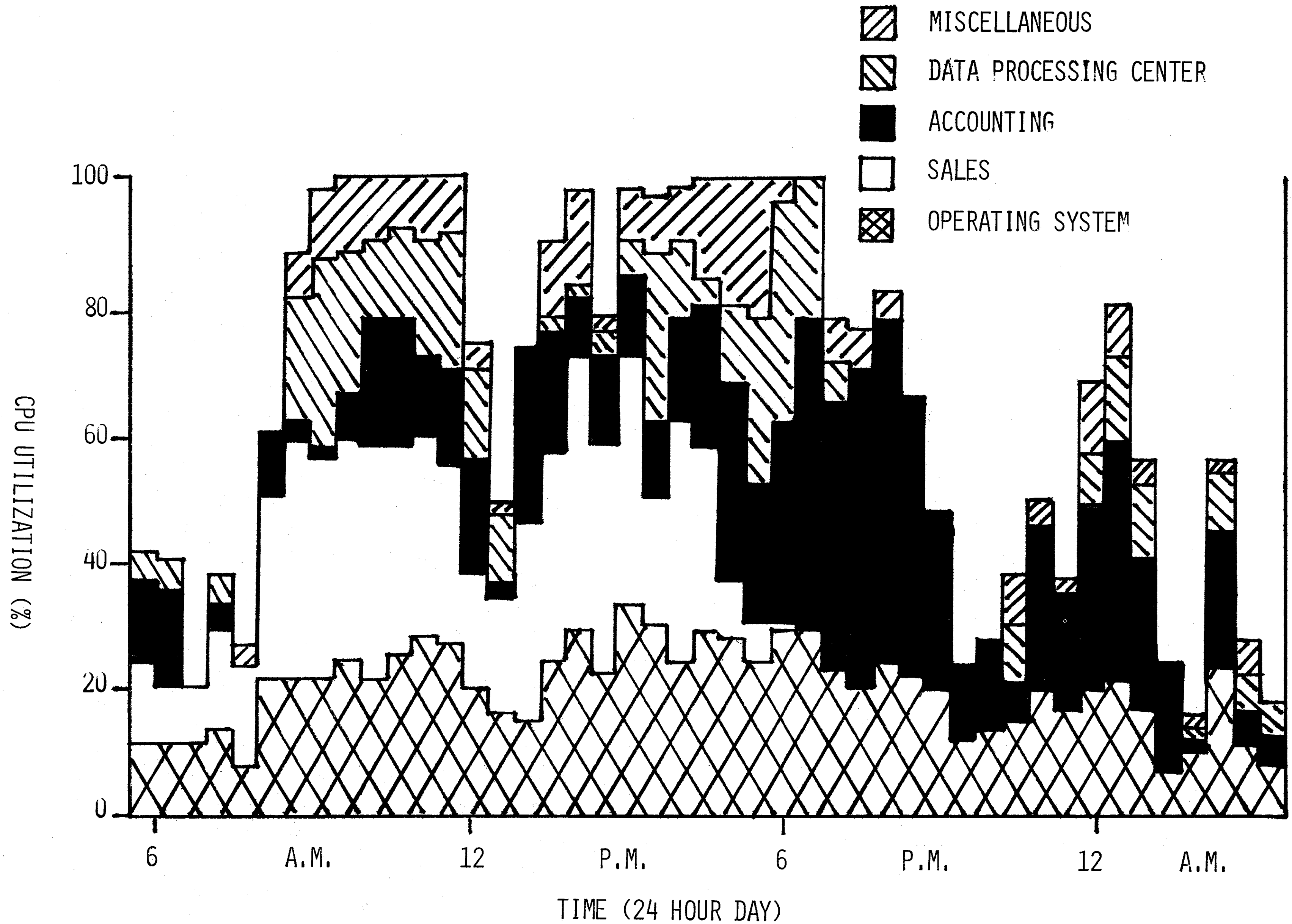
- O WORKLOAD
  - O JOBS/HOUR
  - O TRANSACTIONS/SECOND
- O RESOURCE UTILIZATIONS
- O RESPONSE/TURNAROUND TIMES
- O MAIN MEMORY USE
  - O PAGE-IN RATES
  - O ETC.

OTHER CAPACITY PLANNING CONSIDERATIONS

- O SPECIFIC TIME WINDOWS
- O PREDECESSOR JOB REQUIREMENTS
- O SPECIAL I/O REQUIREMENTS (FORMS)
- O NUMBER OF TAPE MOUNTS
- O UTILIZATION OF TAPE MOUNTERS
- O OFFLINE PRINT/PUNCH WORKLOAD
- O ETC.

DATA REQUIREMENTS

FIGURE - 6



CPU CONSUMPTION BY APPLICATION.

FIGURE 7

Although system availability, workload characterization and resource utilization are very important factors in understanding computer system capacity, the key to establishing current and future capacities is the user service requirements. Without a firm fix on user requirements, the capacity of a computer system will be very nebulous and in effect float between many different values as system requirements change. For example, before moving from the capabilities of one computing system to that of greater capability (e.g., from a system driven by a 3032 to one driven by a 3033), the service (response/turnaround times) being provided the critical batch and online applications on the current system should be established. In this light, the future capacity requirements are forecast with these performance parameters (response/turnaround times) as a base. It is usually decided what new applications are possible with the new configuration and the growth to be accommodated in old applications. One of the critical factors used in determining whether the new configuration will live up to its expectation is the adherence to the old service requirements. Capacity is not normally allocated for users of current applications to move to a drastically improved service. Obviously, their service will be improved as part of the migration, but a large improvement in user service performance normally accompanying a new CPU tends to leave a user with a false impression of his service requirement. A user may feel he wants to maintain his drastically improved service, even when he is impacted by the implementation of a new planned application. What this means is that a user, not aware of a specific service objective planned for his application, will reject a plan to return him to some lesser service which is the service he was happy with on the old configuration. This may mean a large portion of the new capacity planned for new applications will be lost. The capacity of a system is caught up in the negotiations and agreements on user service requirements between DP Operations and the user community. The implications here are that the user is well aware of the service contracted, it may drastically improve for any number of reasons; but, when it returns to the objective value his expectation is maintained.

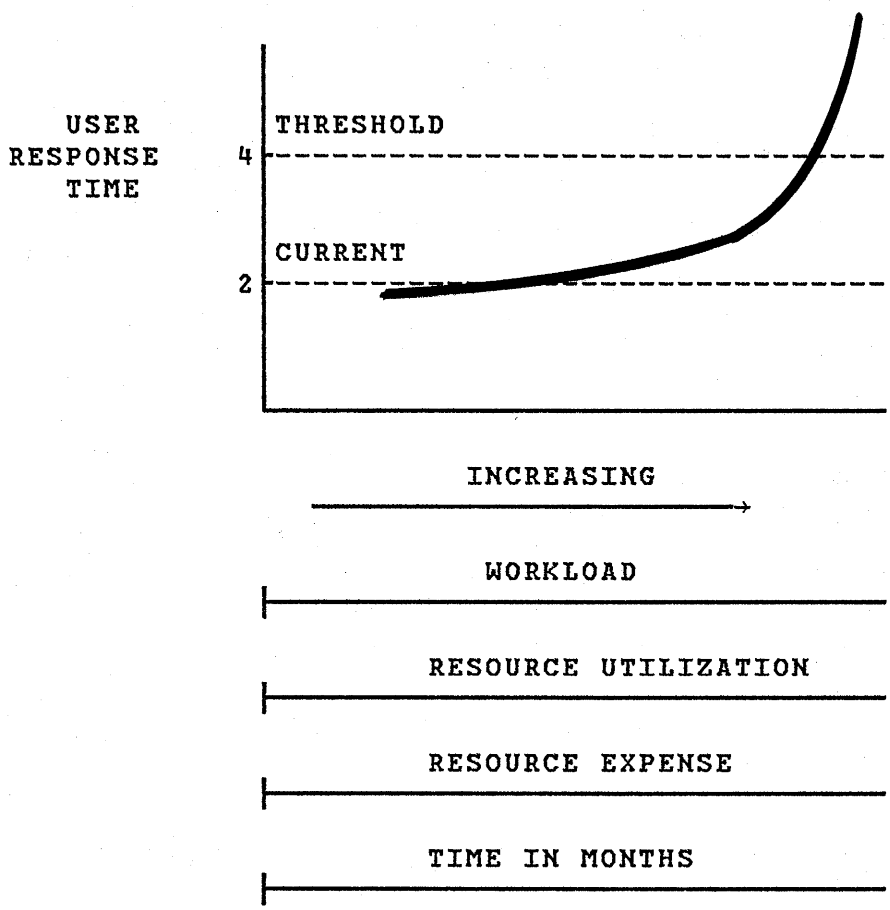
A key concern being expressed by many users over the past several years is consistency of service rather than an improved service. They are requesting that the service once established be maintained. This applies primarily to an on-line environment where certain work procedures are developed around a particular user service (response time). When the response time values change significantly (improves or degrades), procedures can be greatly impacted.

In a discussion, as given above, the following question arises., if capacity is driven by user service objectives,

"What is the appropriate service for online COBOL programmers (testing), clerk inquiry application, etc.?" It would be nice to put a value on each application, 4 seconds for the programmers, 3 seconds for the inquiry, etc. This is a very difficult problem to solve but the following discussion of one customer's approach to solving this problem might be helpful.

To establish reasonable user service requirements within their installation, an organization implemented the following procedure. During a period of time when capacity was not a real problem and with little or no service complaints, each user's service was measured. These service levels were taken as indicative of satisfied users and these values were discussed with each user. Obviously, this would not mean every user, but the environment should be analyzed and classes of users selected by a selection procedure. The number of user classes addressed must be kept to a manageable size. In disclosing the level of service (response/turnaround time) being provided a given user, the primary question was the satisfaction being perceived. If the user was satisfied and agreed that their current value of service was adequate for their application, this value became the objective or requirement. However, if the user was not satisfied, a request for improved service was discussed with operations. If operations felt such a request was reasonable, the change would be implemented. But, on the other hand, unreasonable requests were reviewed and discussed with the user. It might require several iterations between operations and the user until agreement could be reached. It was the final agreed upon value that became the user's service objective. This was a problem where the user shared in the solution. Establishing user service objectives is not only a technical question, but there are many psychological, political and economical factors involved.

Another question arises concerning system capacity and user service objectives, "How is future capacity planned using response and turnaround times?" There are several methods available in which a computer system's workload (Transactions/Sec, Jobs/Hour, Etc.) is increased and the change in response or turnaround time is predicted. From a theoretical point of view, queueing analysis or discrete simulation may be used. A model is developed and various known values of load are used as input. Knowing the current user service being provided and having some threshold value (Figure 8) which can not be exceeded, the model workload is varied until the threshold is reached. At this load, which is indicative of a period of time in the future, resource utilizations may be noted from the model and the expense of relieving any resource bottlenecks can be evaluated.



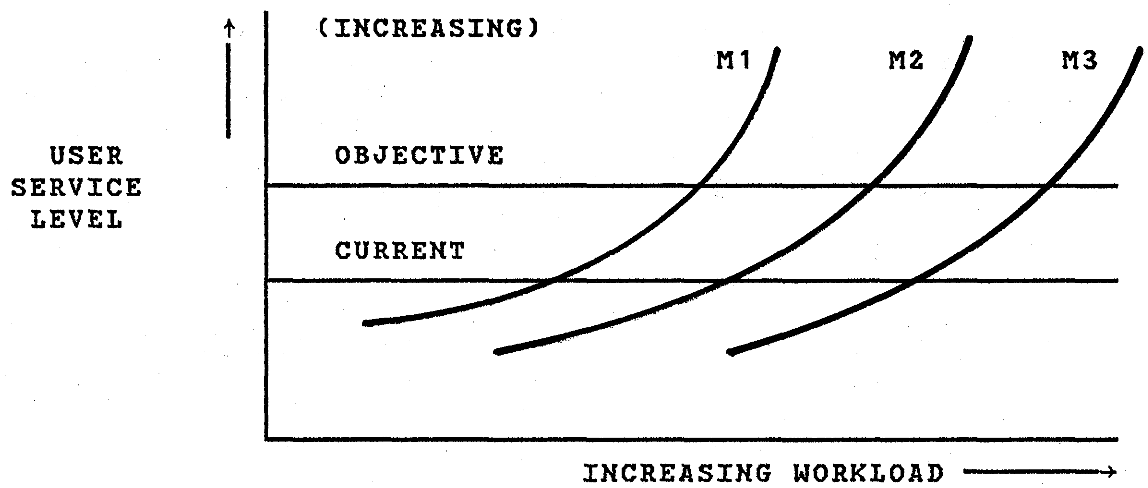
SYSTEM CAPACITY (THEORETICAL)

FIGURE - 8

User service versus load curves (Figure 9) may also be empirically generated from benchmarking or actual historical data gathered by an installation. The usefulness of empirical curves are greatly enhanced by describing the "operational mode" as accurately as possible. Curves for three systems with operational modes M1, M2, and M3 are shown in Figure 9. Empirically derived curves may be used in much the same fashion as theoretical curves, a known current and a future predicted load may be plotted on the horizontal axis (Figure 9) and the associated user service noted on the vertical axis. The point of critically, the "knee" of the curve, would be noted and the expected time in months before it would be reached could be analyzed.

1.6 Capacity Planning is a Process

A computer installation is a very dynamic environment where you have changing hardware, software, techniques, and people. In this bulletin most of the examples will refer to the MVS operating system, specific pieces of hardware (3033, 3350, etc.) and current software monitors (RMF, SMF, etc.). However, because of the continually changing DP environment, a methodology developed for capacity planning must be as independent as possible of any specific products (MVS, RMF, 3033). Basically, all products should be viewed as inputs to the capacity planning process. For example, the change in the overall capacity planning process should be minimal when it is necessary to change from one operating system to another (i.e., MVT to MVS). Obviously, this is easier said than implemented. I am well aware of the difficulties in moving to a new operating system when capacity planning techniques are relatively well defined under another operating system. In one instance, an installation had developed certain empirical curves (Figure 9) for capacity planning under the MVT operating system. The data for these curves had been gathered over time, analyzed and validated. Since these curves were an integral part of the capacity planning process and directly related to the MVT operating system, the question, when one plans to move to a virtual operating system (e.g., MVS), becomes, "Will the current curves be representative of my system in the virtual environment?" If the answer to this question is "NO", then, "What adjustments must be made to these curves"? The answer to these kinds of questions in a changing environment are not simple. In this particular case, a simple transformation for the curves is not possible. The transformation would probably be gradual with a collection and analysis of data in the virtual environment, then comparisons and adjustments to existing curves. All of this would take place over a period of time and cause a certain amount of disruption to the capacity planning process. The purpose of this subsection is to bring out the necessity for good planning in implementing the capacity planning process to minimize as much as possible disruptions due to product changes.



OPERATING MODE CHARACTERIZATION (M1, M2, M3)

- O CPU SIZE
- O OPERATING SYSTEM
- O NUMBER OF INITIATORS
- O MEMORY SIZE
- O NON-SWAPPABLE IMS MESSAGE PROCESSING PROGRAM
- O NUMBER OF TAPE MOUNTERS
- O ETC.

SYSTEM CAPACITY (EMPIRICAL)

FIGURE - 9

2.0 Capacity Planning Implementation

2.1 Introduction

Systematic capacity planning is possible and it is successfully being used today. The purpose of this section is to outline a procedure for the implementation of the capacity planning process. This section is based on the capacity planning work done at the Washington System Center over the past three years covering national as well as international customers.

From an implementation point of view, the computer market place appears to be broken into three categories of customers involved in the development of capacity planning efforts. The first category, which might be termed the Type-1 Customer, is the largest set (Figure 10). This customer set is considered to be in the initial stages of the development of a systematic approach to capacity planning. He has little or no experience in this area and is looking for specific directions. The customers in the second category, termed the Type-2 Customers (Figure 11), have experience with capacity planning techniques, however, they are receiving very limited results. By virtue of their capacity planning involvement, they are asking more specific questions about measurement tool inconsistencies and other performance parameters. They are looking for more specific capacity planning direction than the Type-1 Customer. For example, why track a specific parameter that appears to be completely random, therefore, it can provide no useful information for their capacity planning efforts. The randomness of such a parameter may not be a characteristic of the system but of the measurement tool. Customers in the last category, termed Type-3 (Figure 12), are by far the smallest set. They have been tracking and gathering system performance data over the years, establishing guidelines, and developing "rules of thumb" by which to manage their systems. They are receiving very good results from their efforts. They are not necessarily seeking capacity planning direction and consultation but are more in a mode of information exchange. They are seeking specific direction during a time of system change as from MVT to MVS, UP to MP or AP, etc. In most instances, these customers are using simple analysis techniques (guidelines, linear analysis, etc.), but are searching for methods to refine their analysis procedures (e.g., queueing, simulation, etc.).

TYPE-1 CUSTOMER

- Initiating Capacity Planning Efforts
- Employing a Number of Measurement Tools
    - Software & Hardware
- Staff Not Too Strong in Measurement and Analysis Area
- Looking for Basic Capacity Planning Direction
    - What Measurement Tools Should be Used
    - What Data Must be Gathered
    - How do I Segment my Workload
    - What Types of Analysis Techniques Should Be Used
    - What Report Writers Should be Used
    - What Reports Should Be Created and How Should They be Formatted
    - Who Within the Installation Should be the Recipient of The Various Reports
    - Who should manage the Capacity Planning effort and where should it reside in my organization

CHARACTERIZATION OF TYPE-1 CUSTOMER

FIGURE 10

TYPE-2 CUSTOMER

- Initiating Capacity Planning Effort
- Employing A Number of Measurement Tools
    - Software & Hardware
- Doing Good Gross Performance Analysis
    - On A Single Resource Basis (CPU, Channel, Etc.)
- Trying Various Capacity Planning Approaches, Getting Very Limited Results
- Asking Very Specific Questions Concerning Measurement Tool Inconsistencies
- Looking For Specific Capacity Planning Direction
- Asking Specific Questions About Vendor Product Line
    - Mass Storage
    - Shared DASD
    - TP (Controllers, terminals, etc.)
    - Etc.

CHARACTERIZATION OF TYPE-2 CUSTOMER

FIGURE 11

TYPE-3 CUSTOMER

- Capacity Planning Effort in Place
- Have Been Tracking and Gathering System Data For Many Years
- Having Good Results (No Real Complex Models)
- Seeking Capacity Planning Direction During Time of Change (MVT-MVS, UP-MP/AP, MVS-MVS/SE, Etc.)
- Asking Very Specific Questions About Vendor Product Line
- Currently Using Simple System Analysis Techniques, Looking For a Better Way
    - Queueing Models
    - Discrete Simulation
    - Benchmarking
    - Etc.
- Exchanging Capacity Planning Expertise

CHARACTERIZATION OF TYPE-3 CUSTOMER

FIGURE 12

In this section, a total system for capacity planning is described. Where the elements of the system are:

- People
- Organizational Structures
- Hardware/Software
- Measurement Tools
- Predictive Tools
- Data Requirements
- Reports/Reporting Process

The primary objective of this section is to define a base process which may be modified to meet the needs of a particular DP installation. In many instances, the organizational structure outlined will require modification or the measurement or predictive tools described are not available and a suitable substitute must be used. Also, the intent of this development is to be as simple as possible and still provide satisfactory results for "practical capacity planning". The term "practical capacity planning" will become much clearer in the following sections.

2.2 The Capacity Planning Process

2.2.1 Personnel Requirements

Before a capacity planning effort is initiated, the people required to staff the project must be selected. The areas providing the principle input for the development of the capacity planning process are:

- Operations Department
- Systems Programming (MVS, SVS, VS1, etc.)
- Applications Programming (BATCH, TSO, IMS, CICS, etc.)

Therefore, the people selected to initiate the project should have some experience in these areas because development of the capacity planning process requires a close coordination with each area outlined above.

Obviously, there are no hard and fast rules as to the number of people required to begin the process or even whether a specific group of people should be set aside to perform the function. The paragraphs that follow will illustrate the experiences from several accounts who initiated capacity planning efforts in their installations.

In most cases to date, organizations are selecting one or two people and establishing a new department called the capacity planning department. One person is selected to head the project and assume primary responsibility for the department activities. The individuals chosen are experienced DP professionals and have a strong background in at least two of the technical areas sited above. This means the capacity planning group will begin with knowledgeable people able to provide a good interface into operations, systems and applications departments. As the implementation process is developed in the following sections, the details of why this expertise is required will become clear. As to whether one or two people are required, it seems that operations and systems or operations and applications expertise is required which would imply that two would be the more reasonable requirement. Obviously, the size of the account and DP staff might dictate that only one full time person is available. With the proper consultation, one person to initiate the capacity planning efforts is not unreasonable. This part of the process is concerned with understanding the current capacity planning efforts and developing a plan for enhancement or a new development. As the plans begin to be implemented, the capacity planning groups will probably require additional people. But, during the time plans are being developed, personnel needs are outlined. Hopefully justification for additional personnel will be no problem.

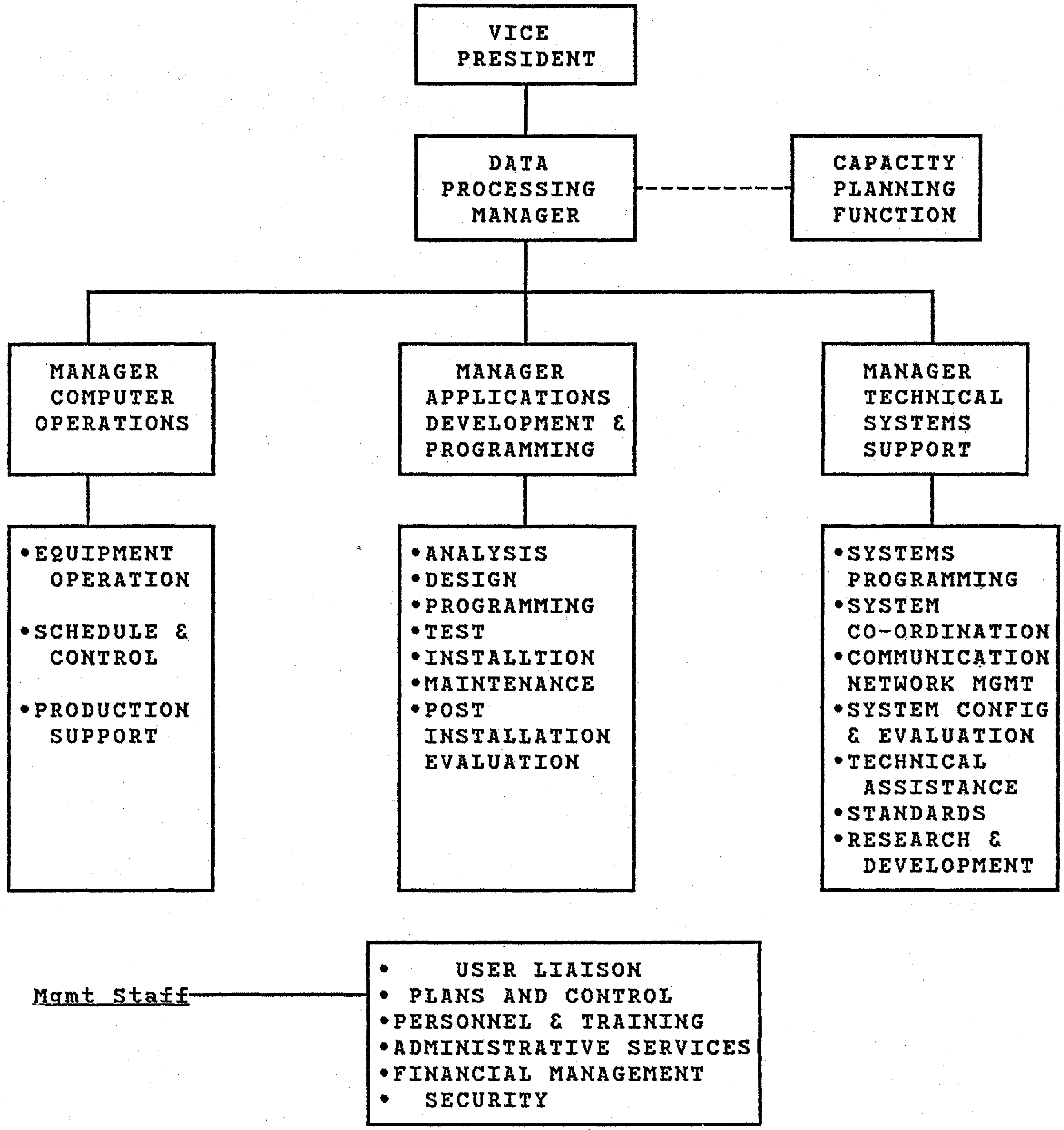
The requirements outlined above, in general, apply to organizations initiating capacity planning efforts. As pointed out in the introduction to this section, there are many installations that are already doing capacity planning and have groups already formed. These groups, in many cases, also have responsibility for system performance evaluation (section 1.4.1) and have grown to five or seven people.

In summary, the personnel guidelines to develop a capacity planning effort are that the process can be initiated with one or two experienced DP people. The experience should be in the operations, systems and applications areas. After the process has been initiated, the growth of this department can be justified and controlled. All indications

are that as the process matures and more elements of the methodology are implemented, the group may grow to between five and ten people. This growth figure will depend on the size of the installation and other functions performed by the group (e.g., performance evaluation, tuning, etc.).

2.2.2 The Data Processing Organization

Having selected the personnel to staff the capacity planning group, the question arises as to the placement of this group in the DP organizational structure (Figure 13). There is definitely no hard and fast rule as to the placement of this group. However, there are a number of good reasons why having this group report to a level of management above the manager of operations, applications development and programming, and technical system support would be very beneficial to the process. In Figure 13 the capacity planning function is shown as a dotted line activity from the data processing manager. This indicates the uncertainty of a direct line management function for the group where it might function in a staff capacity. The uncertainty comes from the fact that this type of structure is not currently being observed in the field. But, there are two primary factors that would make such a structure quite reasonable. First, the insight and information the capacity planning group is expected to have concerning the overall operation of the DP installation (e.g., operations, user community, etc.). It seems reasonable that the DP manager who finds himself having to respond to his upper management more frequently concerning various aspects of the DP installation would want the capacity planning group much more accessible for consultation. The capacity planning group will be very ineffective in developing the process and interacting with other departments (operations, applications development, systems programming, various users, etc.) if their function is not perceived as being strongly backed and committed to by upper management. By reporting to a level of management above these departments (excluding certain users), capacity planning is viewed as a more important function requiring main line support by each DP department. In order to make a decision concerning the placement of this group in your organization, the important thing to consider concerning the mission of this group is that very close integration and co-ordination is required with personnel in the operations, applications development and technical systems support groups.



DATA PROCESSING ORGANIZATION

FIGURE - 13

In several situations where capacity planning projects are still faultering after being in existance for one year, the capacity planning group did not receive the required co-operation from other areas. Many times personnel in other groups are even hostile toward the capacity planning people (i.e., wrestling for certain political powers and recognition). Therefore, the capacity planning group must have the recognition at a high enough management level to be effective.

From past experiences, many DP installations are placing the capacity planning group under the manager of technical systems support. The function will receive a great deal of recognition and support from this parent group. However, operations and applications development and programming may view capacity planning as a secondary activity. Although, it may be perceived as being an important function by these groups, they feel no real line responsibility for its accomplishment. Capacity planning as a fruitful process will fail until the DP organization as a whole views it as a vital function. This means that each group will contribute as a normal part of their activities to the capacity planning process. For example, operations might correlate manually recorded unscheduled IPL (Initial Program Loadings) accounting with certain measurement data reported to the technical systems support group (i.e., for validation purposes).

In summary, the primary consideration concerning the placement of the capacity planning group is for placement to show upper managment commitment if possible. However, if such placement is not possible, upper management must be clear in establishing their commitment to the capacity planning effort.

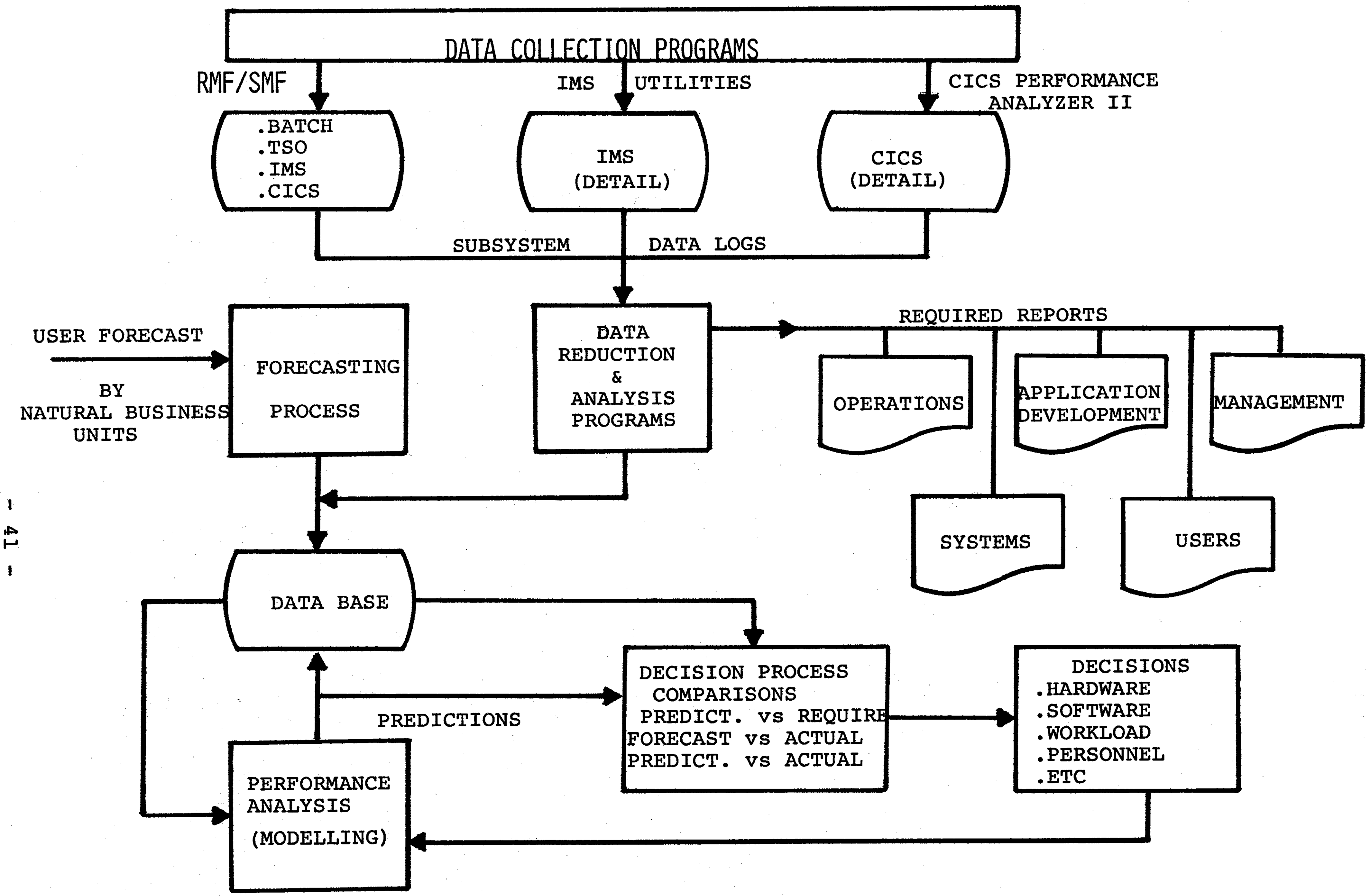
2.2.3 Implementation

Capacity planning is an ongoing process which must be installed and managed by the data processing installation. Consultation can be very helpful but the task of implementation sits squarely on the shoulder of the <u>DP Installation</u>. Capacity planning must be viewed as much more than a data gathering and performance prediction exercise. The process should be viewed as an integration of the following components:

- DP Management
- Technical DP Personnel
- User Community
- Computer Hardware and Software
- Measurement Tools
- Data Collection and Reporting
- Workload Characterization
- System Modelling and Performance Prediction

These components should be systematically structured and controlled to provide an effective capacity planning program. One example of a systematic approach to ongoing capacity planning is outlined in Figure 14. It is the user community that drives the DP installation. As shown in Figure 14, the user workload would be forecast in natural business units (NBU), such as a number of new accounts, number of invoices, etc. This forecast is input to a process designed to convert NBU into data processing units (DPU), such as, transactions per second, jobs per hour, earliest start time, latest end time, etc. This process of selecting NBU's and converting them to DPU's and using the results for capacity planning is still very much an "Art". For example, the best approach for implementing such a process is by trial and error. By analysis and elimination, select those NBU's which appear to be the dominant ones (those NBU's that account for the major portion of the DP workload). Then, implement a plan for tracking the performance of the NBU's (current volumes as well as growth) against the DPU performance. If certain selected units do not appear to be dominant (DPU's are not tracking in any reasonable way with the NBU's), reassess your environment and make other selections. Assuming, as shown in Figure 14, the forecast process is adequate and NBU's can be reasonably converted into DPU's, then current data as well as growth factors over time are input to the data base.

Any capacity planning process requires a family of collectors to collect, reduce and report upon the required performance parameters. Specific requirements will be outlined for the timely collection and reporting of data (i.e., daily, weekly, monthly). Also, a customized set of reports must be defined for each area (operations, systems, application development, users, management).



CAPACITY PLANNING PROCESS

FIGURE 14

Certain data stored in the data base will be used for performance analysis of the computer system, namely,

- Model Development (Figure 58)
- Model Calibration (Prediction vs. Current)
- Performance Prediction (Workload Forecasts)
- Model Validation (Preditions vs. Future).

There are many different modelling techniques available and several are discussed in Section 2.2.4. The three phases of modelling are calibration, prediction and validation (tracking); these are shown in Figure 15 with data selected from an actual modelling effort performed at the Washington Systems Center. The three graphs at the top of Figure 15 depict four load points for an IMS, TSO and Batch environment. The load points are for a given base time (point 0) and three months into the future (Reference Base Time). The workloads are given in values of transactions per second for IMS, interactions per second for TSO, and jobs per second for batch. The second set of graphs in the middle of Figure 15 displays the response times and turnaround time for each load point. This is response and turnaround times at the host system (not terminal values). The corresponding CPU, tape and DASD utilizations are given in the graphs at the bottom of the Figure.

The heart of the capacity planning process, as depicted in Figure 14, is the data base which contains the following data:

- Current and Forecast Workloads
- Current and Historical Performance Data
- Performance Predictions (Calibration and Validation Data)
- Data for Reports

This description of the capacity planning process is only an introduction. The objective of the remainder of this section is to develop in detail a basic structure in which capacity planning may be implemented.

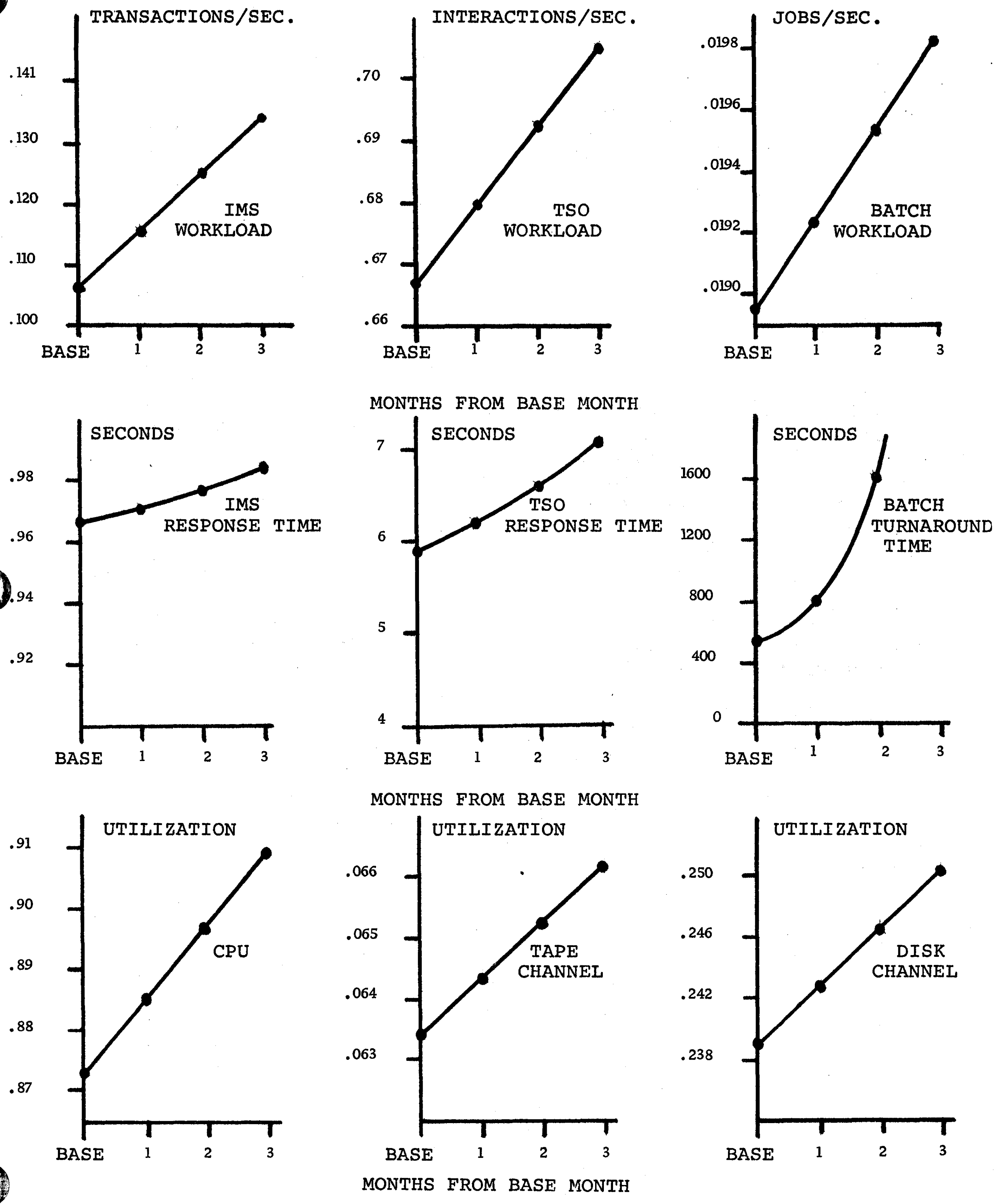


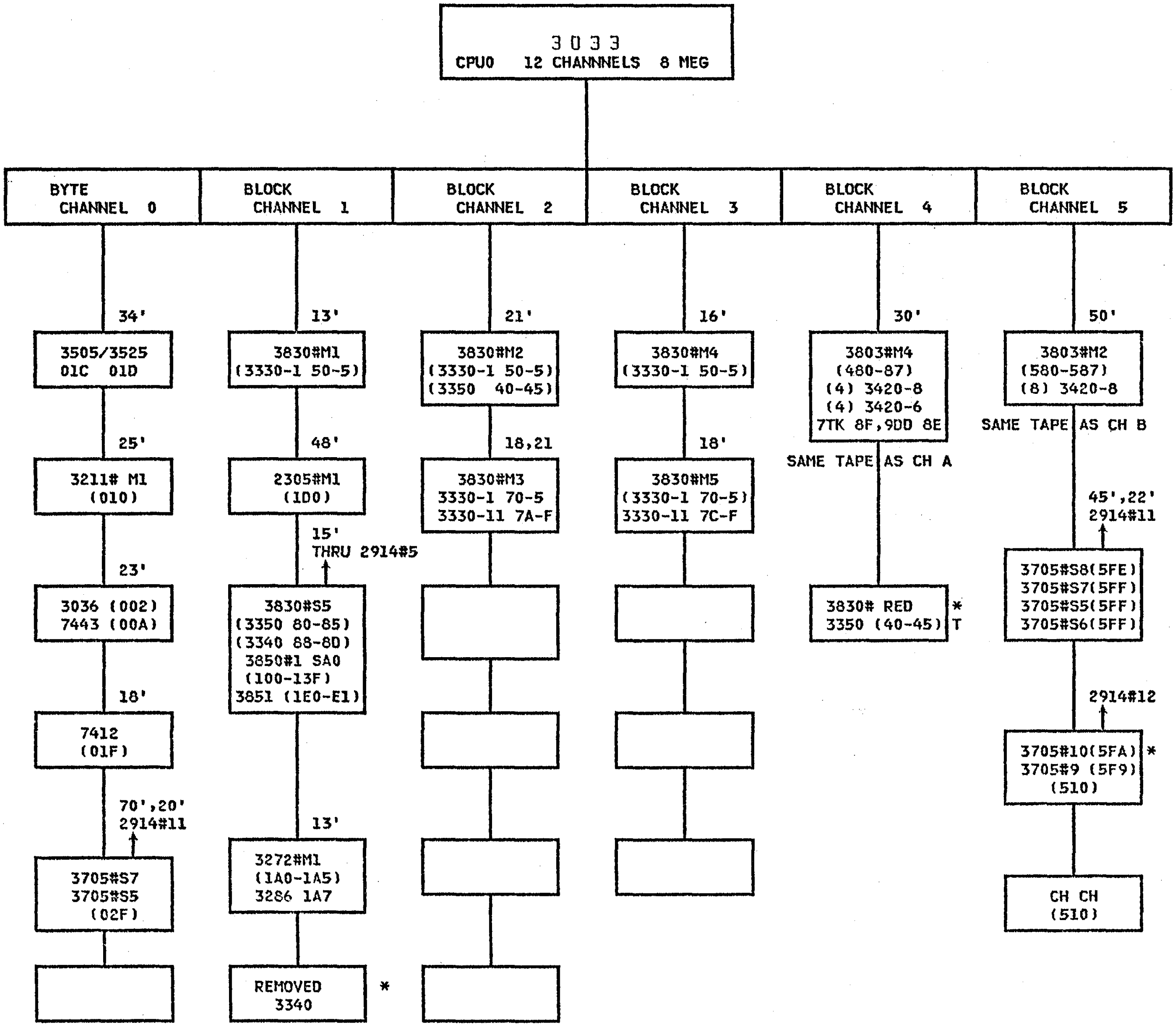
FIGURE 15 : SYSTEM STATISTICS

To initiate a capacity planning effort, there are certain preliminary items required (Figure 16). A clear system configuration diagram is a very important item. This diagram should include hardware identification (CPU, channels, control units, etc.), model numbers, and CPU, channel, control units and I/O device connections. As an example of a configuration diagram, Figure 17 describes a 12 Channel, 8 Megabyte, 3033 Configuration resident in the Design Center at the Washington System Center. It should be noted that the configuration diagram was printed by an IBM 3800 printer. System configuration diagramming has been computerized by the Design Center. In a large DP installation, as the Design Center (7 large computers, 158-3033), configuration changes frequently occur and manual updating can become very cumbersome. Obviously, it is very crucial to the capacity planning process that analysis be accomplished on the most current system configuration. Also, this means that software monitor data will be correctly correlated with the proper device for the user application being analyzed. In configurating hardware, usually an installation will configure portions for each subsystem (BATCH, IMS, etc.). Therefore, within the total configuration it should be understood what resources apply to the various subsystems. Since total configurations as well as configurations for subsystems will change frequently, this must be given particular attention when data is being monitored on a continuing basis. Certain data discrepancies in measurement tool output may be directly traced to a configuration change. In Figure 17, a block will usually contain information on a control unit as well as the attached I/O devices. When certain I/O devices are shared by channels, it is so noted below the block.

- O CLEAR SYSTEM CONFIGURATION DIAGRAM
  - O HARDWARE SIZES AND TYPES
- O SCP TYPE AND RELEASE LEVEL
- O SUBSYSTEMS INSTALLED
  - O JES
  - O BATCH
  - O TSO
  - O IMS
  - O CICS
  - O ETC.
- O NUMBER OF TERMINALS INSTALLED
  - O RJE, TSO, DB/DC
  - O LOCAL/REMOTE
- O DAILY BREAKDOWN OF WORKLOAD BY SUBSYSTEM SHIFTS
  - O BATCH
  - O TSO
  - O IMS/CICS
  - O ETC.
- O SYSTEM SCHEDULING PROFILE OF CRITICAL APPLICATION TYPES AND SERVICE OBJECTIVES (TURNAROUND AND RESPONSE TIMES)

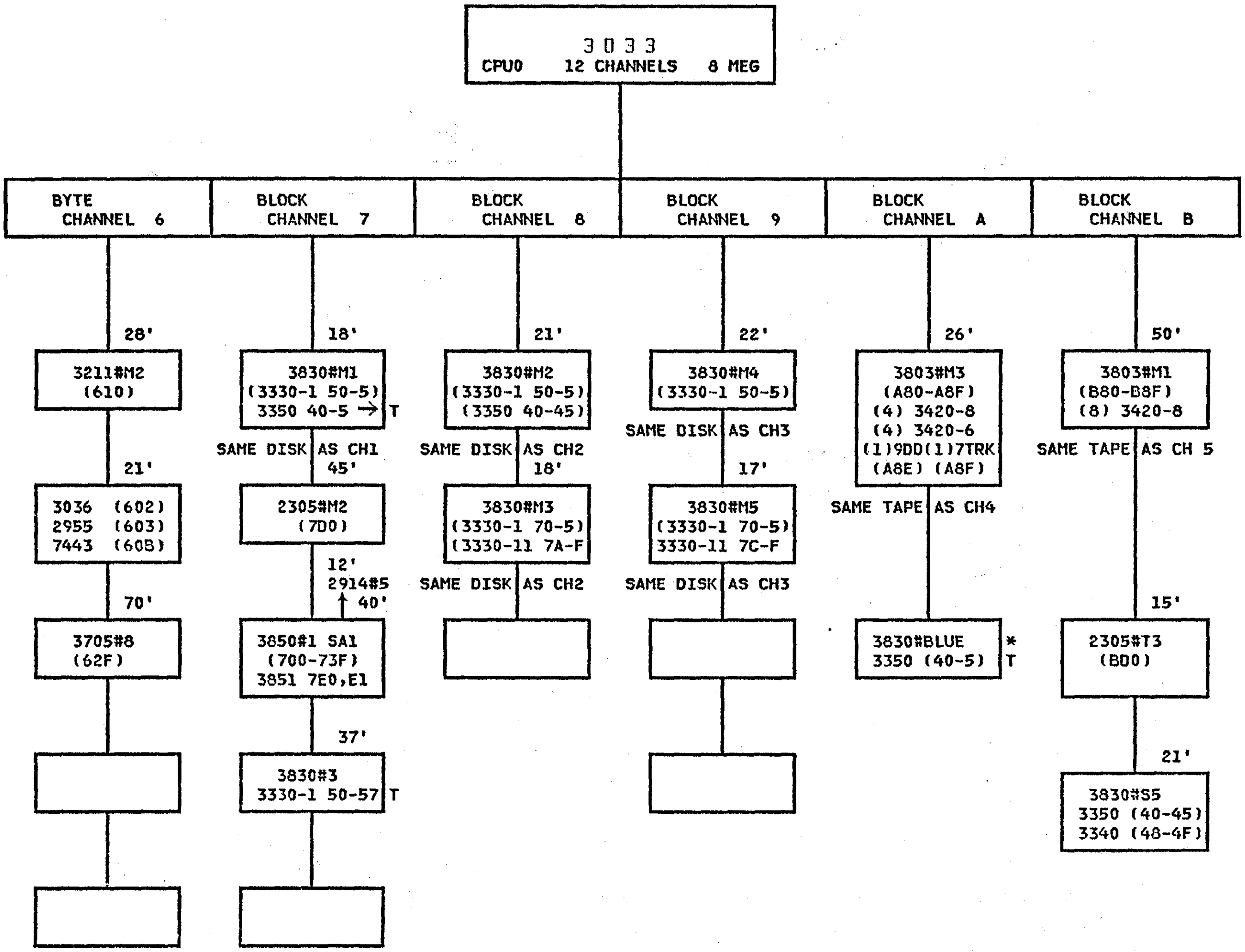
PRELIMINARY DATA REQUIREMENTS

FIGURE - 16



EXAMPLE DESIGN CENTER CONFIGURATION (Part 1)

FIGURE 17



EXAMPLE DESIGN CENTER CONFIGURATION (PART 2)

FIGURE 17

Another important item is a schedule over the period determined to be your critical period of resource consumption. This is the period of time when insufficient computer capacity would make it necessary to purchase additional resources. For example, some DP installations view their month end closing (last three days of month) as that critical period of time, or the first week of the month, or every Thursday. In any event, the schedule should reflect your period of interest. For example, a schedule for the operation of an actual data processing installation is given in Figure 18. This is a weekly schedule for the times of operation of the operating system (MVS) and subsystems (BATCH, TSO,IMS). Also, the time allotted for use of the Advanced Text Management System (ATMS). Although each day has the same profile (i.e., MVS, BATCH, TSO, IMS, ATMS), the workloads (e.g., number and types of transactions and batch jobs) may vary drastically between daily shifts or between different days of the week. This is a 24 hour schedule and critical periods are not indicated. However, if there are critical periods to be analyzed, they must be identified. The key to using a schedule for capacity planning is understanding the workload profile (i.e., types of BATCH, TSO, IMS applications being processed and at what time of the day) and other performance indicators (e.g., CPU time, elapsed times, etc.) across the 18 hours of scheduled operation.

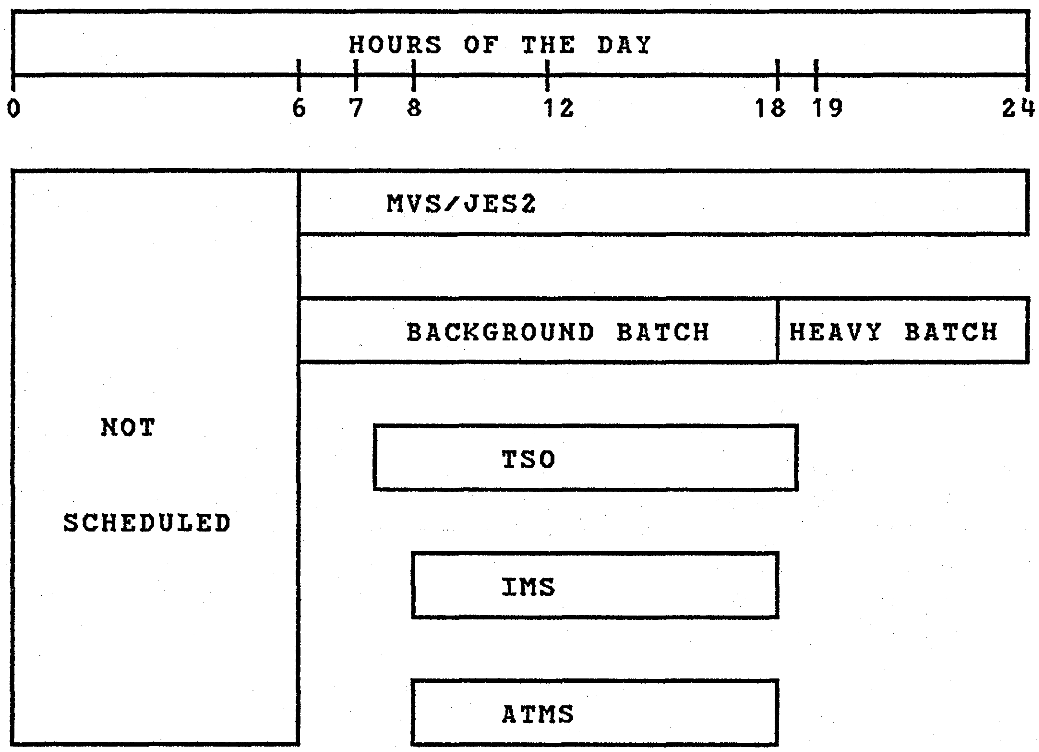
The purpose of an accurate schedule is to indicate which subsystems (BATCH, TSO, IMS, CICS) and critical applications will run on which computer system and the hours of the day they are active.

Capacity planning in this technical bulletin is concerned with managing the resources outlined in Figure 19. Because of the lack of adequate measurement tools for certain system resources as well as a definite attempt to keep the process simple, performance data is collected principally for the following devices:

- CPU
- CHANNELS
- I/O Devices
  - DISK
  - DRUM
  - TAPE
  - PRINTER
- TERMINALS

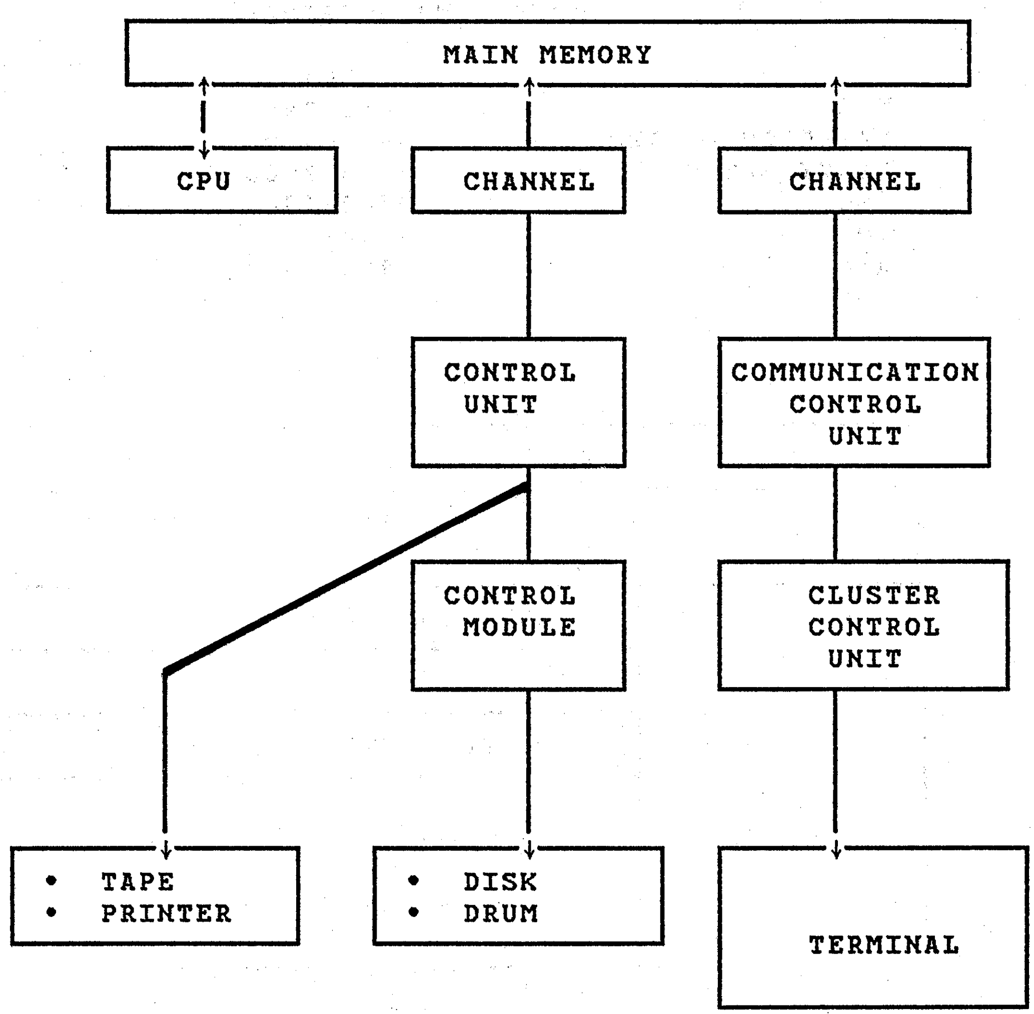
SCHEDULED HOURS OF OPERATION
(MONDAY THRU FRIDAY)

	HOURS	HOURS/WEEK
MVS	0600-2400	90
BACKGROUND BATCH	0600-1800	60
HEAVY BATCH	1800-2400	30
TSO	0700-1900	60
IMS	0800-1800	50
ATMS	0800-1800	50



EXAMPLE DATA PROCESSING SCHEDULE

FIGURE 18



RESOURCE CONSUMPTION

FIGURE 19

Performance of other devices, which are not measured directly, is implicit in other system parameters. For example, system paging rates are indicative of main memory requirements for a particular workload. Also, if response time is known at the host, a measure of user response time at the terminal is indicative of network delays due to lines and controllers.

Having outlined the system configuration and the resources across which measurement data will be collected, the process of data reporting will be discussed. Because the size of DP installations (hardware resources) can be very large inducing a complexity all its own and the number of people within the enterprise (Figure 1) requiring DP information can also be large (another level of complexity), the reporting process must be kept as simple as possible, if it is to be manageable. Therefore, this is the context in which the remainder of this section is being written.

The primary input to the reporting process will be software monitor data, although, some manual input will be required. There are a large number of software monitors available [1] to aid the capacity planning process, but, as pointed out in Reference 1, the number of software monitors installed in a DP installation can be an added source of complexity and confusion for the capacity planning effort. Figure 20, which is extracted from Reference 1, is a small sample of the measurement tools available to collect performance data.

It is hoped that a well planned and implemented capacity planning effort will aid in reducing some of the complexity related to managing a DP installation. In the systematic approach being described, the measurement tools required have been reduced to a minimum. However, in outlining the performance data collection requirements, certain data may not be collected by the tools selected. For example, the tools outlined in Figure 21 will not provide a means for measuring user response time at the terminal (to be discussed later). But, experience has shown that implementation of the capacity planning process requires trade-offs between data availability and cost of acquiring data as well as overall system overhead. Also, a trade-off is required between increased complexity of implementation and the number of measurement tools used.

PERFORMANCE MEASUREMENT TOOLS

- O COLLECTORS
  - O HARDWARE MONITOR
  - O SMF (SYSTEM MEASUREMENT FACILITY)
  - O GTF (GENERAL TRACE FACILITY)
  - O TS TRACE (TIME SHARING TRACE)
  - O IMS/VS SYSTEM LOG
- O ANALYZERS
  - O HARDWARE MONITOR REPORT PROGRAM
  - O SGP (STATISTICS GENERATING PACKAGE)
  - O SMF GRAPHICAL ANALYZER
  - O CAPACITY MANAGEMENT AID
  - O IMS/VS LOG TRANSATION ANALYSIS
  - O IMS/VS STATISTICAL ANALYSIS
- O COLLECTOR - ANALYZER
  - O MF/1 (MEASUREMENT FACILITY 1)
  - O RMF-II (RESOURCE MEASUREMENT FACILITY-II)
  - O SVSPT (VS2 PERFORMANCE TOOL)
  - O VS1PT (VS1 PERFORMANCE TOOL)
  - O SIR (SYSTEM INFORMATION ROUTINE)
  - O CICS PERFORMANCE ANALYZER II
  - O CICS PLOT
  - O CICS DYNAMIC MAP
  - O IMS/VS MONITOR REPORT PRINT PROGRAM
  - O IMS/TRAPDL 1
  - O APL SYSTEM
  - O UTILITY IEHLIST (LIST VTOC)

PERFORMANCE MEASUREMENT TOOLS

FIGURE - 20

- o RMF, SVSPT, VS1PT
- o SMF (BATCH & TSO)
- o CICS/VS PERFORMANCE ANALYZER II
- o IMS/VS LOG ANALYSIS
- o MANUAL LOGS

RECOMMENDED MEASUREMENT TOOLS

FIGURE - 21

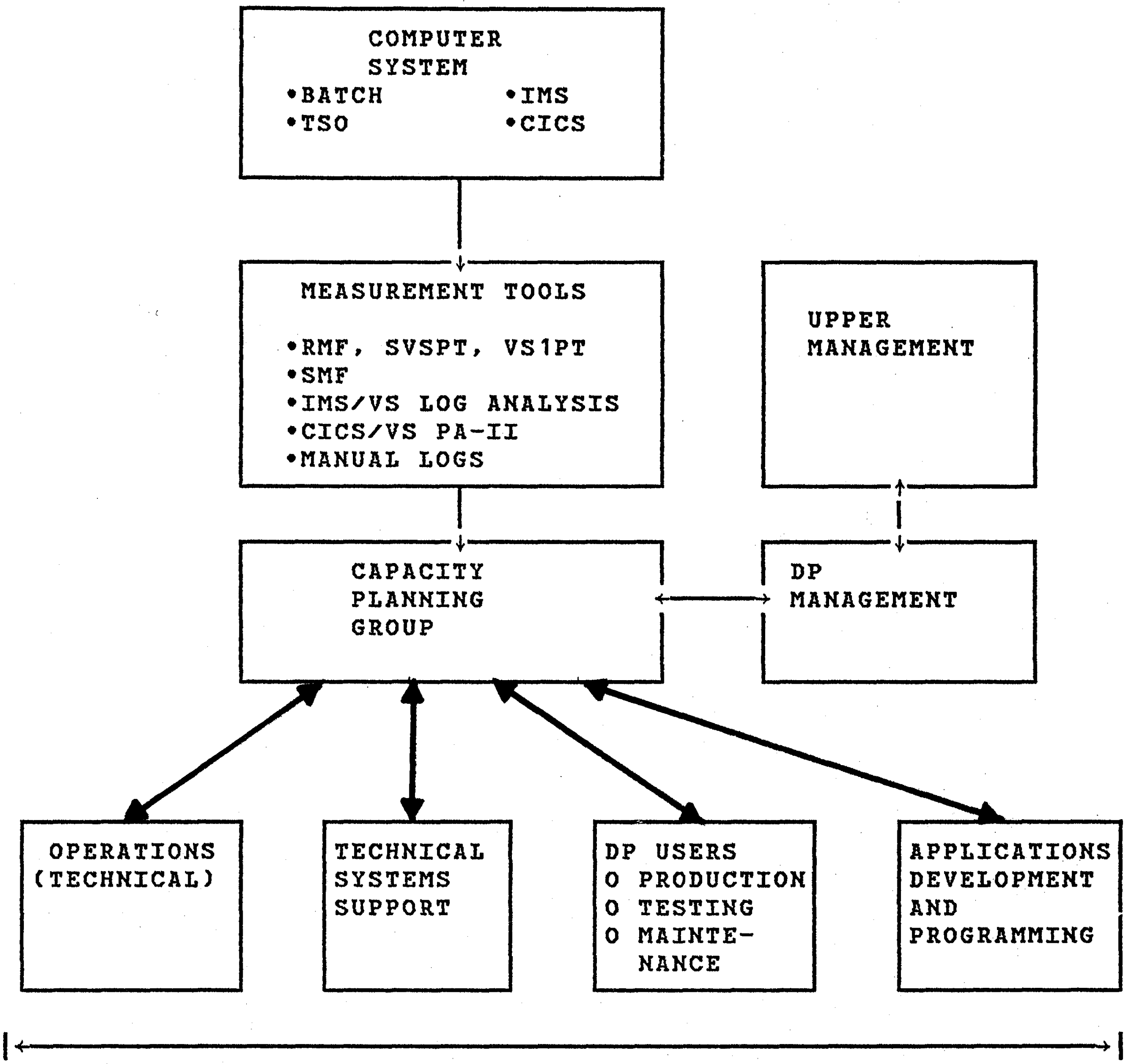
Within this set of measurement tools (Figure 21), a capacity planning process will be described. Although the overall methodology would apply to other subsystems, the subsystems considered by this bulletin are outlined below:

- BATCH
- TSO
- IMS
- CICS

Having outlined the available measurement tools and the subsystems to be measured, the organization requirement for reporting will be discussed. As shown in Figure 22, performance measurement tools are available to provide information for the following areas:

- UPPER MANAGEMENT
  - CORPORATE
  - USER
  - DATA PROCESSING
- DP MANAGEMENT (BELOW VP/DIRECTOR LEVEL)
- CAPACITY PLANNING GROUP
- OPERATIONS
- TECHNICAL SYSTEMS SUPPORT
- APPLICATIONS DEVELOPMENT AND PROGRAMMING
- DP USERS

Each area has certain unique data requirements whereas a great deal of overlap will be noted. The primary factors to consider in the reporting process is what data is pertinent for the recipient, how should this data be displayed (reporting format) and above all what function is the recipient expected to perform once the data is received. One of the main objectives of this section is to discuss the data requirements, formats and functions expected to be performed for each area outlined above.



CO-ORDINATION OF INFORMATION/KNOWLEDGE ACROSS THESE GROUPS IS CRUCIAL TO THE SUCCESS OF THE CAPACITY PLANNING EFFORT.

ORGANIZATIONAL REPORTING REQUIREMENTS

FIGURE - 22

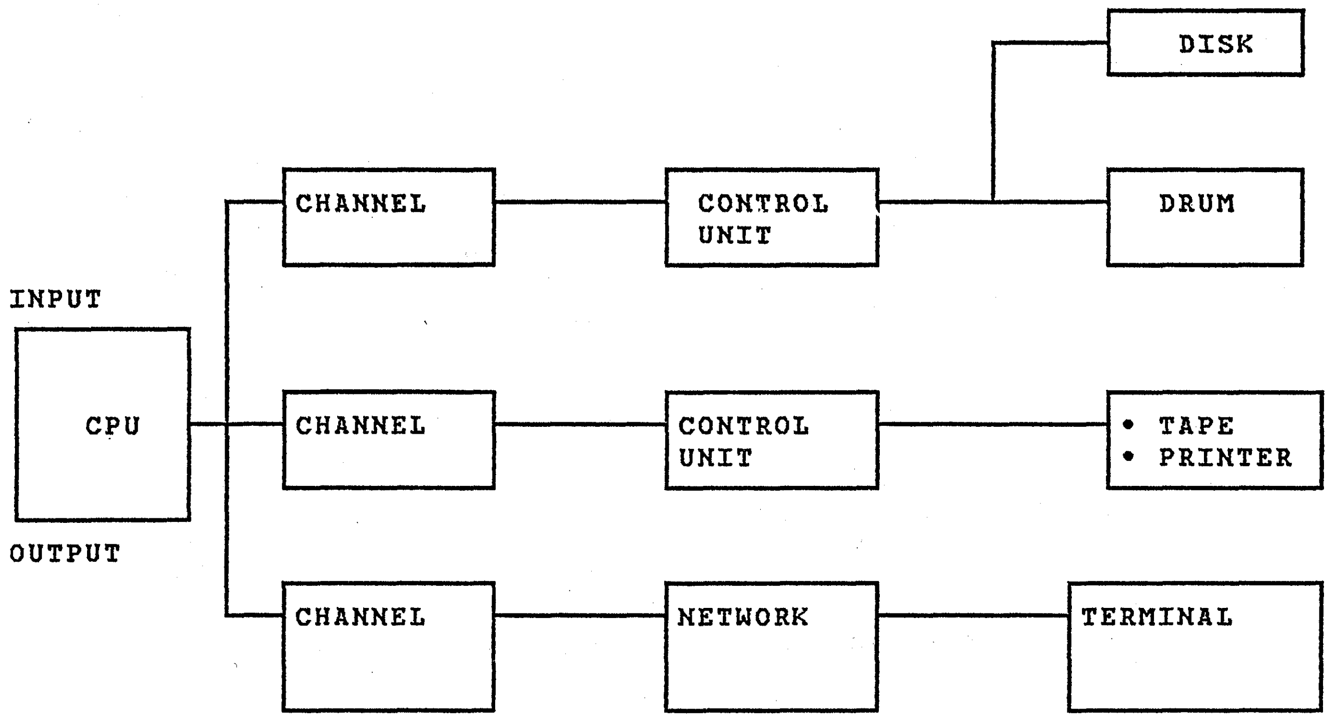
In addressing the problem of defining the pertinent data for capacity planning, there are primarily three elements to be considered. (Figure 23). They are workload, resource utilization and user service data. As shown in Figure 23, the workload, which is a product of the user community, is input and processed causing various levels of resource utilization. The results are returned (output) to the user in a measurable amount of time (response/turnaround time). These parameters are the key factors to be established in the current DP environment and tracked on a continuing basis.

The measurement tools outlined in Figure 21 make it possible to collect system as well as subsystem data. In the case of subsystem data, it is important to understand the burden placed on the system by the IMS subsystem as opposed to Batch or TSO requirements when the three subsystem are running concurrently on the same system. System data is a summation of the total. For example, CPU utilization with no attempt at segmentation by subsystems is a system parameter. This categorization of performance data has worked well as a means of understanding the current DP environment as well as providing the data necessary for system modelling and performance prediction. Therefore, the data requirements outlined in the following paragraphs will be categorized in this fashion.

Of primary concern in addressing data requirements from a system perspective is system availability. From a systems understanding point of view, availability of the system to do actual user problem program work is a key area. As pointed out in Figure 24, there are three areas of availability to be addressed:

- Hardware
- Software
- User perception

These areas were discussed in some detail in section 1.4.5. Of greatest importance to the capacity planning effort is the users perception of his availability. It is of little consolation to a user that the hardware and software are up and running (available) whereas one of his critical data bases is down or being reconstructed. This means for the application requiring this data base the system is unavailable and service requirements are not being met. Therefore, from a capacity planning point of view the proper data must be collected to assess system availability.



WORKLOAD	RESOURCE UTILIZATION	USER SERVICE
•JOBS/HOUR	•UTILIZATIONS	•RESPONSE TIME
•INTERACTIONS/SECOND	•CPU	•TURNAROUND TIME
•TRANSACTIONS/SECOND	•CHANNELS	
	•DISKS/DRUMS	
	•TAPES/PRINTERS	

DATA REQUIREMENTS

(WORKLOAD, RESOURCE UTILIZATION, USER SERVICE)

FIGURE - 23

<table>
<tr><td colspan="3">1 DAY, 1 WEEK, 1 MONTH</td></tr>
<tr><td>UNAVAILABLE</td><td colspan="2">ELAPSED TIME PERIOD
(SYSTEM UP TIME)</td></tr>
<tr><td></td><td>WAIT TIME
(UNUSED)</td><td>RESOURCE BUSY TIMES
(SYSTEM PROCESSING TIME)</td></tr>
</table>

EVALUATION OF SYSTEM AVAILABILITY

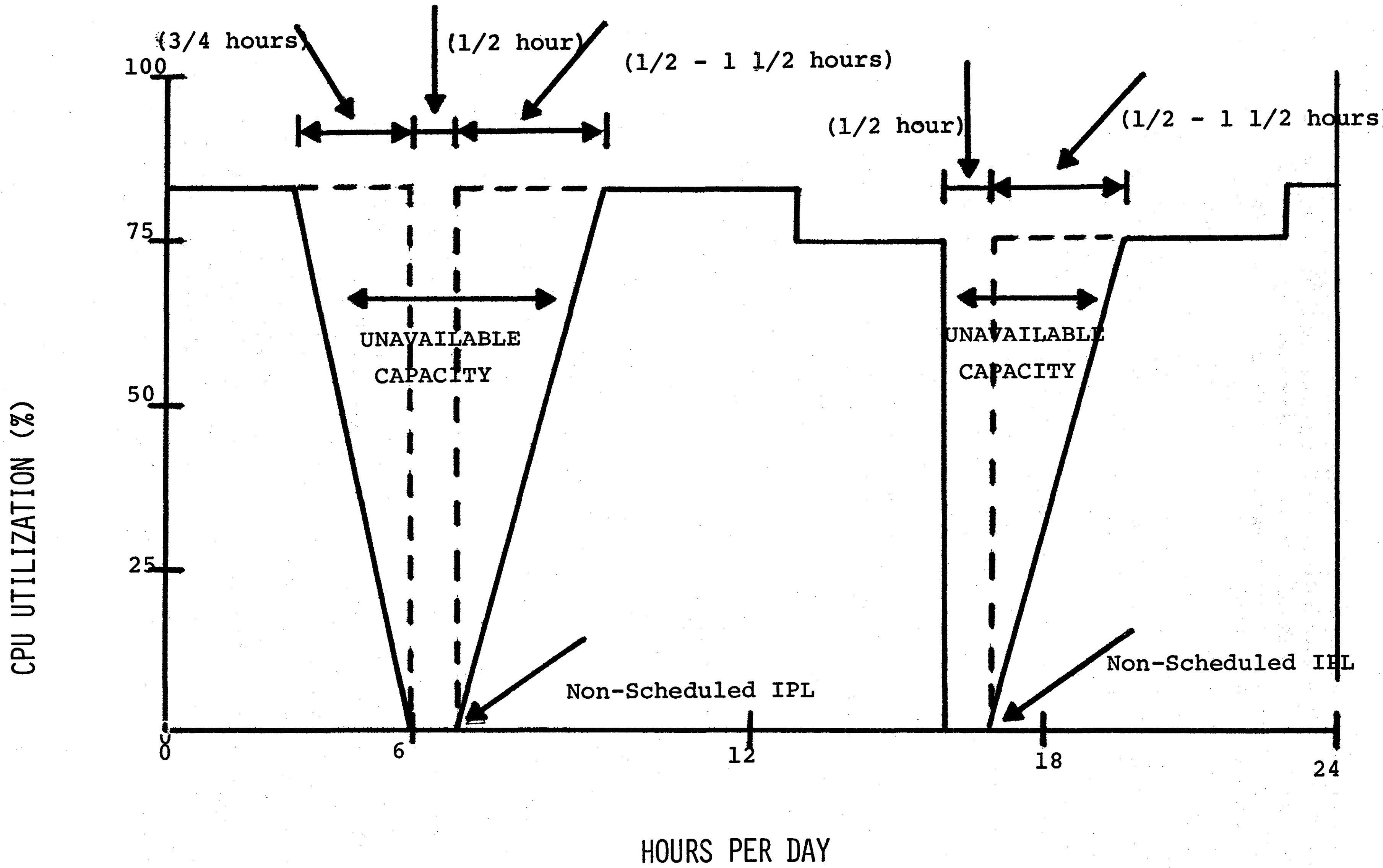
- SYSTEM DOWN TIME
  - PREVENTIVE MAINTENANCE
  - UNSCHEDULED IPL'S
- VARIATIONS IN SCHEDULING DEMANDS
- OPERATIONAL PROBELMS
  - SHIFT CHANGE (PERSONNEL)
- PROGRAM AND DATA PROBLEMS
  - RERUNS (DP/USER CAUSED)
  - MALFUNCTIONING I/O EQUIPMENT
- SYSTEM RECOVERY
  - DATA BASE REQUIREMENTS

SYSTEM AVAILABILITY

(HARDWARE, SOFTWARE, USER PERCEPTION)

FIGURE - 24

In many instances, it is being shown that by keeping the capacity planning process simple, systematically tracking in an ongoing fashion, a clearer understanding of the system's availability is possible. For example, in a case where a DP installation was using SMF as a measurement tool and monitoring on a continuing basis, the following phenomena was sited. Unscheduled IPL's (Initial Program Loadings) can reduce the overall system's availability or capacity (Figure 25). This is a phenomena that is normally treated in the following manner for capacity planning or modelling purposes. Determine the average number of unscheduled IPL's (per day, week, month) and the average amount of downtime, then multiply the two values and the product is the average total downtime or unavailable time. However, some additional factors to be considered are that a system down as shown at hour "6" in Figure 25 will not necessarily follow the dotted line of assumed operation. If a hardware component failed, it would probably follow this dotted line but in many instances when a system fails, it runs for a period of time leading up to the failure in a degraded mode. This is depicted in Figure 25 by the solid negatively sloped line. Therefore, between the solid sloped and dotted vertical line there may be a significant amount of capacity which is not available for problem program work that was assumed available in the simple analysis given above. This would especially be true for accounts with a large number of unscheduled system restarts. This same phenomena (system quiescence) also arises in many cases when a computing system is taken offline for preventive maintenance in a 24 hour, 7 day a week shop. Also, a system which has been down for a period of time as shown at hour "6" (1/2 hour down time), will not necessarily follow the dotted vertical line up to the utilization level it sustained prior to quiescence failure at hour "6". Especially, if a large portion of the user were on-line or interactive. This phenomena has also been noted in pure BATCH environments. The system actually follows the solid line which shows a gradual build-up of CPU utilization. This means users may have left their terminals and are slowly returning as it becomes known the system is available. As in the case of batch work, many jobs that were running before the system failed may not be rerun immediately because other batch work has become more critical. This requires a rescheduling and in many cases a loss of valuable system capacity. Therefore, as indicated in Figure 25, what was thought to be or modelled as a 30 minute outage is a more complex value somewhere between 30 to 165 minutes with a factor for rescheduling.



UNDERSTANDING YOUR SYSTEMS AVAILABILITY/CAPACITY

FIGURE 25

Also, to place another level of complexity on this phenomena, the system may reach full system utilization some 2 3/4 hours later but certain data bases or other application related failures may have occurred and certain users may perceive much longer outages. This example is given to illustrate that within what appears to be overly simplified approaches to capacity planning, a wealth of understanding and insight is possible concerning the complex operation of a DP installation. Be very clear in your understanding about the capacity planning process (measurement, data collection, modelling, prediction), if the system is not understood, the complexity of the modelling technique can not compensate.

The system data required for capacity planning is outlined in Figure 26. The first parameter, which is very important, is total CPU hours available to do problem program work. Several parameters are bracketed off to indicate their interrelationship. The first parameter in the bracket is total CPU utilization. The other parameters are collected as being major sources of CPU utilization. Therefore, the proportion in which each burden the CPU is an important part of understanding and analyzing the CPU. Critical channels and I/O devices are monitored for their busy time or utilizations. In analyzing the user response time (On-line) or turnaround time (Batch), a portion of the delay (queue time) and resource service time is imposed by the I/O portion of the system (channels, control units, I/O device). The RMF measurement tool provides system data to aid in calculating the I/O portion of response/turnaround time. These three parameters are bracketed beginning with "Average Device Queue Length". The remaining nine parameters are reported to provide some insight into multiprogramming. Basically, these parameters establish the average system multiprogramming level and the portion attributable to the various subsystems (BATCH, TSO, IMS, CICS).

Data requirements for each subsystem are necessary in addition to the overall system data. Although, data requirements for the BATCH, TSO, IMS, and CICS environments are discussed, the basic view in data collection is that the user generated workload can be characterized in some reasonable fashion and the associated resource service quantified. The quality of the data collected when quantifying resource utilization (CPU, channels, I/O Devices, etc.) by subsystem or user is governed by the accuracy of the meausrement tools. For the subsystems to be measured and the measurement tools outlined above, the data requirements for BATCH, TSO, IMS, and CICS reflect only those values it is possible to separate at this time.

- TOTAL CPU HOURS AVAILABLE
- TOTAL CPU HOURS USED BY PERIOD
- NUMBER SWAPS PER TIME PERIOD (OUT)
- NUMBER OF SWAPS PER TIME PERIOD (IN)
- NUMBER OF PAGES PER SWAP OUT
- NUMBER OF PAGES PER SWAP IN
- DEMAND PAGING RATE (OUT)
- DEMAND PAGING RATE (IN)
- VIO PAGING RATE (OUT)
- VIO PAGING RATE (IN)

- PERCENT BUSY CHANNEL
- PERCENT BUSY BY I/O DEVICE

- AVERAGE DEVICE QUEUE LENGTH
- DEVICE ACTIVITY RATE
- NUMBER OF SIO'S PER CHANNEL
- AVERAGE MULTIPROGRAMMING LEVEL PER TIME PERIOD
- MAXIMUM NUMBER OF INITIATORS ASSIGNED
- AVERAGE NUMBER OF ACTIVE INITIATORS
- MAX NUMBER OF TSO USERS
- AVERAGE NUMBER OF ACTIVE TSO USERS
- MAX NUMBER OF MPP (MESSAGE PROCESSING PROGRAMS)
- AVERAGE NUMBER OF ACTIVE MPP
- MAX NUMBER OF CICS TASKS
- AVERAGE NUMBER OF CONCURRENT CICS TASKS

SYSTEM DATA

FIGURE - 26

Batch data requirements are outlined in Figure 27. Most of this data is self explanatory but, the trade-offs between collecting Job or Job Step information might require some discussion. From a capacity planning and tracking point of view, it appears that job data is adequate. Although, for a more detailed analysis, it may be necessary to understand what jobs are consuming which resources and at what time of the day. It may not be enough to know that a job performs so many thousand I/O's during its execution time. If the last job step does essentially all the I/O and this is done during the last 30 minutes of a 2 hour run, it may be necessary to understand job steps. Also, from a planning point of view, if the workload of a DP installation can be categorized into a small number of categories or clusters [12] of job step types (CPU Bound, I/O Bound, etc.) and historical data indicates how these categories have grown in the past, job step information can provide valuable capacity planning informataion. The data required for the TSO, IMS, and CICS environments is outlined in Figures 28, 29, 30. The data items are self explanatory except possibly for recording TSO workload in interactions instead of commands. A user sitting at a terminal may transmit and receive several times for one command. In that case for performance purposes, the concern is for the number of interactions and not the command. For example, in addressing user service, the concern is for response time associated with each interaction as opposed to the sum of interaction across the command. The TSO information recorded by RMF addresses interaction activity. The use of this data will be discussed when the functional requirements of the recipients are outlined.

In the following discussions on data requirements for various groups and individuals nothing is said about the frequency (daily, weekly, monthly, quarterly) of reporting. In some cases, certain basic guidelines may be given. It seems from experience that the DP installation is best at establishing these values. Using RMF as a base, Figure 31 outlines measurement data recording cycles, reporting intervals and frequency of reporting. As a basic guideline, recording data for tuning purposes is done in the range of 250 milliseconds and reported on 15 minute intervals. But in the case of capacity planning, which is an ongoing effort, a recording cycle of 1 second and a reporting interval of 1 hour is quite reasonable. In cases where certain peak load phenomena is analyzed, a variation to these guidelines might be required. Also, the cycle, interval or frequency values are not rigidly fixed and if a period of data collection indicates initial setting are not satisfactory, change them. This is the reason for emphasizing continuous ongoing system monitoring. Many parameter requirements and reporting questions which at first would appear to be trivial, are not and systematic tracking and reporting will provide some insight.

WORKLOAD	SERVICE

WORKLOAD

- O JOBS
  - O SHORT
  - O MEDIUM
  - O LONG
- O JOB STEPS (CATERGORIZED)
  - O CPU BOUND
  - O I/O BOUND
- O EXCP'S PER JOB
- O EXCP'S PER JOB STEP
- O AVG. TAPE MOUNTS/JOB
- O PRINTER SET-UP TIME/JOB
- O AVG. NUMBER OF PRINT LINES/JOB
- O RESOURCE REQUIREMENTS/JOB
  - O CHANNELS
  - O DISKS/DRUMS
  - O TAPES
  - O PRINTERS
- O I/O RESOURCE CONSUMPTION BY JOB IF POSSIBLE

SERVICE

- O TOTAL BATCH CPU HOURS/DAY
- O CPU TIME (TCB & SRB)/JOB
- O CPU TIME (TCB & SRB)/JOB STEP
- O EARLIEST START TIME/JOB
- O LATEST END TIME/JOB
- O ELAPSED TIME/JOB
- O ELAPSED TIME/JOB STEP

BATCH DATA REQUIREMENTS

FIGURE - 27

WORKLOAD	SERVICE
O NUMBER OF INTERACTIONS O TRIVIAL O MEDIUM O COMPLEX	O TOTAL TSO CPU HOURS/DAY O CPU TIME (TCB & SRB) BY INTERACTION TYPE
O EXCP'S BY INTERACTION	O RESPONSE TIME BY INTER-ACTION TYPE O CPU O TERMINAL
O NUMBER OF SESSIONS/PERIOD O LOCATION O DURATION O NUMBER OF INTERACTIONS	
O RESOURCES REQUIRED BY SESSION O CHANNELS O DISKS/DRUMS O TAPES O PRINTERS	
O I/O RESOURCE CONSUMPTION BY INTERACITON TYPE IF POSSIBLE	

TSO DATA REQUIREMENTS

FIGURE - 28

WORKLOAD

- O TRANSACTIONS (BY LOCATION AND TYPE)
  - O SHORT
  - O MEDIUM
  - O COMPLEX
- O EXCP'S BY TRANSACTION TYPE
- O REQUIRED BY CRITICAL APPLICATIONS
  - O CHANNELS
  - O DISKS
  - O TAPES
  - O PRINTERS
- O I/O RESOURCE CONSUMPTION BY TRANSACTION TYPE IF POSSIBLE

SERVICE

- O TOTAL CPU HOURS/DAY
- O CPU TIME/TRANSACTION TYPE
- O RESPONSE TIME BY TRANS. TYPE
  - O CPU
  - O TERMINAL

IMS DATA REQUIREMENTS

FIGURE - 29

WORKLOAD

- O TRANSACTIONS (BY LOCATION AND TYPE)
  - O SHORT
  - O MEDIUM
  - O COMPLEX
- O EXCP'S BY TRANSACTION TYPE
- O REQUIRED BY CRITICAL APPLICATIONS
  - O CHANNELS
  - O DISKS
  - O TAPES
  - O PRINTERS
- O I/O RESOURCE CONSUMPTION BY TRANSACTION TYPE IF POSSIBLE

SERVICE

- O TOTAL CPUR HOURS/DAY
- O CPUR TIME/TRANSACTION TYPE
- O RESPONSE TIME BY TRANS. TYPE
  - O CPU
  - O TERMINAL

CICS DATA REQUIREMENTS

FIGURE - 30

- RECORDING CYCLE
  - 250 MILLISECONDS
  - 500 MILLISECONS
  - 1.0 SECOND

- REPORTING INTERVAL
  - 15 MINUTES
  - 30 MINUTES
  - 1 HOUR

- REPORTING FREQUENCY
  - DAILY
  - MONTHLY
  - WEEKLY
  - QUARTERLY

DATA COLLECTION AND REPORTING

(RMF BASE)

FIGURE - 31

In defining the data reporting requirements for operations, it is felt that the key to good capacity planning will be found in the operations area. In capacity planning, where modelling and prediction is used for planning future systems (hardware and software), the capacity planning group has a particular view or perspective on how the computer system operates. This system view is primarily developed from gross measurement data. By reporting the proper data to operations, two things are accomplished. First, it will provide personnel in operations with a better understanding of the user workload characteristics, and the rate at which resources are being consumed. Secondly, it will provide for validation of the measurement data and the conceptualization that the capacity planning group has of the overall computer operation (availability, resource utilizaton, workload, scheduling, etc.). In other words it will allow the group responsible for operations to review some of the data which attempts to characterize DP operations. For example, the daily scheduling of user workload often times is not the way jobs are actually run for that given day. In many cases, the only way to determine how a given user workload was ran on a computer system is to wait until the following day.

For operations, the reported data should provide a better understanding of how the system is being driven (user workload), the rate resources are being consumed and the kind of service being provided to the users. Much to the surprise of many in the capacity planning area, it is found that in some installations operations personnel do not receive many basic performance reports (e.g., resource utilization, workload data, etc.).

The first category of data for operations, which is shown bracketed in Figure 32, is used basically to establish CPU consumption. Namely, the number of CPU hours available over some period of time to do problem program work. Then, within the available hours, segment this total by subsystem. The next four categories are self explanatory. The purpose of this data is to provide operations with a view of resource consumption as it relates to a given user workload. Along with this data, provide the corresponding user service values. In monitoring the CPU, the level of multiprogramming provides some insight into its throughput. As data for operations, shown bracketed in Figure 32, there is the average multiprogramming level over some period of time and the portion of multiprogramming attributable to each subsystem. The remaining data (Figure 32, Part 2) is concerned with configuration information, tape and printer data, all is very relevant to modelling and prediction. It may provide operations with information in certain very large shops, but, in most cases, operations is probably aware of many of these parameters, hence, validation is the primary purpose of its reporting.

- O TOTAL CPU HOURS AVAILABLE/DAY
- O TOTAL CPU HOURS CONSUMED/DAY
- O TOTAL BATCH CPU HOURS/DAY
- O TOTAL TSO CPU HOURS/DAY
- O TOTAL IMS CPU HOURS/DAY
- O TOTAL CICS CPU HOURS/DAY
- O DAILY UTILIZATIONS (INTERVALS LESS THAN HOUR)
  - O CPU
  - O CHANNELS
  - O DISKS, TAPES, PRINTERS
- O WORKLOAD CHARACTERIZATION BY TYPE TRIVIAL (SHORT), MEDIUM, COMPLEX (LONG)
  - O DAILY RATE OF BATCH JOBS
  - O DAILY RATE OF TSO INTERATIONS
  - O DAILY RATE OF IMS/CICS TRANSACTIONS
- O ELAPSED TIME FOR CRITICAL BATCH JOBS
- O RESPONSE TIMES AT CPU AND TERMINAL BY TYPE
  - O TSO, IMS, CICS
- O AVERAGE MULTIPROGRAMMING LEVEL
- O MAXIMUM INITIATORS ASSIGNED
- O AVERAGE NUMBER OF ACTIVE INITIATORS
- O MAXIMUM MPP ASSIGNED
- O AVERAGE NUMBER OF ACTIVE MPP
- O AMAX TASK FOR CICS
- O AVERAGE NUMBER OF ACTIVE TASK FOR CICS

DATA REQUIREMENTS FOR OPERATIONS (PART 1)

FIGURE - 32

- RESOURCE REQUIREMENTS FOR CRITICAL APPLICATIONS (BY JOB, SESSION OR TRANSACTION)
  - CHANNELS
  - DISK/DRUMS
  - TAPES
  - PRINTERS
- TSO SESSION INFORMATION
  - LOCATION
  - DURATION
  - NUMBER OF INTERACTIONS
  - AVERAGE NUMBER OF ACTIVE SESSIONS
- NUMBER OF TAPE MOUNTS FOR CRITICAL JOBS
- AVERAGE TIME TO MOUNT TAPES (TOTAL)/JOBS
- NUMBER OF PRINT LINES BY JOB
- AVERAGE PRINT EXECUTION TIME BY JOB
- AVERAGE PRINTER SET-UP TIME BY JOB
- AVERAGE DISK MOUNT TIME BY JOB

DATA REQUIREMENTS FOR OPERATIONS (PART 2)

FIGURE — 32

As an aid in developing a capacity planning project, the source of the data required for operation has been outlined in Figure 33. In some cases, the data is not reported in the required format and a column labelled "new report" is indicated as a source. This indicates a new report must be developed if it is to appear in the format indicated in the following paragraph.

To take the capacity planning process a step further, it is felt that in the planning stages, the capacity planning group must not only clearly spell out the data requirements for each functional area (Figure 13), they must outline the reporting formats required. It should be understood, as previously brought out concerning other areas, that certain formats, like data collection requirements, are your best judgement at the time and future knowledge may lead to a change in format. So much of the capacity planning process depends on "try it", "track it", and "change it". But, above all, initiate the process. Many capacity planning projects never get off the ground because the capacity planning group is waiting for data and formatting requirements to be initiated by each functional area (Figure 13). In many cases, these functional groups truly do not know what data is actually available or in what formats they can request it. A large portion of the data reported for operations can be formatted as bar graphs (Figure 34). To reduce the number of reports, subsystem utilization values might be plotted on the same graph showing the break out of each. In cases where the number of graphs become excessive, a cover page summarizing the total might be required and the graphs would become supporting detail.

The following examples of bar graphs are provided:

Figure 35 - Total CPU Hours Consumed

Figure 36 - CPU Hours consumption for a subsystem

Figure 37 - CPU Hours Consumed by subsytems

Figure 38 - Channel Busy

Figure 39 - Transaction Rate (per second)

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA-II	IMS LOG ANALYSIS	MANUAL LOGS	NEW REPORT
UNAVAILABLE CPU HOURS	X		X			X	X
CPU HOURS CONSUMED		X					
CPU HOURS NOT USED (WAIT)		X					
TOTAL BATCH CPU HOURS	X						X
TOTAL TSO CPU HOURS	X						X
TOTAL IMS CPU HOURS	X						X
TOTAL CICS CPU HOURS	X						X
CPU UTILIZATION		X					
CHANNEL UTILIZATION		X					
DEVICE UTILIZATION		X					
BATCH JOB RATE	X	X					
TSO INTERACTION RATE	X	X					
IMS TRANSACTION RATE					X		
CICS TRANSACTION RATE				X			
ELAPSED TIME/JOB			X				X
TSO RESPONSE TIME	X						
IMS RESPONSE TIME					X		
CICS RESPONSE TIME				X			
AVERAGE MPL	X			X	X		X
MAXIMUM INITIATORS						X	
AVERAGE ACTIVE INITIATORS	X		X				X

SOURCE OF OPERATIONS DATA (PART 1)

FIGURE — 33

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA-II	IMS LOG ANALYSIS	MANUAL LOGS	NEW REPORT
MAXIMUM MPP						X	X
AVERAGE ACTIVE MPP					X		X
AMAX TASKS						X	X
AVERAGE ACTIVE TASKS				X			X
RESOURCE REQUIREMENTS			X				X
TSO SESSION INFORMATION		X	X				X
TAPE MOUNTS/JOB						X	
TAPE MOUNT TIME/JOB	X					X	X
PRINT LINES/JOB			X				X
AVERAGE PRINT TIME/ JOB			X				X
PRINT SET-UP TIME/JOB						X	X
PRINTER UTIL. BY JOB			X				X
DISK MOUNT TIME/JOB	X					X	X

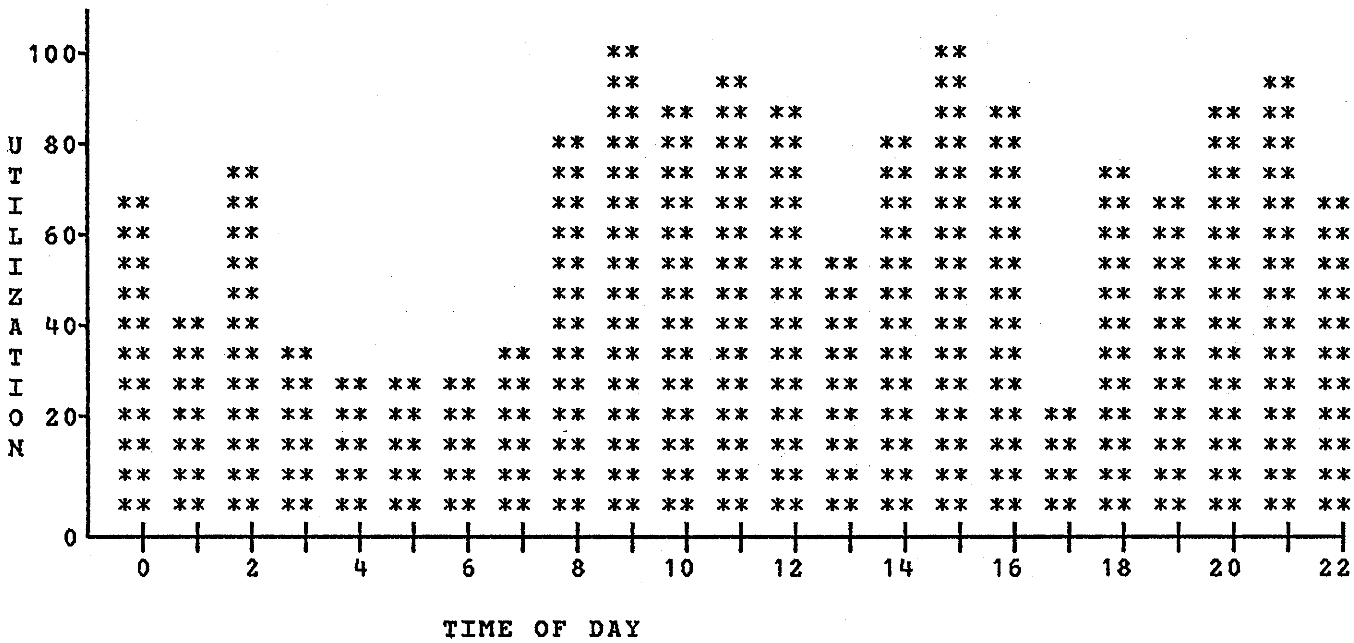
SOURCE OF OPERATIONS DATA (PART 2)

FIGURE — 33

- O TOTAL CPU HOURS CONSUMED/DAY
- O TOTAL BATCH CPU HOURS/DAY
- O TOTAL TSO CPU HOURS/DAY
- O TOTAL IMS CPU HOURS/DAY
- O TOTAL CICS CPU HOURS/DAY
- O DAILY UTILIZATIONS
- O CPU
- O CHANNELS
- O DISKS, TAPES, PRINTERS
- O WORKLOAD CHARACTERIZATION (TRIVIAL (SHORT), MEDIUM, COMPLEX (LONG)
  - O DAILY RATE OF BATCH JOBS
  - O DAILY RATE OF TSO INTERACTIONS
  - O DAILY RATE OF IMS/CICS TRANSACTIONS
- O ELAPSED TIME BY PERFORMANCE GROUP (PERIOD) BY DAY (BATCH JOBS)
- O RESPONSE TIME BY PERFORMANCE GROUP (PERIOD) BY DAY (TSO)
- O RESPONSE TIME BY CATEGORY (TRANSACTION TYPES) BY DAY (IMS, CICS)
- O AVERAGE MULTIPROGRAMMING LEVEL
- O AVERAGE NUMBER OF ACTIVE INITIATORS
- O AVERAGE NUMBER OF ACTIVE MPP
- O AVERAGE NUMBER OF ACTIVE TASKS

OPERATIONS DATA REQUIREMENTS AND GRAPHICAL FORMATS

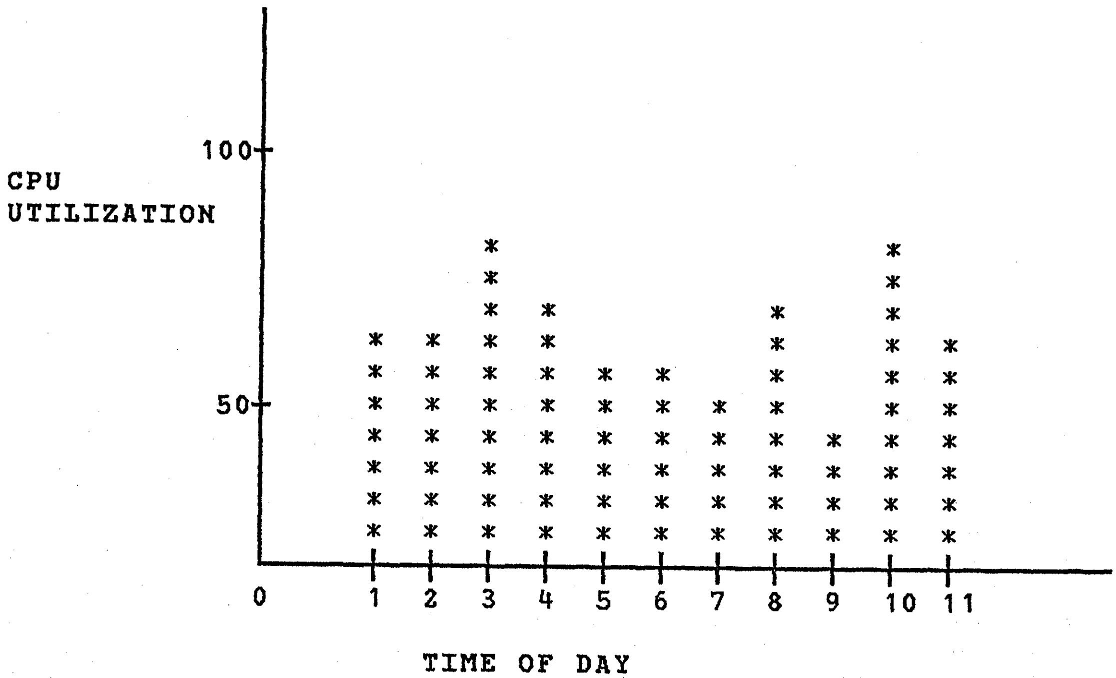
FIGURE - 34



TOTAL CPU HOURS CONSUMED

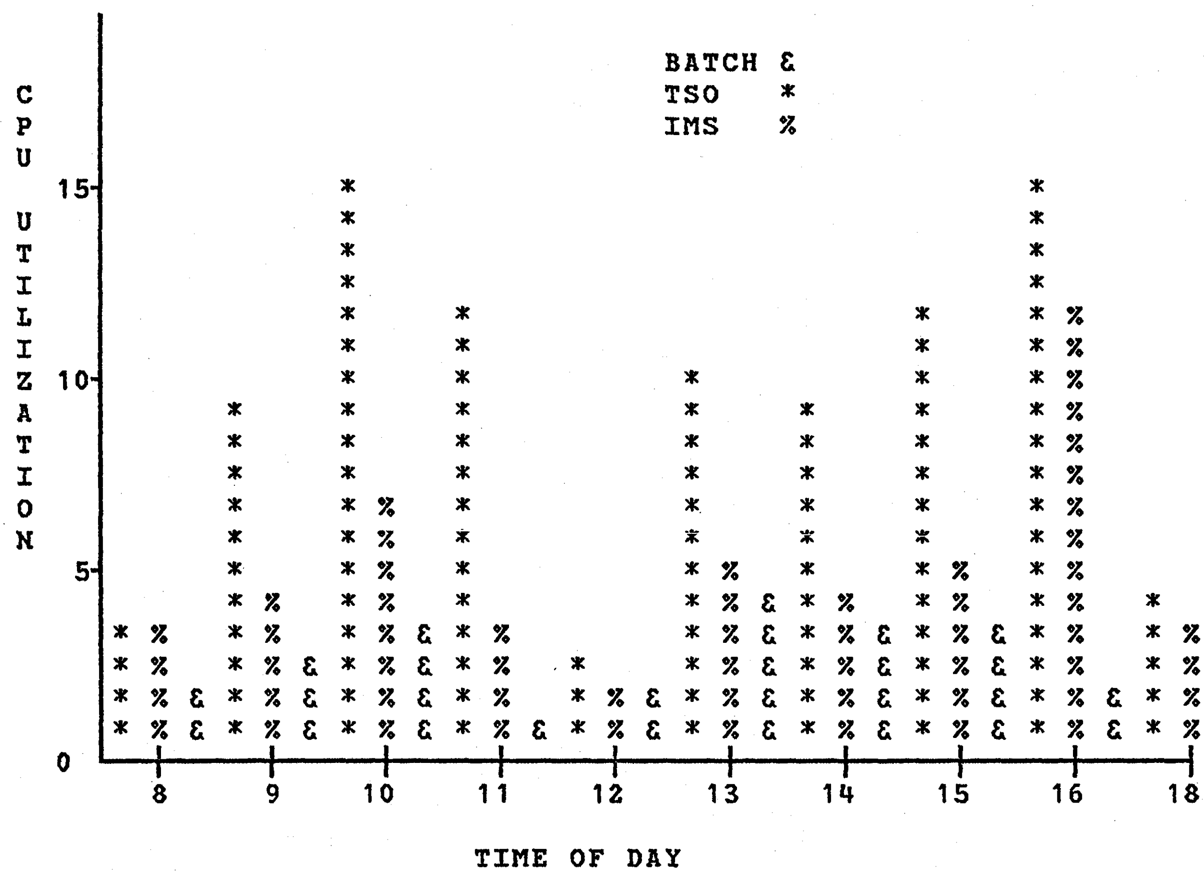
FIGURE - 35

BATCH OR TSO OR CICS OR IMS



CPU HOURS CONSUMED BY A SUBSYSTEM

FIGURE - 36



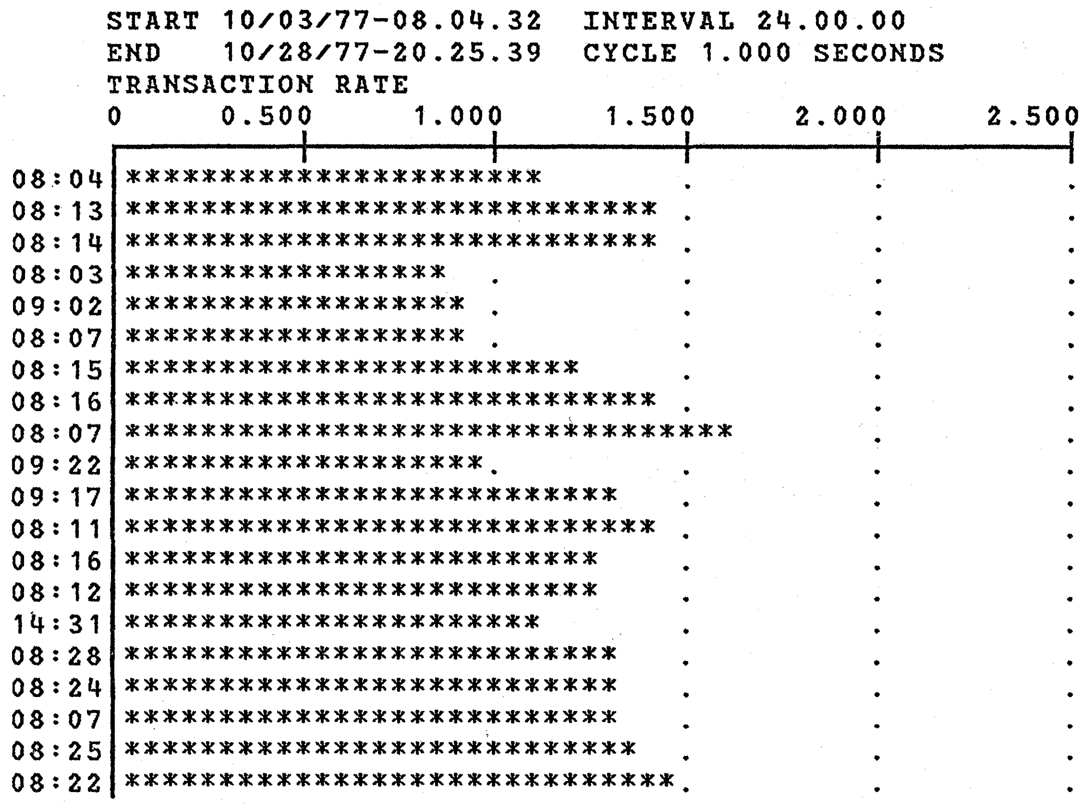
CPU HOURS CONSUMED BY SUBSYSTEMS

FIGURE - 37

```
           START 03/14/78-08.43.41  INTERVAL 00.14.46
           END   03/14/78-20.54.06  CYCLE 1.000 SECONDS
           CPU 1 CHANNEL 01              BUSY PERCENTAGE
           0       0.0      20.0      30.0      40.0      50.0
           +---------+---------+---------+---------+---------+
     08:43 |**********         .         .         .         .
     08:58 |********************         .         .         .
     09:13 |*********************        .         .         .
     09:28 |*****************  .         .         .         .
     09:43 |****************** .         .         .         .
     09:58 |****************************************         .
     10:13 |****************** .         .         .         .
     10:28 |************************     .         .         .
     10:43 |**************************** .         .         .
     10:58 |********************************       .         .
     11:13 |**************************   .         .         .
     11:28 |**************************   .         .         .
     11:43 |**************     .         .         .         .
     11:58 |*************      .         .         .         .
     12:13 |************       .         .         .         .
     12:28 |*******  .         .         .         .         .
     13:09 |*************      .         .         .         .
     13:24 |**************************   .         .         .
     13:39 |**************************** .         .         .
     13:54 |************************************** .         .
     14:09 |***********************************************  .
     14:24 |+++++++++++++++++++   53.0   +++++++++++++++++++++
     14:39 |***************************  .         .         .
     14:54 |***********************      .         .         .
     15:09 |***********************************    .         .
     15:24 |************************     .         .         .
     15:39 |**************************** .         .         .
     15:54 |*******************************        .         .
     16:09 |***************************************.         .
     16:24 |***********************************    .         .
     16:39 |************************************************ .
     16:54 |***************************************.         .
     17:09 |*********************************      .         .
     17:24 |***********************************    .         .
     17:39 |***********************      .         .         .
     17:54 |*************************    .         .         .
     18:09 |*************************    .         .         .
     18:24 |*****************************.         .         .
     18:39 |**************     .         .         .         .
```

CHANNEL BUSY (%)

FIGURE - 38



TRANSACTIONS RATE (PER SECOND)

FIGURE - 39

User service (response/turnaround time) is an important parameter to be monitored by the operations personnel. Normally, operations only become aware of poor user service by telephone calls. In most instances, they get no warning that service is beginning to degrade. A major part of reporting user service on a continuing basis is that operations may begin to see trends in degrading service. In most cases, tracking and providing batch turnaround time is no problem. But with the current state of measurement tools, measurement of user response time for on-line or interactive systems is not a trivial exercise. When analyzing response time values, one must be clear as to the point in the system where the value was measured. In most instances, response time values are reported at the Host System and the network delays are excluded. In capacity planning and from an operations point of view, response time values are needed at the terminal where network delays are included. In all cases, the response time values provided by the measurement tools outlined in Figure 21 are Host values. Therefore, an approach for determining response time at the user terminal must be defined. There are several approaches for determining Terminal Response as outlined below:

- Manual (using a stop watch)
- Host Measurements (guidelines for line and controller delays)
- Minicomputer (i.e., 8100, 3790, System/7, Series 1)
- Monitor Terminal Inhibit Light

In choosing one of the approaches outlined above consideration should be given to cost of implementation, accuracy required and ease of installation. It is possible to systematically structure a manual approach for collecting response time at the terminal. As an example of this process, the following is a summary of an actual installation's approach for manually collecting response time data at the user terminal.

This manual approach was systematically structured and a reporting format developed to lead the person measuring response time completely through the process. In Figure 40, certain pertinent information has been extracted from this report. The essence of the process was to clearly outline the times of the day in which the measurements would be taken, the required physical locations, maintain relevant user scripts (proper mix of transaction types), and establish a process of collection and summarization of the data. A stop watch would be used to time the terminal scripts. These response time values were tracked and correlated with the Host response time values on certain transaction types. If tracking over time indicates that network delays remain relatively constant, then manual response time measurements would not have to be done as often. Such a process establishes network delays for certain transaction types and updates them periodically.

Software monitors only provide transaction response time at the host system. For example, it is possible to use utility programs designed to collect and report IMS response times. If other measures have been used to determine the approximate line and controller delays for these IMS transactions, then the sum of the host response time and approxiate delay value could be used to approximate user response times at the terminal.

A minicomputer can be used to collect user response times by submitting dummy transaction for processing at various times of the day or time stamping normal transactions and summarizing the timing data. When using a minicomputer for measuring user response time, remember that the values collected are only indicative of response time as viewed by users in that part of the network.

Using a measurement tool to monitor the input inhibit light at the user terminal is another approach for measuring response time. In monitoring the inhibit light, one must be aware of the control mechanisms in the application programs (e.g., IMS, CICS). In a situation where IMS is being used in non-response mode (terminal keyboard is unlocked for another transmission before response to last transaction is received), the inhibit light is turned out when the transaction is correctly received at the host, this means host processing will not be included in the total response time value.

- DATE
- SAMPLE TAKEN BY
- LOCATION
- AVAILABILITY CODES
  - TSO, IMS, CICS AVAILABLE
  - CRASH DURING SESSION
  - SYSTEM DOWN
  - TERMINAL NOT AVAILABLE
  - NO TEST TAKEN
- REPORT TIMES: 9:00. 11:00, 14:00, 16:00
- TERMINAL ID OR LINE
- WALL CLOCK TIME
- FIRST ATTEMPT TO LOGON
- LAST ATTEMPT TO LOGON
- NUMBER OF USERS ON SYSTEM
- INPUT SCRIPT AND TIME VARIOUS INTERACTIONS/TRANSACTIONS
- LOGOFF
- WALL CLOCK TIME
- SUMMARY REPORT FOR DAY AT EACH RECORDING TIME
  - LOGON TIME AND DURATION
  - RESPONSE TIME (TRIVIAL, MEDIUM, COMPLEX)
  - NUMBER OF USERS
  - LOGOFF TIME
  - TOTAL TIME

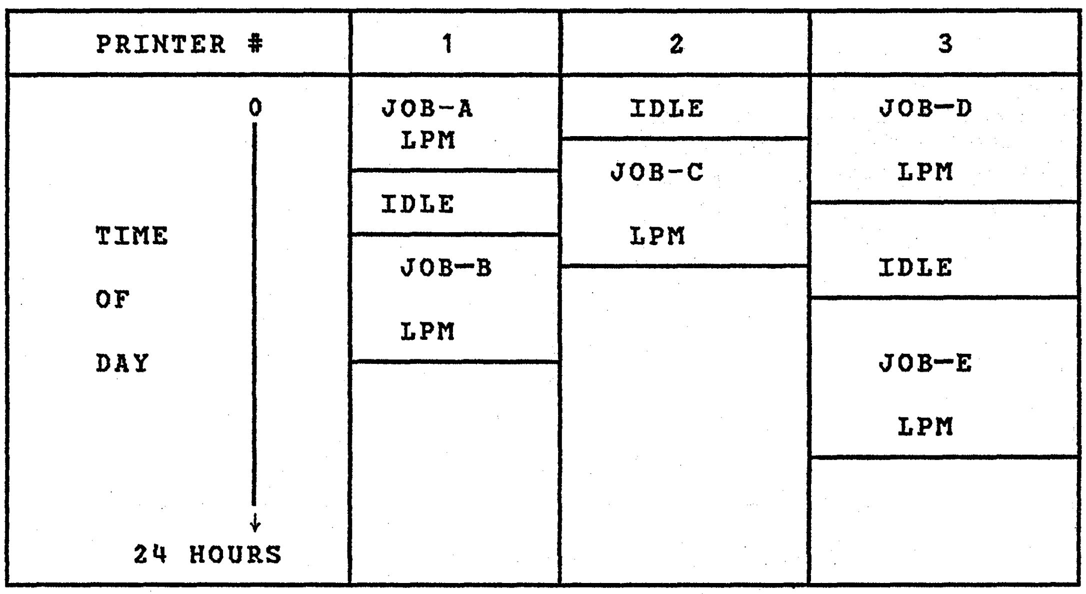
TERMINAL RESPONSE TIME REPORT (MANUAL)

DATA REQUIREMENTS

FIGURE - 40

To make the operations reporting picture complete, printer utilizations and certain magnetic tape data is required as outlined in Figures 41, 42 and 43. This data is self explanatory in that overall system capacity is related to getting the final report off the printer (total turnaround time) as well as magnetic tape mounting being very crucial to a performance analysis when a batch job's overall elapsed time is being determined. In many instances, this data may be known to operations but once again a concurrance on the capacity planners view of operations is very necessary.

System generation involves the specification of particular system options under a starter system or driver. This starter system is used to generate the desired system. System initialization is part of system tailoring that takes place after the system has been generated. The tailoring results from the specification of system parameters, first at intital program loading (IPL) and later when certain operator commands are issued. One of the primary sources of system tailoring information is the data set named "SYS1. PARMLIB". The purpose of "PARMLIB" is to provide many initialization parameters in a prespecified form in a single data set and minimize the need for operator entry of parameters. For example, this data set might include parameters defining for the system resource manager (SRM) the minimum and maximum multiprogramming levels for a given domain. System tailoring by parameter selection also applies to certain application packages as IMS or CICS. System generation and tailoring is normally the job of the system programmer. As part of the capacity planning process, it is very important that the system programmer receive the required system performance data which relates to certain fixed and changed parameters. It is not expected that a system programmer would predict various performance changes due to a change in system parameter. But, systematically tracking and correlation of performance results with parameter selection will provide invaluable insight into system tailoring for performance. Historical data built-up in this fashion is the best input for predicting purposes.



PRINTER UTILIZATION BY JOB

FIGURE - 41

JOBNAME	PRINTER NUMBER	FORMS NUMBER	SET-UP TIME	LINES PRINTED	PAGES PRINTED	PRINT EXECUTION TIME

DETAIL PRINTER REPORT

FIGURE - 42

JOBNAME	START TIME	END TIME	NUMBER OF TAPE MOUNTS	TOTAL TAPE MOUNT TIME	NUMBER OF TAPE MOUNTERS

DETAIL MAGNETIC TAPE REPORT

FIGURE — 43

In a capacity planning environment, where a computer system has been analyzed and its useful life (satisfies the service objectives of the user community) predicted, it is normally assumed that the system is generally "well tuned". Incorrect selection of system parameters may create major system bottlenecks. Therefore, it is critical that the system programmer receive performance information on a continuing basis (relate parameter changes to system operation). In relation to the measurement tools (Figure 21), the data requirements for system programming are outlined in Figure 44 under 5 headings, system data and 4 subsystem areas BATCH, TSO, IMS, and CICS. The primary purpose of this data is tracking performance. As outlined in Figure 44, paging and swapping data, multiprogramming information and basic workload and resource utilization data by subsystem are required. In addition to some of the data formats already outlined (CPU utilization, channel utilization, etc.), Figure 45 thru 50 depicts data displays appropriate for the system programmer. The demand paging rate (Figure 45) plots the number of NON-VIO, NON-SWAP PAGE INS and PAGE RECLAIMS per second. The swapping rate (Figure 46) corresponds to the number of times per second that storage was swapped out and then swapped in. The channel activity rate (Figure 47) plots the number of successful Start I/O instruction issued per second to the channel. The device activity rate (Figure 48) corresponds to the number of successful Start I/O instructions per second issued to a device. Each of the four reports discussed above, are available for MVS users through the RMF plot report. Two other reports outlined in Figures 49 and 50 are complementary. The total number of TSO interactions by hour are plotted broken out into trivial, medium and complex type. Then Figure 49 plots the average response time being experienced by each type of interaction. Also, it should be noted that by Little's Law [13], the average number of interactions in the system by type or total (multiprogramming level for TSO) is equal to the load (interactions per second) times the average response time. These plots for loading and response time were produced by the IBM Service Level Reporter (SLR) program product. Although the plots outlined in Figures 49 and 50 only apply to the TSO environment, system programming should receive the same kind of transaction data for IMS or CICS applications. The sources for the system programming data is outlined in Figure 51.

- O SYSTEM DATA
  - O CPU UTILIZATION
  - O NUMBER OF SWAPS PER TIME PERIOD (IN AND OUT)
  - O NUMBER OF PAGES PER SWAP (IN AND OUT)
  - O DEMAND PAGING RATE (IN AND OUT)
  - O VIO PAGING RATE (IN AND OUT)
  - O AVERAGE MULTIPROGRAMMING LEVEL
  - O MAXIMUM NUMBER OF INITIATORS ASSIGNED (BATCH)
  - O AVERAGE NUMBER OF ACTIVE INITIATOR (BATCH)
  - O MAXIMUM NUMBER OF TSO ASSIGNED (TSO)
  - O AVERAGE NUMBER OF ACTIVE TSO USERS (TSO)
  - O MAXIMUM NUMBERS OF MPP ASSIGNED (IMS)
  - O AVERAGE NUMBER OF ACTIVE MPP (IMS)
  - O AMAX TASKS ASSIGNED (CICS)
  - O AVERAGE NUMBER OF ACTIVE TASK (CICS)
  - O CRITICAL CHANNEL UTILIZATIONS
  - O NUMBER OF SIO's PER CHANNEL
  - O CRITICAL I/O DEVICE UTILIZATIONS
  - O DEVICE ACTIVITY RATE (SIO's PER SECOND)
  - O AVERAGE DEVICE QUEUE LENGTH

- O BATCH DATA
  - O NUMBER OF JOBS BY TYPE PER TIME PERIOD
  - O EXCP's BY JOB TYPE*
  - O EXCP's BY JOB, CHANNEL, DEVICE
  - O CPU TIME BY JOB TYPES
  - O CPU TIME BY JOB
  - O ELAPSED TIME BY JOB

*TYPE (SHORT, MEDIUM, LONG)

DATA REQUIREMENTS FOR SYSTEM PROGRAMMING (PART 1)

FIGURE - 44

- TSO DATA
  - NUMBER OF INTERACTIONS BY TYPE BY PERIOD
  - EXCP's BY INTERACTION TYPES*
  - EXCP's BY CHANNEL, DEVICE
  - CPU TIME BY INTERACTION TYPE
  - CPU TIME BY INTERACTION
  - RESPONSE TIME BY INTERACTION TYPE (HOST, TERMINAL)

- IMS DATA
  - NUMBER OF TRANSACTIONS BY TYPE BY PERIOD
  - EXCP's BY TRANSACTION TYPE*
  - EXCP's BY TRANSACTION
  - EXCP's BY CHANNEL, DEVICE
  - CPU TIME BY TRANSACTION TYPE
  - CPU TIME BY TRANSACTION
  - RESPONSE TIME BY TRANSACTION TYPE (HOST, TERMINAL)
  - RESPONSE TIME BY TRANSACTION (HOST, TERMINAL)

- CICS NUMBER
  - NUMBER OF TRANSACTIONS BY TYPE BY PERIOD
  - EXCP's BY TRANSACTION TYPE*
  - EXCP's BY TRANSACTION
  - EXCP's BY CHANNEL, DEVICE
  - CPU TIME BY TRANSACTION TYPE
  - CPU TIME BY TRANSACTION
  - RESPONSE TIME BY TRANSACTION TYPE (HOST, TERMINAL)
  - RESPONSE TIME BY TRANSACTION (HOST, TERMINAL)

*TYPE (TRIVIAL, MEDIUM, COMPLEX)

DATA REQUIREMENTS FOR SYSTEM PROGRAMMING (PART 2)

FIGURE - 44

R E P O R T

```
START 03/14/78-08.43.41  INTERVAL 00.14.46
END   03/14/78-20.54.06  CYCLE 1.000 SECONDS

DEMAND PAGING RATE
      0      10.0      20.0      30.0      40.0      50.0
      +---------+---------+---------+---------+---------+
08:43 |*****    .         .         .         .         .
08:58 |************       .         .         .         .
09:13 |***************    .         .         .         .
09:28 |****************   .         .         .         .
09:43 |**************     .         .         .         .
09:58 |*****************  .         .         .         .
10:13 |***************    .         .         .         .
10:28 |**************     .         .         .         .
10:43 |*****************  .         .         .         .
10:58 |*******************.         .         .         .
11:13 |*****************  .         .         .         .
11:28 |*****************  .         .         .         .
11:43 |**************     .         .         .         .
11:58 |***********        .         .         .         .
12:13 |*************      .         .         .         .
12:28 |*********.         .         .         .         .
13:09 |*******  .         .         .         .         .
13:24 |**********         .         .         .         .
13:39 |*************      .         .         .         .
13:54 |*******************.         .         .         .
14:09 |*******************.         .         .         .
14:24 |****************   .         .         .         .
14:39 |*****************  .         .         .         .
14:54 |**********         .         .         .         .
15:09 |***************    .         .         .         .
15:24 |***********        .         .         .         .
15:39 |**************     .         .         .         .
15:54 |**************     .         .         .         .
16:09 |****************   .         .         .         .
16:24 |*************      .         .         .         .
16:39 |***************    .         .         .         .
15:54 |*******************.         .         .         .
17:09 |****************   .         .         .         .
17:24 |****************   .         .         .         .
17:39 |**************     .         .         .         .
```

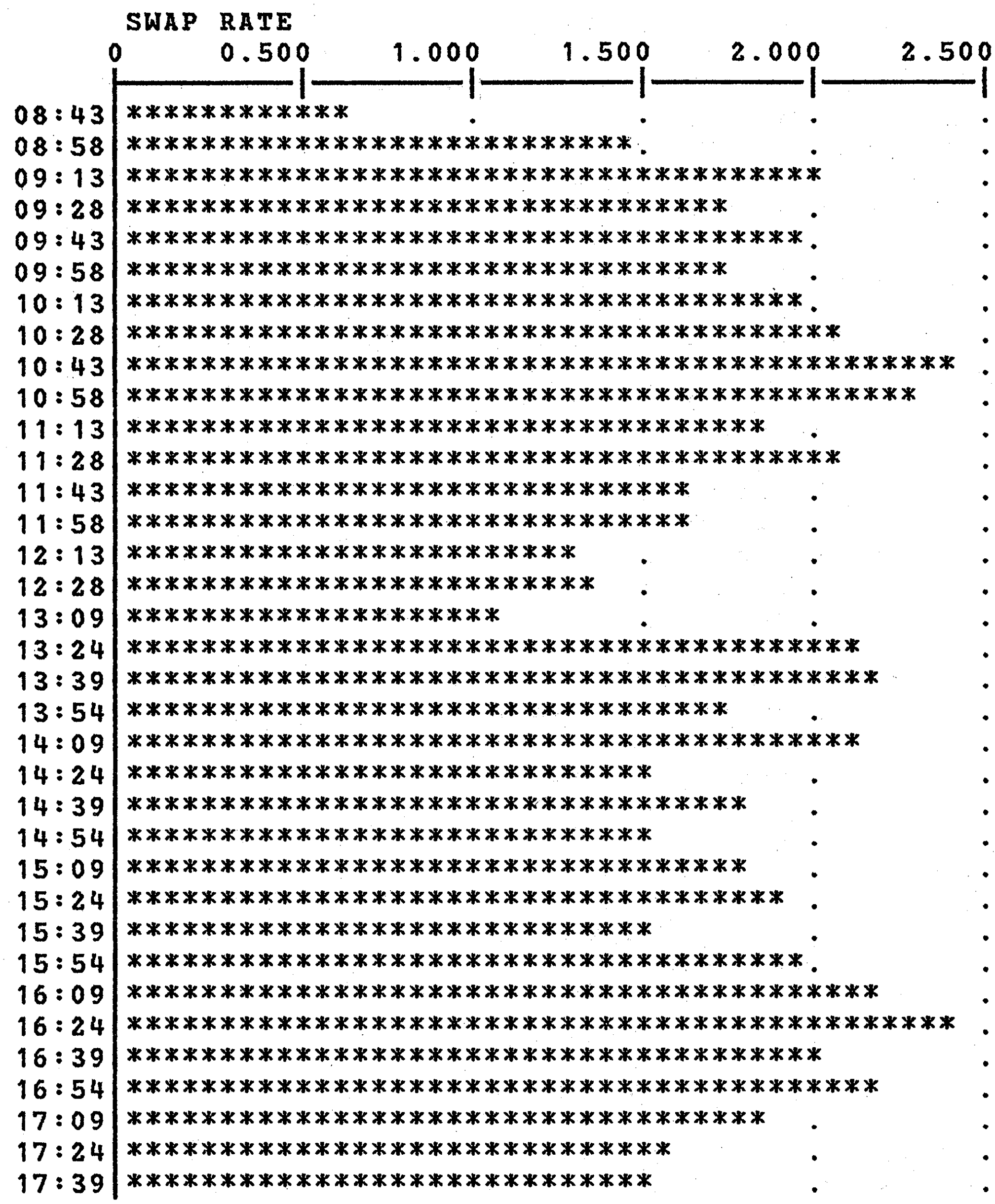
DEMAND PAGING RATE

FIGURE - 45

R M F   P L O T

OS/VS2
RELEASE 03.7A

SYSTEM ID T62P
RPT VERSION 03



SWAPPING RATE

FIGURE - 46

R E P O R T

```
START   10/03/77-08.04.32   INTERVAL  24.00.00
END     10/28/77-20.25.39   CYCLE   1.000 SECONDS

CPU 1 CHANNEL 01                  ACTIVITY RATE
0         10.0         20.0         30.0         40.0         50.0
|----------|------------|------------|------------|------------|
03  08:04|********  .           .            .            .            .
04  08:13|***********           .            .            .            .
05  08:14|**********            .            .            .            .
06  08:03|************          .            .            .            .
07  09:02|***       .           .            .            .            .
10  08:07|******    .           .            .            .            .
11  08:15|***********           .            .            .            .
12  08:16|*********.            .            .            .            .
13  08:07|******************    .            .            .            .
14  09:22|***********           .            .            .            .
17  09:17|************          .            .            .            .
18  08:11|************          .            .            .            .
19  08:16|***********           .            .            .            .
20  08:12|*******   .           .            .            .            .
21  14:31|*********.            .            .            .            .
24  08:28|**********            .            .            .            .
25  08:24|****************      .            .            .            .
26  08:07|**********            .            .            .            .
27  08:25|************          .            .            .            .
28  08:22|*************         .            .            .            .
```

CHANNEL ACTITIVITY RATE

FIGURE - 47

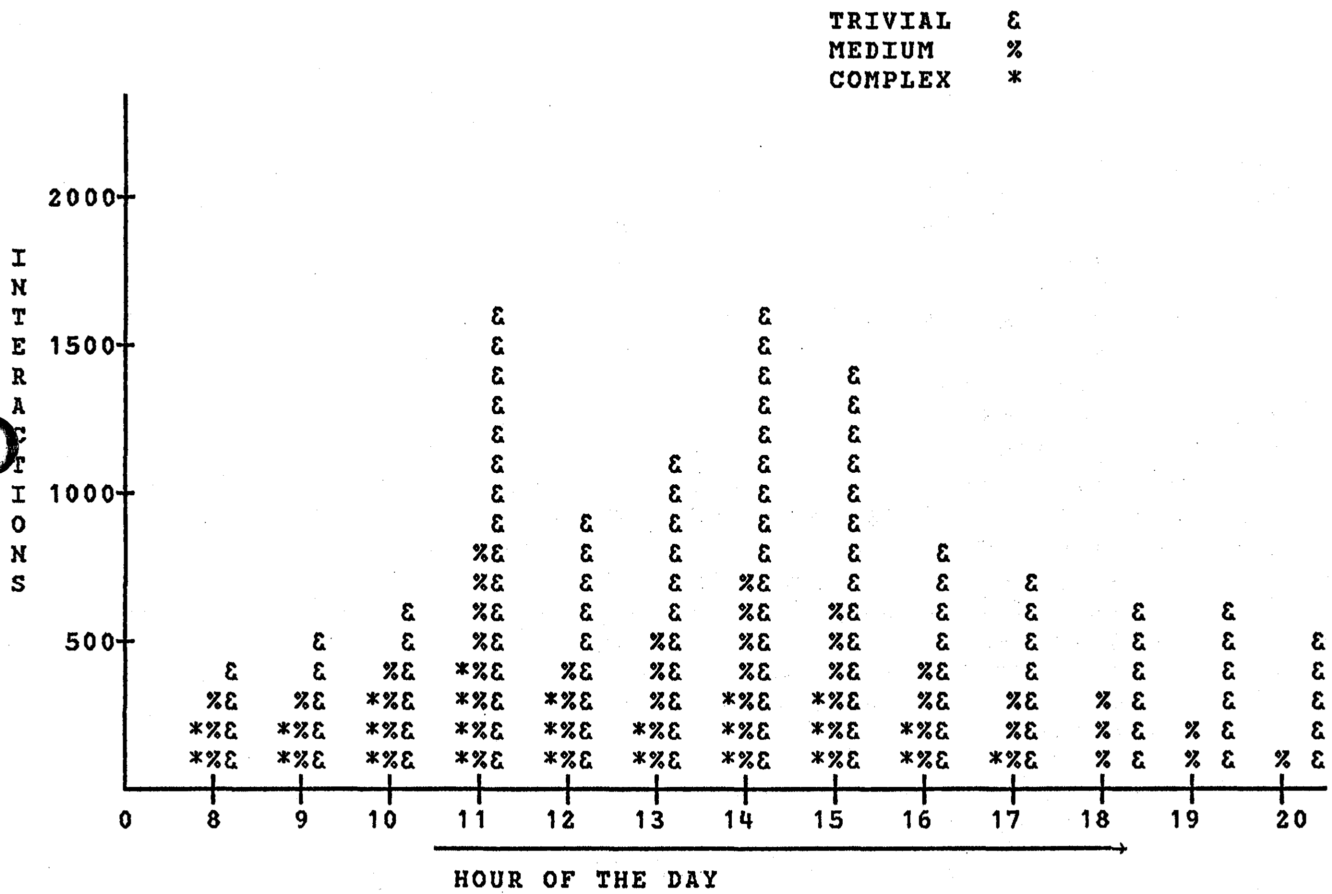
```
          R E P O R T

          START  10/03/77-08.04.32    INTERNAL 24.00.00
          END    10/28/77-20.25.39    CYCLE   1.000 SECONDS

          DEVICE 154                 ACTIVITY RATE
          0      10.0      20.0      30.0      40.0      50.0
          |---------|---------|---------|---------|---------|
03   08:04|******** .         .         .         .         .
04   08:13|***********        .         .         .         .
05   08:14|**********         .         .         .         .
06   08:03|************       .         .         .         .
07   09:02|***      .         .         .         .         .
10   08:07|******   .         .         .         .         .
11   08:15|***********        .         .         .         .
12   08:16|*********.         .         .         .         .
13   08:07|****************** .         .         .         .
14   09:22|***********        .         .         .         .
17   09:17|************       .         .         .         .
18   08:11|************       .         .         .         .
19   08:16|***********        .         .         .         .
20   08:12|*******  .         .         .         .         .
21   14:31|*********.         .         .         .         .
24   08:28|**********         .         .         .         .
25   08:24|****************   .         .         .         .
26   08:07|**********         .         .         .         .
27   08:25|************       .         .         .         .
28   08:22|*************      .         .         .         .
```

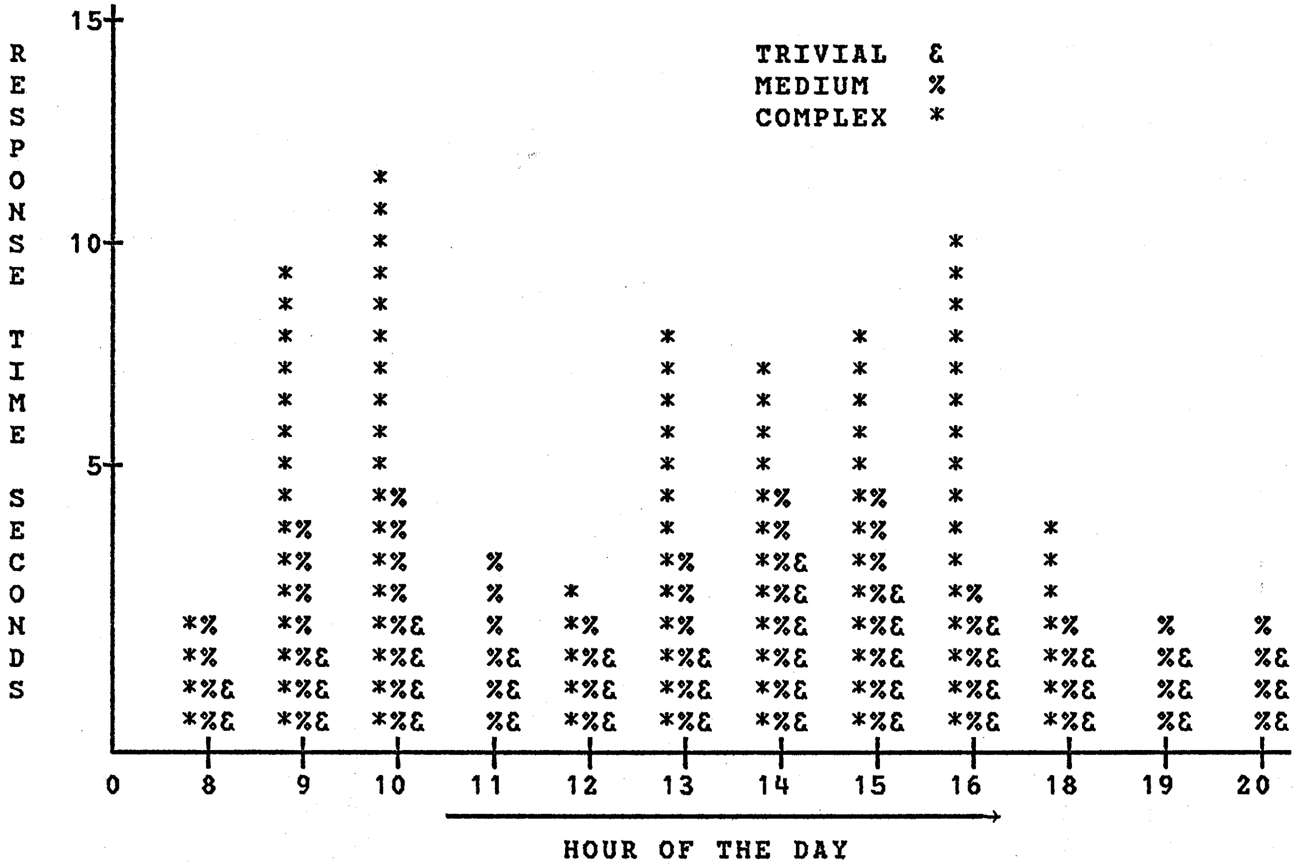
DEVICE ACTITIVTY RATE

FIGURE - 48



TOTAL NUMBER OF TSO INTERACTIONS BY TYPE

FIGURE - 49



AVERAGE TSO RESPONSE TIME

FIGURE - 50

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA-II	IMS LOG ANALYSIS	MANUAL LOGS	NEW REPORT
<u>SYSTEM DATA</u> CPU UTILIZATION		X					
NUMBER OF SWAPS (IN & OUT)	X						
NUMBER OF PAGES SWAP (IN & OUT)	X						
DEMAND PAGING RATE (IN & OUT)	X						
VIO PAGING RATE (IN & OUT)	X						
AVERAGE MPL			X	X	X		X
MAXIMUM NUMBER OF INITIATORS						X	
AVG. NUMBER OF ACTIVE INITIATOR	X		X				
MAX. NUMBER OF TSO USERS						X	
AVG. NUMBER OF ACTIVE TSO USERS	X						
MAXIMUM NUMBER MPP						X	
AVG. NUMBER OF ACTIVE MPP					X		
AMAX TASKS						X	
AVG. NUMBER ACTIVE TASKS				X			
CRITICAL CHANNEL UTILIZATION	X	X					

SOURCE OF SYSTEM PROGRAMMING DATA (PART 1)

FIGURE — 51

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA—II	IMS LOG ANALYSIS	MANUAL LOGS	NEW REPORT
<u>SYS. DATA (CONT)</u> NUMBER OF SIO's PER CHANNEL	X	X					
CRIT. I/O DEVICE UTILIZATIONS	X	X					
DEVICE ACTIVITY RATE	X	X					
AVG. DEVICE QUEUE	X						
<u>BATCH DATA</u> NUMBER OF JOBS BY TYPE PER PERIOD			X				
EXCP'S BY JOB TYPE	X		X				
EXCP'S BY JOB, CHANNEL, DEVICE			X				
CPU TIME BY JOB TYPE	X		X				
CPU TIME BY JOB			X				
ELAPSED TIME BY JOB			X				
<u>TSO DATA</u> NUMBER OF INTER-ACTIONS BY TYPE	X		X				
EXCP'S BY INTER-ACTION TYPE	X						
EXCP'S BY CHANNEL, DEVICE			X				

SOURCE OF SYSTEM PROGRAMMING DATA (PART 2)

FIGURE — 51

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA-II	IMS LOG ANALYSIS	MANUAL LOGS	NEW REPORT
<u>TSO DATA (CONT)</u> CPU TIME BY INTERACTION TYPE	X						
RESPONSE TIME BY INTERACTION TYPE	X						
<u>IMS DATA</u> NUMBER OF TRANSACTIONS BY TYPE					X		
EXCP'S BY TRANSACTION TYPE	X						
EXCP'S BY TRANSACTION	X						
EXCP'S BY CHANNEL, DEVICE			X				
CPU TIME BY TRANSACTION TYPE	X				X		
CPU TIME BY TRANSACTION	X				X		
RESPONSE TIME BY TRANSACTION TYPE					X		
RESPONSE TIME BY TRANSACTION					X		
<u>CICS DATA</u> NUMBER OF TRANSACTIONS BY TYPE				X			
EXCP'S BY TRANSACTION TYPE	X						

SOURCE OF SYSTEM PROGRAMMING DATA (PART 3)

FIGURE — 51

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA—II	IMS LOG ANALYSIS	MANUAL LOGS	NEW REPORT
<u>CICS DATA (CONT)</u> EXCP'S BY TRANSACTION	X						
EXCP'S BY CHANNEL, DEVICE			X				
CPU TIME BY TRANSACTION TYPE				X			
CPU TIME BY TRANSACTION				X			
RESPONSE TIME BY TRANSACTION TYPE				X			
RESPONSE TIME BY TRANSACTION				X			

SOURCE OF SYSTEM PROGRAMMING DATA (PART 4)

FIGURE — 51

One of the most difficult problems in the capacity planning process is predicting computer system requirements for new applications. The most readily used means of predicting new application requirements is comparing the proposed application to an existing application where certain performance parameters (CPU time required, response or elapsed time at various loadings, channel and device activity rates, etc.) are known. There are no nice neat equations/models available which output various performance requirement for a certain prescribed input. The only way to improve new application performance predictions is by measurement and maintenance of pertinent historical data files. For example, a new application can only be successfully compared to an existing application and its performance data used for prediction, when such data has been accurately measured and maintained. Also, validation of the method is possible only by measuring the performance parameters of the new application after it is implemented and comparing it to the old base application. Therefore, to improve new application predictions, the application development group should receive certain performance data on a continuing basis.

The data requirements for the application development department are outlined in Figure 52. The primary factors to be considered are workload to be processed (interactions/second, transactions/sec, number of batch jobs), CPU time or utilization, elapsed or response time and I/O loading. This reporting process is aimed at building a performance profile on critical batch and on-line applications. The source of this performance data is outlined in Figure 53.

- NUMBER OF INTERACTIONS PROCESSED BY APPLICATION
- NUMBER OF TRANSACTIONS PROCESSED BY APPLICATION
- CPU TIME BY APPLICATION
- ELAPSED TIME BY BATCH JOB WITHIN APPLICATION
- RESPONSE TIME BY INTERACTION WITHIN APPLICATION
- RESPONSE TIME BY TRANSACTION WITHIN APPLICATION
- EXCP's BY BATCH JOB WITHIN APPLICATION
- EXCP's BY TRANSACTION WITHIN APPLICATION
- CHANNEL AND DEVICE UTILIZATIONS BY APPLICATIONS, IF POSSIBLE
- TAPE MOUNTS BY JOB WITHIN APPLICATION
- PRINTER UTILIZATION BY APPLICATION, IF POSSIBLE
- SPECIAL PRINTER FORMS BY JOB WITHIN APPLICATION

REQUIREMENTS FOR APPLICATIONS DEVELOPMENT

FIGURE - 52

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA-II	IMS LOG ANAL	MANUAL LOGS	NEW REPORT
NUMBER OF INTERACTIONS PROCESSED BY APPLICATION (TSO)	X	X	X				
NUMBER OF TRANSACTIONS PROCESSED BY APPLICATION (CICS,IMS)				X	X		
CPU TIME BY APPLICATION	X	X	X	X	X		
ELAPSED TIME BY BATCH JOB WITHIN APPLICATION	X		X				
RESPONSE TIME BY INTERACTION WITHIN APPLICATION (TSO)	X						
RESPONE TIME BY TRANSACTION WITHIN APPLICATION(CICS,IMS)				X	X		
EXCP'S BY BATCH JOB WITIN APPLICATION	X		X				
EXCP'S BY TRANSACTION WITHIN APPLICATION	X		X				
CHANNEL & DEVICE UTILIZATIONS BY APPLICATATION (INDIRECT)	X		X				
TAPE MOUNTS BY JOB WITHIN APPLICATION						X	X
PRINTER UTILIZATION BY APPLICATION			X			X	X
SPECIAL PRINTER FORMS BY JOB WITHIN APPLICATION						X	X

SOURCE OF APPLICATION DEVELOPMENT DATA

FIGURE — 53

Understanding the users of the data processing facility is key to a successful capacity planning effort. The primary problem to be addressed is characterization of the user workload. This characterization is much more than quantifying the number of transactions being processed per hour. Characterization of a workload involves scheduling; when are the peak transaction loads, when must certain critical or heavy use batch jobs be run. Also, it may be necessary to define predecessor and feeder job requirements in a batch environment. These are only a few of the additional factors that may be required for user workload characterization. A characterization is a function of the industry (banking, retail, petroleum, etc.), user (clerk, technician, manager, etc.), and the actual customer within an industry. Workload characterization is an art and in most instances will require variations as greater knowledge of the user and his application is gained by ongoing tracking of the environment.

With the measurement tools outlined in Figure 21, the data requirements necessary to aid in understanding the user environment is given in Figure 54. The two critical requirements outlined in this figure are the necessity of obtaining loading data (numbers of transactions by type over a period of time) and the CPU service required to process the load. It is necessary to have this service information for certain critical channels and I/O devices. However, it may not be necessary to return this information to the user. But, this is valuable information for the capacity planner. As noted at the top of Figure 54, data will be categorized by function (production, testing, maintenance) and departments within function. The sources of this user data is given in Figure 55.

The primary purpose of user data is to provide a means of validating and tracking user growth projections. Users will be required to make their growth projections and the data outlined in Figure 54 will provide the necessary feedback. This reporting and interaction between the capacity planning group and users will provide the required input for good user growth projections. The growth discussed in this paragraph is that termed "natural growth" (growth of existing applications) as opposed to new applications which was previously discussed.

Data will be categorized by function (production, testing, maintenance) and each category broken down by user departments, groups, etc.

- TOTAL CPU HOURS CONSUMED
- TOTAL CPU HOURS CONSUMED BY SUBSYSTEM
  - BATCH, TSO, IMS, CICS
- AVERAGE CPU TIME CONSUMED BY CRITICAL APPLICATIONS
- CRITICAL BATCH JOB EARLIEST START AND LATEST END
- CRITICAL BATCH JOB START AND END TIMES
- RESPONSE TIME OBJECTIVE BY INTERACTION TYPE (TSO)
- RESPONSE TIME BY INTERATION TYPE (TSO)
- RESPONSE TIME OBJECTIVE BY TRANSACTION TYPE (IMS, CICS)
- RESPONSE TIME BY TRANSACTION TYPE (IMS, CICS)
- RESPONSE TIME OBJECTIVE FOR CRITICAL TRANSACTIONS (IMS, CICS)
- RESPONSE TIME FOR CRITICAL TRANSACTIONS (IMS, CICS)
- NUMBER OF BATCH JOBS BY TYPE
- NUMBER OF INTERACTIONS BY TYPE
- NUMBER OF TRANSACTIONS BY TYPE
- NUMBER OF CRITICAL TRANSACTIONS

DATA REQUIREMENTS FOR DP USERS

FIGURE - 54

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA-II	IMS LOG ANAL	MANUAL LOGS	NEW REPORT
TOTAL CPU HOURS CONSUMED		——(SEE FIGURE - 29)——					
TOTAL CPU HOURS CONSUMED BY SUBSYSTEM, APPLICATION		——(SEE FIGURE - 29)——					
AVERAGE CPU TIME CONSUMED BY CRITICAL APPLICATION			X	X	X		
BATCH EARLIEST START, LATEST END						X	
BATCH START AND END TIMES			X				
RESPONE TIME OBJECTIVE BY INTERACTION TYPE						X	
RESPONSE TIME BY INTERACTION TYPE	X						
RESPONSE TIME OBJECTIVE BY TRANSACTION TYPE						X	
RESPONSE TIME BY TRANSAGTION TYPE				X	X		
RESPONSE TIME BY CRITICAL TRANSACTIONS				X	X		
NUMBER OF BATCH JOB BY TYPE			X				
NUMBER OF INTERACTIONS BY TYPE	X						
NUMBER OF TRANSACTIONS BY TYPE				X	X		
NUMBER OF CRITICAL TRANSACTIONS				X	X		

SOURCE OF DATA FOR DP USERS

FIGURE — 55

The data to be reported to upper management (DP and Corporate) must be clear, concise and represent the most pertinent factors (workload, user service, availability, etc.) on system performance. In most cases, when a DP installation finds its system is out of capacity upper management is the last to know. Then, the primary reason for informing upper management is that they must sign or approve the order for a new CPU. One of the primary purposes of the capacity planning process is to keep upper management informed on a continuing basis as to the status of the available system capacity. Therefore, equipment requests will not be a surprise.

Using the measurement tools outlined in Figure 21 as the primary data gathering instrument, the data required to keep upper management informed on a continuing basis is outlined in Figure 56. A good constraint for reporting data might be that any information not clearly contained on a single page report will not be reported to upper management. Several installations are moving to a summarization of their measurement data specifically to be a one page report for their management. In working with some of these accounts, the data items thought to be critical are outlined in Figure 56. In the case of workload, it is very important that upper management gain some perspective on the number of transaction or batch jobs being processed. This indicates a growth in workload and a gradual consumption of resource or system capacity. It would be preferrable to relate workload to the natural business units (e.g., new accounts, checks, engines, etc.). This would provide upper managers with a better frame of reference than growth in batch jobs or transactions processed. However, in most instances DP installations do not have the historical data (natural business units correlated with data processing units) required to support such a reporting scheme. Although a CPU is available across some period of time (week, month, year), availability numbers (hardware, software, user perception) will provide basic information on just how much time the CPU is actually available to do problem program work. Also, a critical item to the reporting process for upper management is the question of user service. The reporting process must include user service objectives (response/turnaround times) as well as the actual values at the current transaction or job volumes. Upper management must begin to associate a system's capacity with the workload to be processed and the user service being provided. In certain instances, other data might be reported to upper management on a one time basis. This data is recorded under miscellaneous in Figure 56. The sources for the management data is outlined in Figure 57.

ONE PAGE REPORT

- O WORKLOAD
  - O NUMBER OF BATCH JOBS BY TYPE
  - O NUMBER OF INTERACTIONS BY TYPE
  - O NUMBER OF TRANSACTIONS BY TYPE
  - O NUMBER OF CRITICAL TRANSACTION TYPES
- O AVAILABILITY
  - O AVAILABLE CPU HOURS
  - O CPU HOURS CONSUMED (TOTAL)
  - O CPU HOURS CONSUMED BY SUBSYSTEM
- O USER SERVICE (OBJECTIVE AND ACTUAL)
  - O TURNAROUND TIME OBJECTIVE FOR CRITICAL BATCH JOBS
  - O TURNAROUND TIMES FOR CRITICAL BATCH JOBS
  - O RESPONSE TIME OBJECTIVE BY INTERATIONS TYPE
  - O RESPONSE TIMES BY INTERATION TYPE
  - O RESPONSE TIME OBJECTIVES BY TRANSACTION TYPE
  - O RESPONSE TIMES BY TRANSACTION TYPES
  - O RESPONSE TIME OBJECTIVE FOR CRITICAL TRANSACTIONS
  - O RESPONSE TIME FOR CRITICAL TRANSACTIONS
- O MISCELLANEOUS SECTION
  - O MAJOR NEW SYSTEMS INSTALLED DURING REPORTING PERIOD (HARDWARE OR SOFTWARE)
  - O MAJOR SYSTEM CHANGES DURING REPORTING PERIOD
  - O AVAILABILITY PROBLEMS
    - O HARDWARE
    - O SOFTWARE
    - O APPLICATION (e.g. DATA BASE RECOVERY)
  - O SIGNIFICANT UNRESOLVED PROBLEMS
  - O MAJOR SHIFTS IN WORKLOAD FROM ONE DATA CENTER TO ANOTHER
  - O IDENTIFY SENSITIVE NEW APPLICATIONS/PROJECTS
  - O FROM A SCHEDULING POINT OF VIEW IDENTIFY SPECIFIC TIME WINDOW CONSTRAINTS
    - O PREDECESSOR/POST JOB REQUIREMENTS
    - O ETC.

DATA REQUIRED FOR MANAGEMENT

FIGURE - 56

DATA ITEM	RMF REPORT	RMF PLOT REPORT	SMF REPORT	CICS PA-II	IMS LOG ANAL	MANUAL LOGS	NEW REPORT
NUMBER OF BATCH JOBS BY TYPE	X		X				
NUMBER OF INTERACTIONS BY TYPE	X						
NUMBER OF TRANSACTIONS BY TYPE				X	X		
CPU HOURS CONSUMED	X		X	X	X		
AVAILABLE CPU HOURS (COMPUTED)	X					X	X
USER SERVICE	(SEE FIGURE 29)						

SOURCE OF DATA FOR MANAGEMENT

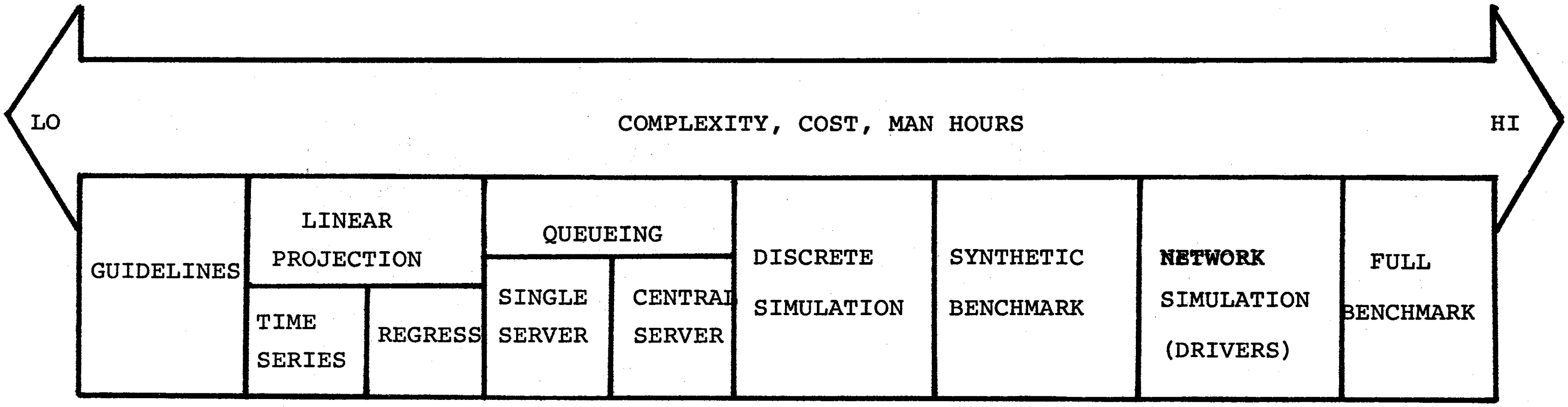
FIGURE — 57

The capacity planning group will be the co-ordinator, controller and maintainer of the data outlined for the various groups above. Since the primary tool of the capacity planner is some type of model used for predictive purposes, all the data discussed thus far is of interest, especially as it relates to a definition of the computer system and the workload it must process. It is the ongoing data gathering and co-ordination that will improve the model and its predictive capabilities. Modelling and its relation to the overall capacity planning process is discussed in detail in the following section.

2.2.4 Performance Analysis for Capacity Planning

In the preceding section, the data requirements to drive the capacity planning process were discussed. The primary purpose of the data gathering activity was to obtain a better understanding of the operation of the data processing installation. As a part of the capacity planning effort, there is the requirement to provide data to develop models for prediction, for input and to validate these models. Models may take many different forms and have varying degrees of detail. Models may be very simple in that they are basically guidelines developed by monitoring the operation of the computer system over time. For example, when the available number of TSO users exceed 40 on a system, operations will normally receive more user complaints. A workload of a certain number of transactions per second might be associated with this level of user discontent. This can be thought of as a model to be used as an aid in analyzing the TSO environment. Then, for a more costly and detailed form of modelling, a given computer system might be structured with the actual hardware and software, where the actual application programs will be run for a period of time and certain performance data collected and analyzed. This is termed a benchmark and in many situations affords the most accurate modelling and predictive process. Accuracy would imply that the applications were well defined and integrated on the hardware and software to be used in production.

A spectrum of performance analysis techniques which can be used for capacity planning are outlined in Figure 58. As indicated in the figure, as one moves across the spectrum complexity, cost and man hours increase with respect to the technique used. No indication is made of the fact that accuracy will increase as one moves across the spectrum. For example, referencing benchmarking at the high end, to perform a benchmark it is necessary to select that subset of applications and workload from the DP environment which characterizes the total operation. Then, this subset must be taken and processed on a computer system as close as possible to the actual hardware and software being proposed.



SPECTRUM OF
PERFORMANCE ANALYSIS TECHNIQUES

FIGURE 58

After the benchmark has been accomplished, it is necessary to analyze the results and reason what the results of this subset environment means in terms of the total. In most instances, only installations with a great deal of understanding of their DP operations can adequately use the results of benchmarking as their only capacity planning tool. Also, in many cases, the level of understanding a DP installation has about their environment would dictate a much less costly and complex means of analysis. It has been shown in several cases that DP installations using simple guidelines, monitoring their system on a continuing basis and using linear projections for future requirements are doing very credible capacity planning. Also, included in this spectrum of analysis techniques are statistical processes using linear time series or regression analysis. Queueing analyses depicted here are single server and closed queueing models of the central server type.

In many instances, the accuracy of single server queueing models are questioned. But, there are many analyses or environments where the model accuracy is not the real question. Because, many of the parameters required to develop the model are in suspect of being grossly inaccurate and in some cases unknown. What is needed at this point is a simple modelling technique (more functionally adequate than theoretically correct) that provides the capability of stepping a workload through the DP environment and basically being able to relate to the relationships between the model and computer system. This analysis technique is viewed as being much closer to "guidelines" or an "empirical" type of analysis. As you move away from the single server analysis in moving across the spectrum (Figure 58), it becomes increasingly more difficult to compare internal model operations to actual system operations. It is just the fact that the model is becoming internally more complex and normally requires a computer for solution. Many single server models are also computerized but one is not as completely overwhelmed by the model complexity. In modelling computer installations, it is clear that the dynamics (continuously changing environment) and interactions (people, hardware, software, etc.) presents a complexity that regardless of the theoretical accuracy of a model, it is impossible to model what is not understood. Hence, the capacity planning process is an attempt at gaining a better understanding of the overall data processing environment.

2.2.5 Initiation of The Capacity Planning Process

To initiate a capacity planning effort requires that an implementation plan be developed. Three critical items to be included in this plan are outlined in Figure 59. Normally, there is some form of capacity planning or performance analysis being performed within a DP installation. The activity uses measurement tools with reports being generated and circulated throughout the organization. This being the case, it is not reasonable to think that a capacity planning effort can be initiated without understanding the current process. From a political and economical point of view, it may be necessary to retain functions already implemented or use people involved in the current process. For example, reporting is one of the critical parts of capacity planning and may require that a report already in existence be inhanced rather than replaced. This would imply that the reporting process can not be developed in a vacuum and suddenly imposed on people already receiving various reports. It is very important that the personnel involved in the capacity planning process are receptive to its concepts and reporting mechanisms. In its initial stages, reception of the capacity planning effort as a viable process is very critical to its life. Therefore, a reasonable amount of effort should be expended in studying and flowing out the current "Capacity Planning" process, the types of measurement tools being used, data being gathered and reporting formats. The recepients of the various reports should be established as well as the functions they are performing with the reported data.

The next item of concern in initiating a capacity planning effort is the workload and CPU consumption. It should not be misunderstood when it is recommended that the CPU be the only resource considered. It is very clear that a complete capacity planning program will eventually include channels, I/O devices and the network. But, the initiation of capacity planning effort should be kept as simple as possible. Therefore, focusing on the CPU is aimed at simplifying this process. Also, this is not unreasonable if the approach includes a certain amount of tuning for other auxiliary devices. In essence, the idea is to remove all bottlenecking so the CPU would only be constrained by the workload it is required to process.

1. DEVELOP A PLAN

2. DESCRIBE IN DETAIL FLOW YOUR CURRENT CAPACITY PLANNING INFORMATION PROCESS

   - O INSTALLED MEASUREMENT TOOLS
   - O FORMAT OF VARIOUS REPORTS
   - O RECIPIENTS AND ACTION TAKEN ON VARIOUS REPORTS

3. OUTLINE IN AS GREAT A DETAIL AS POSSIBLE WORKLOAD AND CPU REQUIREMENTS BY APPLICATION AREAS AND SUBSYSTEM (BATCH, IMS, ETC.)

   - O WORKLOAD CHARACTERIZATION (BY LOGICAL SHIFT)*
   - O CPU UTILIZATION (BY LOGICAL SHIFT)

4. PERFORMANCE ANALYSIS

   - O DETAILED/SUPERFICIAL
     - O CURRENT SYSTEM PERFORMANCE
     - O PREDICTIONS/FORECASTS
       - O INCREASED WORKLOAD
       - O NEW WORKLOAD

* LOGICAL SHIFT - A PERIOD OF TIME IN THE 24 HOUR DAY THAT CORRESPONDS TO A DATA PROCESSING FUNCTION (E.G. ON-LINE PROCESSING - 7:00 A.M.-7:00 P.M.) RATHER THAN THE NORMAL PERSONNEL SHIFT (E.G., 8:00 - 5:00 P.M.)

CAPACITY PLANNING IMPLEMENTATION

(INITIAL EFFORT)

FIGURE - 59

2.3 The USAGE Technique

The USAGE (Understand your System and Application Growth Environment) technique is a capacity planning process which focuses primarily on the CPU and is an excellent way to initiate a capacity planning effort. The USAGE program has been used extensively in 1977 and 1978, with over 200 joint IBM/customer workshops completed. In this section, we will discuss:

- Potential Benefits and Purpose
- Study and Terminology
- Organizational Phase
- Current Workload Analysis
- Forecasting Future Workload
- Configure to meet the Forecast

2.3.1 Potential Benefits and Purpose

USAGE is a methodology used to forecast growth in CPU capacity requirements over a period of 1 to 2 years. It is based on readily available data (SMF records). Experience has shown USAGE to be easy-to-use, easy-to-understand, and a very effective forecasting tool.

It generally can lead to better planning and the methodology itself can be used as a continuing planning/tracking vehicle.

A successful implementation of USAGE will provide the following:

1. An accounting for all computer time used by an installation during a given sample period, ususally one month. This will include all application programs, and other system functions.

2. An understanding of the current workload with a breakdown among production work (application, both batch and online), development work (enhancements and new applications), and support work (maintenance, reruns, reports, etc.).

3. A basis (from 1 and 2 above) for forecasting future growth of current work and new workload to be generated by present application development, or changes in the business served by the DP system.

USAGE is a methodology intended for management use; and therefore, understanding the total workload and forecasting for future requirements are the keys to the exercise. Implicit in the above statement is the requirement that you complete the exercise, including the forecasting step. As you will see, USAGE has provisions and guidelines to estimate missing data and future workload requirements to allow you to complete the study in a relatively short period of time.

However, it is important to put USAGE in perspective. We have discussed what it is used for, and can identify what it is not used for:

- Tuning - USAGE is not a tuning tool and will not assist you in this regard. USAGE assumes at the outset that your installation is an "average" tuned shop, and it is expected that you will continue to tune the installation.

- Long-Range Planning - USAGE is best suited for a 1 to 2 year forecast cycle. It is not intended as a

replacement for a 5 to 7 year plan or as a technique to establish a corporate DP strategy (such as an Executive Strategy Session).

- Simulation & Modeling - USAGE will not provide the accuracy nor investigate the detail provided by a simulation or modeling exercise designed for specific performance analysis requirements.

2.3.2 Study and Terminology

The USAGE study involves analyzing current data to understand how the data processing equipment is being used, to forecast the effect of future growth, and to configure equipment that will satisfy that capacity requirement. In this section, we will introduce the terminology employed in a USAGE study and review the process.

Before we work through the USAGE methodology in a step-by-step fashion, it will be useful to look at some terminology. The purpose of this section is to define and understand those basic concepts which are an essential part of the USAGE approach.

Analysis Period

The analysis period is simply the unit of time for which you intend to study and collect data. With USAGE, an analysis period of one month seems to be the best choice, both in terms of being long enough to provide a representative sample of programs and data, and yet short enough to allow for an easily managed project.

SMF Time

Recorded SMF CPU time is the basic source of data for a USAGE study. Specifically, SMF records Type 1 (System Wait Time), Type 4 and 5 (Problem Program Time), and Type 34 and 35 (TSO Time) are the records which are utilized. If MVS is the SCP, then Type 70 records replace the Type 1 records.

Production Time Periods

Earlier it was indicated that a USAGE study is based on an analysis period of one month. From a month's worth of SMF data you can determine what your average CPU workload requirements were for that month, and you can forecast average CPU workload requirements for future months.

From the standpoint of understanding your capacity requirements, an average monthly workload number is probably not very useful to you. As you know, your workload (and the utilization of the CPU resource) will significantly vary during a month, depending upon such factors as:

- The time of day (2 p.m. vs. 2. a.m.)
- The day of the week (Monday vs. Sunday)
- When the online systems are available.

What is useful, therefore, is to separate the collected SMF data into several Production Time Periods, which can now be defined as units of time representing logically different units of work. The selection of appropriate production time periods is very important to the successful implementation of USAGE, and, of course, will be determined based on the customer environment. The choice, however, should be based on the following two major elements:

1. The characteristics of the workload - for example, a heavy online period of time versus a primarily batch-oriented period.

2. The volume of the workload - for example, a heavy batch-oriented workload period versus a light batch-oriented period.

The length of the production time periods should be in the range of 6 to 12 hours, and three to five periods should be adequate for most situations. An example of a month separated into useful production time periods might be:

- Weekdays, 8 a.m. to 5 p.m. (primarily on-line)
- Weekdays, 5 p.m. to 1 a.m. (heavy batch)
- Weekdays, 1 a.m. to 8 a.m. (light batch)
- Weekdays and Holidays (light batch)

Wall Clock Time

This is simply the chronological time which will help you compute and understand CPU availability and CPU utilization for a particular production time period. For example, if you choose a weekday time period of 8 a.m. to 4 p.m. and there were 21 workdays in the month you are evaluating, that period's Wall Clock Time would be calculated as follows: 8 hours times 21 workdays = 168 Wall Clock Hours.

Elapsed Time

Elapsed time is defined as the time that CPU is available to do work, and can be computed from SMF record Type 1 (Type 70 in MVS). If during a specific time period no downtime or IPL's occured, elapsed time would equal Wall Clock Time (168 hours in the previous example). A simple measure of availability can be obtained by dividing elapsed time by Wall Clock Time for a particular time period.

Wait Time

Wait time can be defined as the amount of CPU time that a CPU is available but is not processing work. Wait time is accurately recorded by SMF Type 1 (Type 70 in MVS), and is used to calculate total CPU time of a particular time period.

Total CPU Time

Total CPU time is the amount of time the CPU actually processes work during a particular time period. By subtracting Wait Time from Elapsed Time you can calculate total CPU Time. Following our earlier example, and assuming Wait Time to be 40 hours, 168 elapsed hours minus 40 wait hours = 128 total CPU hours used during that 8 hour production time period for that month. By dividing 128 total CPU hours by 168 elapsed hours, you have a measure of percent utilization (76% in our example) for that period. Utilization computed in this way should be consistent with measurements obtained from other sources, such as a hardware or software monitor.

Paging Time

Paging time is the portion of total CPU time generated by demand paging. Demand paging is the paging which the system does to dynamically manage real memory. Where real memory is assumed to be less than the specified amount of virtual memory. Demand paging does not include functional paging which is caused by TSO swapping and/or VIO. To calculate paging time, the demand paging rate (i.e., number of pages in and out per second) must be estimated. The demand paging rate can be measured with tools such as RMF, MF1, SVSPT, VS1PT, etc. The paging time is calculated as the product of the paging rate and the CPU processing time per demand page per second. The CPU processing time per page per second is relatively constant for a given operating system release and CPU (e.g., MVS 3.7 on a 3033, SVS 1.7 on a 158, etc.). Generally, demand paging is a separate line item in the usage study. Demand paging should be measured for each period since it varies by workload mix and is normally different from period to period.

Business Elements

Most DP managers have some idea as to their CPU utilization. However, they tend not to know the degree to which particular applications or business areas contribute to that amount of utilization in a month. In other words, they don't know that payroll requires 4 hours of CPU time per month, credit management requires 8 hours, TSO requires 20 hours, batch testing requires 30 hours, and so on.

In order to understand current capacity requirements and, more importantly, to be able to forecast future requirements, it is critically important to segment a total workload into smaller understandable units. Within the USAGE methodology, the term Business Element was chosen to describe these workload units. The careful selection of appropriate Business Elements is an important part of the USAGE approach.

In choosing Business Elements in your environment, you should identify three major groups as follows:

1. Production Work - Production work should account for about 70% of the total CPU time utilized by your shop. Within the production areas, you should identify 5 to 10 business elements that account for the majority of the workload. The choice of unique elements will be based on your business (a bank will have different elements than a chemical company), whether the applications are online or batch, and whether the elements are likely to grow at different rates.

2. Testing - this is another major unit of work that probably requires about 20% of your CPU resource. Testing for new application development is evaluated as a separate element because it represents future production work as well as being a significant load on the current system. Business Elements under the major testing heading should be separated by online and batch testing, and by major new applications project.

3. Operations Support - the operations area should be analyzed separately because it represents neither production work, nor future production work. However, it is necessary to the functioning of the shop and it does require a significant part of the CPU resource (probably about 10%). Items included under the operations heading are:

- Reruns
- Job Scheduling
- System Programmer Time
- Data Base reorganizations
- Tape Management Systems
- Maintenance Programming
- Job Accounting Routines
- Performance Tools

<u>Capture Ratios</u>

SMF does not record all the CPU time expended on behalf of a particular job. Its purpose is to be consistent rather than complete, in order to satisfy requirements of charge back systems. In addition the portion of total time captured by SMF varies with the workload type and the specific SMF implementation for the various SCP's.

If only one application is running in a system, the capture ratio can be measured. It is the total SMF Task Control Block (TCB) time divided by the Total CPU Time less the Paging Time. If the SCP is MVS and Service Request Block (SRB) time is being measured and used, then the capture ratio is the TCB time plus SRB time divided by the Total CPU Time less the Paging Time.

The capture ratio varies between approximately 0.25, or 25% captured, and .97, or 97% captured. Included in the uncaptured time is scheduling done by HASP, JES2, and JES3, attention processing time for TSO, and some IOS time. Thus the capture ratio can help you allocate uncaptured CPU time to specific applications.

It is important to note that True CPU hours is the real measure of job CPU utilization. For example, if batch and TSO both have the same amount of SMF hours, it is possible for the TSO load to be twice as large in True CPU hours. Assume that batch has a capture ratio of 0.8 and that TSO has a capture ratio of 0.4. Then, for the same amount of reported SMF TCB time, TSO has twice as much True CPU time as batch. For example:

Workload	SMF TCB Hours	CR	True CPU Hours
TSO	10	.4	25
Batch	10	.8	12.5

The True CPU hours, after adjustment via the capture ratios, more accurately describe the relative impact of the two types of workload. Consequently, forecasting based on the True CPU time will more accurately reflect the effects of a changing workload profile and lead to better conclusions.

The summatation of all True CPU time should approximate the measured Total CPU Time. Paging Time, as previously defined, is a true CPU time value. If the calculated amount is within 5% of the measured amount, the correspondence is good. If it is within 10%, it may be acceptable. If the error is larger, then further analysis is required.

If the calculated True CPU time is low, then SMF may be missing some portion of the workload, or the capture ratios may not adequately reflect the workload characteristics. If the calculated True CPU time is high, then the capture ratios should be investigated. After capture ratios are chosen, they should not be changed unless there is a very good reason. If the capture ratio is changed, it should be carefully noted since it will be difficult to compare future USAGE studies to past ones.

2.3.3 Organizational Phase

The following section describes how to organize the actual data collection/reduction study. Several members of the DP staff should become involved in the data gathering. First, someone who can handle the SMF and related job accounting information is key. Also, a certain amount of time from several DP managers, such as application development, operations, systems programming, security, data base, and end-user interface managers will be required. These individuals can contribute vital information about how things are operating, how much effort is spent on development, application schedules, equipment changes, future plans and other significant performance or operational constraints.

Objectives of the Study

One central objective is to identify the major business elements now in production. These should be the same as used by top management to describe what DP does. In other words, you list a small number of business elements which account for 80-90% of the production workload and which are understood by the entire organization. The basic need is for communications. Don't use code names or designations that are strictly DP terms. Use common business oriented nomenclature. Names like "Engineering Design", "Accounts Payable", "Sales Analysis", or "Bill of Materials", are

functions that are descriptive and widely understood within the firm.

The identification of time periods should be based on how the system is operated. If there are significant difference, like online periods versus batch only, then they should define the time period. Weekends and long periods of repeated peak workload are meaningful to identify as individual time periods. For instance, if there are two or three extremely busy work days each month with very stringent deadline or response time requirements, then you should identify this period as a separate data collection period. For example, in the banking industry the "proof and transit" application has a 2 a.m. deadline which might be defined as the boundary between the second and third periods. Your planning must recognize the unique as well as the average workloads.

The study will need to identify the availability and quality of data. This should include the job accounting standards, identification of batch and interactive testing, maintenance versus enhancement workload, operations support, application rerun time, and any "special" workload which may be included in the collected data.

The month chosen for collection of SMF should have some general characteristics. Use a normal month for the study. If every month is different, then it will not matter. But, if it is a retailer who specializes in toys, you will not want to use December as a normal month. In fact you may want to use December, but as a special time period, which requires special handling in respect to capacity. Another month's data can be used for the normal workload measurement. Avoid, if possible, a period which had unusual "change" in the DP environment. Don't use the first month after installing a new SCP, or processor, or major I/O subsystem like mass storage. Also, avoid any period which had a severe down-time problem, or significant vacation/holiday impact.

Forecasting should always involve management input. Questions about the future are answered by management decisions made now and in the future. You should plan to include all DP managers so the forecast will be developed in a "real time" environment.

The total time frame for a USAGE study should be one to two months from start to finish.

People time required for a good study is approximately four weeks for technical support, with one week of management time, plus one full day for all DP management at the workshop.

There are certain key time qualifiers which can extend the time frames outlined above. The first qualifier is the quality and availability of SMF data. In a few cases, the control on these data may not be adequate. You need a month's worth of processable data. The definition of Business Elements and Production Time Periods may be lengthy if the business environment and operations are not clearly understood by the data gathering team. Job naming conventions may be non-standard; and therefore, require some extra effort in data reduction. Finally, the level of understanding of SMF and normal operating procedures will affect the data collection and the data reduction phases of the study.

At the end of the organizational phase of the USAGE study you should have agreed on:

- Terminology
- Business Elements
- Production Time Periods
- Participants
- Work Assignments.

2.3.4 Current Workload Analysis

In the USAGE Study Planning Meeting the Business Elements, the Period of Analysis, the Production Time Periods were selected. Definitions and terminology were also agreed upon. During the analysis you should review the results of the planning meeting as well as the data reduction programs used. In the data gathering and tabulating steps, when problems are found with the original data, definitions, and assumptions, keep a list of these problems, new assumptions or methodology changes for review in the workshop, because they may substantially change the degree of confidence that customer management has in the results.

The analysis of the systems current workload is divided into serveral steps:

1. Find total elapsed time in month.
2. Find total wait time in month.
3. Calculate Total CPU Time and utilization percentage.
4. Gather raw SMF time.

5. Allocate raw SMF time to Business Elements.

6. Divide Business Element SMF time into Production Time Periods.

7. Assign a Capture Ratio to each Business Element.

8. Calculate True CPU Time by Business Element.

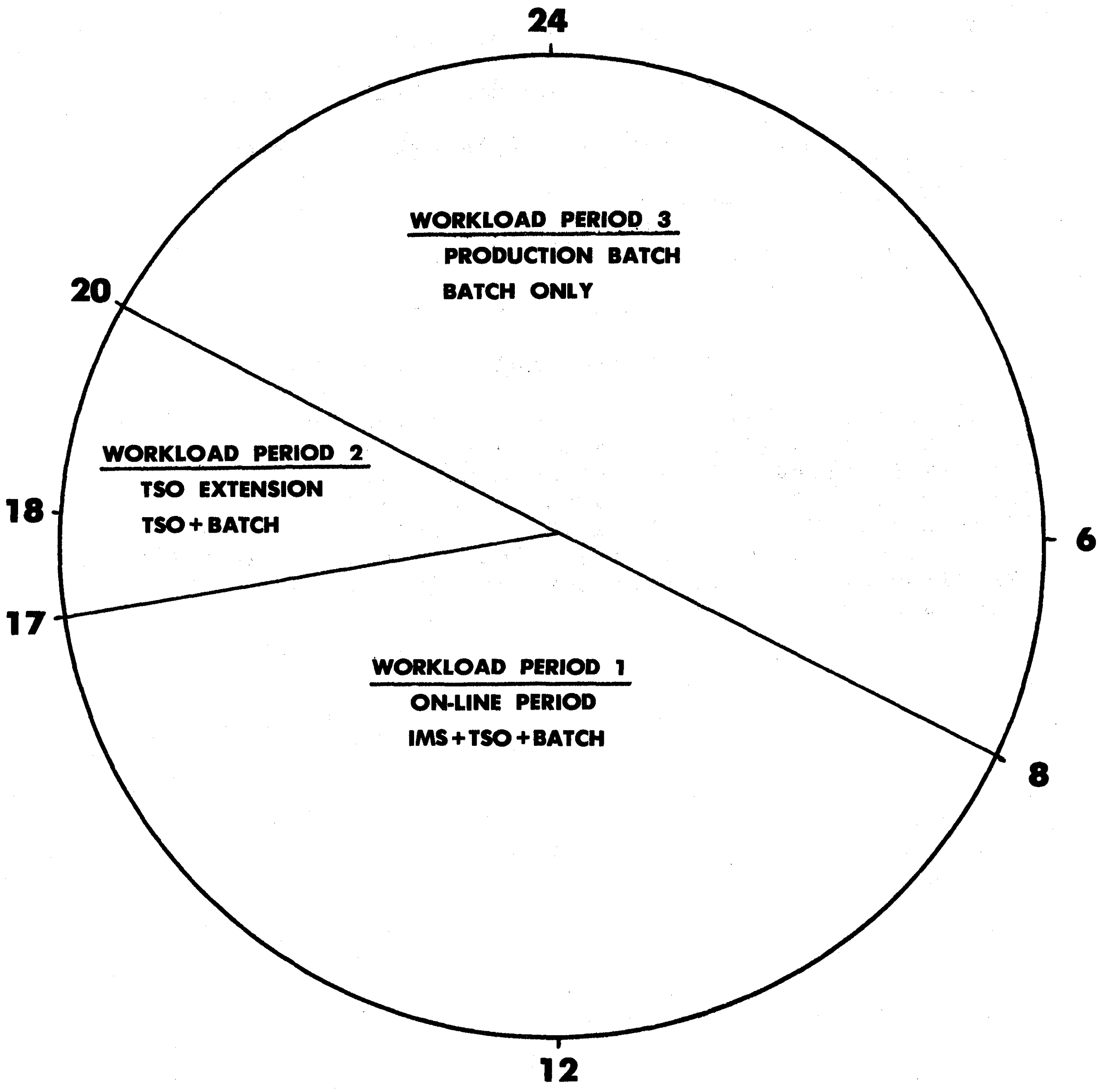
9. Estimate True CPU Time for paging.

10. Tabulate and Analyze Current Utilization.

    - Consolidating multiple CPU shop data
    - Subtotals of Business Categories
    - Bottom line calculations
    - Cross checks of your data
    - Decide what the analysis means.

In Figure 60, an example is given of a good technique for displaying installation production periods. For these production periods, Figure 61 is a tabulation of the business element and associated raw SMF data. The difference in Total CPU Time (455.5) and Raw SMF Time (264.0) should be noted. Figure 62 shows the True CPU Time (SMF Times after capture ratios are applied) for each business element.

WORKLOAD PERIOD: A RECURRING SPAN OF TIME DURING A DAY OR WEEK WHEN CERTAIN WORK MUST BE COMPLETED.



USAGE PRODUCTION PERIODS

FIGURE 60

	P1	P2	P3	
AVAIL HRS	189	60	252	
HRS OPER	185	57	248	
WAIT	6.7	12.2	15.7	
ACTIVE	178.3	44.8	232.3	455.4
BUSINESS UNITS				
ON-LINE (IMS/VS)	71.4	0	0	
TSO	18.1	8.5	0	
ENG	2.3	.7	50.0	
MFG	2.9	1.0	10.4	
PURCH	0	0	4.3	
MKTG	0	0	9.4	
PERS	0	2.6	11.5	
MISC	0	0	18.5	
DEVEL BATCH	0	0	16.7	
DEVEL TSO	2.0	.8	0	
ENHANCE BATCH	0	0	2.2	
ENHANCE TSO	0.4	0.2	0	
MAINT BATCH	1.0	0.8	5.4	
MAINT TSO	2.1	1.3	0	
SYS PROG	0.6	0.4	4.6	
DB/DC SUPP	0	2.4	8.8	
RERUNS, ETC.	0	0	8.7	
TOTAL	100.8	18.7	144.5	264.0

USAGE: CURRENT WORKLOAD WITH RAW SMF TIME

FIGURE 61

	P1	P2	P3	
AVAIL HRS	189	60	252	
HRS OPER	185	57	248	
WAIT	6.7	12.2	15.7	
ACTIVE	178.3	44.8	232.3	
			455.4	(TOTAL CPU TIME)

BUSINESS UNITS				CAPTURE RATIO
ON-LINE (IMS/VS)	102.0	0	0	.7
TSO	51.7	24.3	0	.35
ENG	2.9	.9	62.5	.8
MFG	4.8	1.7	17.3	.6
PURCH	0	0	7.2	.6
MKTG	0	0	15.6	.6
PERS	0	4.3	19.2	.6
MISC	0	0	30.8	.6
DEVEL BATCH	0	0	23.8	.45
DEVEL TSO	6.7	2.7	0	.3
ENHANCE BATCH	0	0	4.9	.45
ENHANCE TSO	1.4	.7	0	.3
MAINT BATCH	2.2	1.8	12.0	.45
MAINT TSO	7.0	4.3	0	.3
SYS PROG	1.2	.8	9.2	.5
DB/DC SUPP	0	4.8	17.6	.5
RERUNS, ETC.	0	0	14.5	.6
PAGING	2.1	1.0	0	
TOTAL	182.0	47.3	234.6	463.9 (TRUE CPU TIME)

USAGE: CURRENT WORKLOAD WITH TRUE CPU TIME

FIGURE - 62

In this example, summation of all components of True CPU Time (463.9) has exceeded the Total CPU Time (455.4.). This is approximately a 2 percent difference and well within the 5 percent guideline. With the data outlined in Figure 62, the next step is to move on to the forecasting phase.

2.3.5 Forecasting Future Workload

The purpose of doing the forecast steps in the USAGE Workshop is to:

- Explore potential/latent demand of end user groups
- Develop a 24 month forecast of demand by Business Element
- Determine when various capacity thresholds begin to be exceeded or when other warning signs show in the customer's forecast
- Evaluate various hardware/software configuration alternatives
- List management decision alternatives and lead time requirements for ordering, installation or implementation
- Make a capacity planning recommendation (e.g., a proposal for hardware, software, application, or Installation Management procedure).

The steps in the forecast process are as follows:

1. Gather interview data for forecasts.
2. Select forecast target dates (or cycle) and configurations.
3. Classify Business Elements into forecasting categories and then calculate future demand for each Business Element in terms of True CPU Time in hours/month based on one of the following classifications:
    a. Stable applications
    b. Partially installed batch
    c. Partially installed online
    d. New Batch
    e. New Online

f. Latent demand in:

- Development
- Personal computing
- Administration
- Inquiry
- DB/DC
- Outside Customer
- Data center support
- Operations Analysis
- Maintenance

4. Tabulate and Interpret the forecast

5. Configure to meet the forecast.

Gather Interview Data for Forecasts

This step of the study is the opportunity for you to quantify accurately the future requirements for DP within that location and to quantify new business for the DP Center.

As in the Analysis Steps, you will want to get enough information to double check your estimates. For this reason you should plan to see the people who understand the growth plans:

- DP Manager
- Project Office Manager
- User Interface/Accounts Manager
- Manager of Plans, Forecasts, & Venture Analysis

Outside DP you will be seeing:

- Major users
- Known growth areas
- Potential users

Ask questions that identify and quantify expected business growth or actual data reflecting the volumes, shape, and frequency of production arrival (jobs or transactions), CPU processing, and delivery (reports or messages). The contingencies of these new requirements and management levels of confidence should be noted with the answers to all questions.

The information needed will come from many sources. Some of the information sources as well as information required are listed below:

- Project request lists.
- Project Lists - work in test, or planned, and manpower levels.
- Organization and headcounts for Design, Development, Test and Maintenance.
- Batch Load estimates in terms of current experience.
- Survey Based or pathlength estimates from IMS Design Reviews.
- Transaction based estimates from end users.
- Data base calls per transaction.
- Degree of utilization of terminals (light or heavy load, many or few users).
- Multiple or single transactions.
- Data Entry, Query, Update, Calculations, Reporting, Message Switching, Program Development, Problem Solving.
- Business 2-5 yr. plan.
- Annual Report.

<u>Select Forecast Dates and Configurations</u>

Your may decide to forecast on a fixed cycle time to conform to the planning cycle (i.e., every 6 months). The other method would be to look for a combined load forecast at the cutover time of (1) significant new application(s) (i.e., + 3 months when IMS Order Entry is completed; + 18 months when the Customer Information file is in production; and + 24 months when the Integrated Financial Forecasting System is completed). In addition to picking the cycle of the

forecast periods you will want to decide whether you are doing a consolidated forecast or a forecast of individual CPU's (i.e., an isolated TP application). This depends on your operations or scheduling obligations and constraints.

Business Element Forecasting Steps

Each existing or new Business Element will be categorized into one of the following areas. What you know about it will have a large bearing on how you forecast the growth of an application.

1. Stable Applications

   Most likely this would be a batch job or an emulation job, probably for a non-changing function. The only growth would be that of changing I/O or transaction volumes. Your forecast should be conservative. At most it should change the number of True CPU hours per month at the same rate as the volumes change (adjusted for inflation in the case of dollar budgets).

2. Partially Installed Batch

   For applications that are being implemented, the calculations are similar to the stable applications. The rate of growth is just more rapid. The growth should be in direct proportion to the number of users or new volumes if the entire function is partially installed.

3. Partially Installed Online

   This is calculated in the same manner as partially installed batch:

$$\text{Projected load} = \frac{\text{Current Load}}{\text{Current Transactions}} \times \text{Forecast Transactions}$$

   - Where current and projected load are given in True CPU hours.

   - Where Current Transactions are a weighted sum of current transactions. The weighting factor should be based on relative complexity (i.e., path length to process a transaction).

   - Forecast Transactions are a similar weighted sum.

   Remember that most DB/DC applications do evening batch update, backup, catalog and data maintenance,

and management reporting. This takes as much time as the day time online load and should be included as the online use grows.

4. <u>New Applications Online</u>

This load projection may be figured one of three ways based on:

- Equivalence to a known current application
- Modelling
- Path length estimating from a DB/DC Design Review.

5. <u>Latent Demand</u>

If the data center has been operating at capacity for any extended period of time or the expansion of terminals has been limited by the budget or policy, then your installation may have latent demand. Latent demand is a quantity of work in an installation that is being surpressed due to insufficient resources.

To calculate the amount of work being surpressed is an "ART" but it can be estimated from some basic guidelines on user workloads generated when adequate resources are available. For example, guidelines are given below for five different online applications with adequate resources.

Application	Users/termminal	Transaction Vol/day	Complexity per Transaction
Data Entry	1	1000	Simple*
Inquiry	1	300	Medium**
Production DB/DC	1	500	Simple
Programmer	2.5-3	1200	Not characterized
Personal Computing	5	500	Not characterized

* Simple Transaction - less than 6 DB calls

** Medium Transaction - 6 to 12 DB calls

These are only guidelines and may not satisfactorily define your environment. If better data is available, use it. The guidelines are expressed in the approximate number of users per terminal for a given application and the estimated number of transactions generated by the terminal per day. For example, the guideline for the personal computing application is for 5 users assigned to one terminal, approximately 500 transactions will be generated in one day.

As a simple example of the computation of latent demand, the data entry application will be used. Assume that a data processing installation has been operating at capacity for some period of time and currently there are 5 terminals used for data entry (i.e., 700 simple transactions per day per terminal). Also, this example installation operates 22 days per month. The amount of supressed workload defined in CPU hours is based on a CPU which processes 1 simple transaction in 1 second. Note that the current transaction load is only 70 per cent the volume per terminal given in the guidelines above (i.e., 700 vs 1000). This implies the system may contain latent demand. To calculate the potential workload due to latent demand with adequate resources (i.e., new CPU) available in the future, the monthly workload, in CPU hours, must be calculated for the current transaction load and for the load defined in the guidelines. The difference in the two workload calculations is the potential workload due to latent demand. The calculations for the example outlined in this paragraph are given below:

Current Monthly Workload

$$= \frac{\text{22 days X 5 terminals X 700 Transactions X 1 second}}{\text{3600 Seconds/Hour}}$$

= 21.39 CPU Hours/Month

Projected Monthly Workload

$$= \frac{\text{22 Days X 5 Terminals X 1000 Transactions X 1 second}}{\text{3600 Seconds/Hour}}$$

= 30.56 CPU Hours/Month

Potential Workload Due to Latent Demand

= Projected Workload - Current Workload

= 30.56 - 21.39 = 9.17 CPU Hours/Month

Hence, there are 9.17 CPU Hours/Month of potential workload due to latent demand in the data entry application. This should be accounted for in the capacity planning effort. This is only one view of looking at latent demand and should be used only if it is appropriate for the environment to be studied.

Tabulating and Interpreting the Forecast

At this stage you should have all the data elements to complete a worksheet for each forecasting period.

To interpret the forecast the following questions should be asked:

When does the total requirement exceed the capacity threshold in the day time or night time?

- When does the TP requirement exceed the TP capacity threshold?
- What level of capacity will the current DASD support?
- What are the hardware growth options? What is their life span? Consider CPU, Memory, Channels and DASD.
- What are the software growth options (i.e., MVS for large memory systems)?
- What are the management options (Scheduling, Tuning, Availability, Rerun, Productivity) whose closer tracking and improvement will also improve capacity?
- What Service Level Requirements are needed to improve user satisfaction? Is there latent demand? What resources are needed to satisfy these requirements?
- What Project Management or Return on Investment objectives need to be improved and what resources are needed to satisfy these requirements?

These questions are the major problem, attention or decision areas. The answers to these questions are the foundation for the actions desired from the USAGE study such as:

- Order new hardware or software
- New procedures to better utilize the current configuration
- Order MVS or SNA to improve Host system capacity
- DP planning & measurement tracking using USAGE methodology.

Figure 63 is a tabulation of the forecasted CPU hours for the example. The future times chosen are 12 and 24 months, the total workload grows from 466 CPU hours for the full month to 697 hours.

BUSINESS UNITS	CURRENT	+12 MOS	+24 MOS
IMS	102	127	167
TSO	76	126	171
ENG	67	71	75
MFG	24	21	29
PURCH	7	7	8
MKTG	15	17	19
PERS	31	33	34
INV CTL	0	6	7
SALES ANAL	0	3	14
BOM	0	7	18
BATCH TEST	29	31	14
INT TEST	11	44	88
MAINT BATCH	16	16	16
MAINT INT	11	11	11
OTHER OPS	48	62	68
PAGING	6	0	0
TOTALS	466	621	697

USAGE: WORKLOAD PROJECTION IN TRUE CPU HOURS

FIGURE - 63

CONFIGURE TO MEET THE FORECAST

Now that the workload has been forecasted for the next 24 months, the next job is to select the CPU, memory and channels. It is assumed in all cases that the system will be running MVS.

CPU Capacity

The following are four considerations for CPU selection:

1. Total CPU Requirement - you should plan to install the CPU so that at installation time, the total load would be no more than 55 - 65% of capacity.

2. Any online system which requires an average utilization greater than 40% of the CPU during the production period, should be a candidate for running by itself during peak periods in order to accommodate peak loads and simplify operational aspects. The residual capacity during non-peak periods can be used for batch systems with non-critical deadlines.

3. In the case of multiple online systems, where each by itself is less than the 40%, you can add two or more to arrive at the 40% threshold value discussed in item 2 above. TSO and RJE with critical turnaround requirements should be considered as online.

4. For stability and availability, serious consideration should be given to multiple systems for physical isolation of testing and production. These systems should be integrated for operational reasons through shared spool, or JES3 and SNA-3 and/or IMS Multiple Systems Coupling.

From these guidelines, you will be able to select CPUs and the integration technique. This should now be reviewed for migration considerations to arrive at the installation date.

3.0 Summary

Although many factors of the capacity planning process are discussed in this technical bulletin, its primary purpose is to outline the capacity planning implementation process. The basic guidelines for implementation were developed from experiences with large DP installations over the last two years. These installations are involved in developing ongoing capacity planning efforts.

Implementation of the capacity planning process requires a major commitment on the part of upper management within the organization. The cooperation and coordination required between various departments (operations, systems, application development, users) to effectively develop a program will happen only if each department sees capacity planning as a major commitment. Upper management must see that the right talent (people), hardware and software is provided to implement the process.

To begin capacity planning in a DP installation, several key items (Figure 64) must be considered. The current process of "Capacity Planning" or by whatever name the process is known, must be outlined and understood. For example, what measurement tools are currently installed, what reports are being generated and who are the recipients, what functions are being performed with the reported data, what special procedures are being followed, etc. The purpose of this exercise is twofold. First, it can not be assumed that a capacity planning process structured seperate from existing activities and implemented at some later time can survive the pressures and politics of the organization. Secondly, there may be some concepts, guidelines, reports, people, etc., that will aid and possibly make the implementation process easier if known during the development process.

Assuming that the computer system is maintained in a "well tuned" state, CPU usage must be accounted for by subsystem (BATCH, TSO, IMS, CICS) and major applications within subsystem. Basically, capacity planning is initiated on the CPU and gradually includes the channels, I/O devices and network. The "USAGE" technique outlined in section 2.3 is an excellent way to initiate a capacity planning effort that starts with a focus on the CPU.

It has been pointed out that the key to understanding the capacity of a computer system is understanding the user service requirements (response and turnaround times). Therefore, a primary item in implementating a capacity planning effort is the establishment of critical application service objectives. Then, the service provided to these critical users must be tracked on a continuing basis.

1. OUTLINE IN DETAIL CURRENT CAPACITY PLANNING FUNCTIONS AND THE REPORTING PROCESSES.

2. ACCOUNT FOR AS MUCH CPU USAGE AS POSSIBLE (CPU HOURS) BY SUBSYSTEM (BATCH, TSO, IMS, ETC.) AS WELL AS BY MAJOR APPLICATION AREA WITHIN EACH SUBSYSTEM.

3. ESTABLISH SPECIFIC USER SERVICE OBJECTIVES (RESPONSE AND TURNAROUND TIMES).

4. PERFORM SOME TYPE OF INITIAL PERFORMANCE ANALYSIS AND FORECASTING.

5. DEFINE PARAMETERS FOR QUANTIFYING NATURAL LOAD GROWTH FOR EACH APPLICATION AREA IN TERMS DIRECTLY RELATED TO THE VALUES ESTABLISHED IN (2) ABOVE.

6. DEFINE PARAMETERS FOR QUANTIFYING NEW APPLICATIONS FOR EACH AREA IN TERMS DIRECTLY RELATED TO THE VALUES ESTABLISHED IN (2) ABOVE.

7. DEFINE A PROCEDURE FOR TRACKING PERFORMANCE ON A CONTINUING BASIS.

8. ESTABLISH REPORT FORMATS AND DEFINE THE REQUIRED TOOLS FOR IMPLEMENTING ITEMS 2, 3, 4, 5, AND 6 ABOVE.

SUMMARY: PLAN FOR INITIAL IMPLEMENTATION OF

CAPACITY PLANNING PROCESS

FIGURE - 64

One of the elements of the implementation process is a performance analysis technique. This technique may be detailed (discrete simulation, benchmarking) or less technical (linear projection, simple queueing analysis). To decide which approach to adopt, consideration should be given to the confidence one has in the data used to develop the model and the current understanding of the systems operation (workload characterization, resource consumption by application, availability (hardware, software, user perception). It appears, the best return on investment is accruing to those capacity planning efforts where performance analysis is kept as simple as possible and ongoing performance tracking is a major commitment.

In reference to items 5 and 6 of Figure 64, the primary point to be noted is the relationship of these items to item 2. Whatever units are being used to quantify CPU consumption, here it is CPU hours, these same units should be used to quantify natural workload growth of existing applications and new applications workloads. This means it is possible to track forecast workloads. In many instances, workloads are forecast in units not currently being collected or tracked by the installed measurement tools.

Tracking or ongoing performance monitoring is a key to good and improved capacity planning. It is this daily, weekly, and monthly performance data that begins to unravel some of the mysteries of data processing. Therefore, a basic requirement for implementing a capacity planning process is the development of a plan for ongoing performance monitoring.

In the development of a capacity planning effort, the reporting process must be given particualr attention. For example, the reported information is dictated by the measurement tools installed. Since it is recommended that measurement tools be kept to a minimum (as a maximum 3 or 4, depending on the number of subsystems, BATCH, TSO, IMS, CICS), certain data will not be available. Therefore, as pointed out in this bulletin, certain alternatives must be chosen. Data must be reported in the proper formats to be easily readable. Consideration should be given to the recipients of each report and the function they are expected to perform with the data reported. The reports and the reporting structure (frequency of reporting, recipients, formats, data, etc.) will be a primary factor in the viability of the capacity planning effort.

Through the involvement with many large DP installations over the last three years in a capacity planning role, it is clear that a viable capacity planning program is paramount for understanding and managing today's complex data processing environments.

4.0 References

1. Bronner, L., "Capacity Planning, An Introduction", IBM Washington Systems Center Technical Bulletin #GG22-9001, January, 1977.

2. Boyce, J., Belhumeur, R., Raimer, P., Shute, T., "IBM System/370, Tracking and Resolving Problems and Co-ordinating Changes", IBM Poughkeepsie Systems Center, Technical Bulletin #GG22-9000, June, 1975.

3. IBM Corporation Installation Management Brief, "Problem and Change Control at the State of Washington Data Processing Service Center", manual #GK20-1073, August, 1977.

4. Price Waterhouse & Co., "Management Controls for Data Processing", IBM Installation Management Manual, #GF20-0006, 1976.

5. Avizienis, Algirdas, "Approaches to Computer Reliability - Then and Now", AFIPS Conference Procedings, Vol. 45, PP. 401-411, 1976.

6. Short, R. A., "The Attainment of Reliable Digital Systems Through the Use of Redundancy - a Survey", IEEE Computer Group News, Vol. 2, No. 2, pp. 2-17, March, 1968.

7. Borgerson, B. R., "Dynamic Confirmation of System Integrity", AFIPS Conference Procedings, Vol. 41, Part-1, pp. 89-96, 1972.

8. Martin, J., "Design of Real-Time Computer Systems", Chapter-6, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1967.

9. Auerbach Information Management Series, "Data Communication Management", Auerbach Publishers Inc., Pennsauken, New Jersey, 1976.

10. Auerbach Information Management Series, "Data Base Management", Auerbach Publishers Inc., Pennsauken, New Jersey, 1976.

11. Martin, J., "Principles of Data-Base Management", Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1976.

12. Artis, Pat, "A Technique for Determining the Capacity of A Computer System", Bell Laboratories, Piscataway, New Jersey.

13. Little, J. D. C., "A Proof for the Queuing Formula: = ," The Operations Research Society of America, Vol. 9, No. 3, 1961.

14. Hippert, R.O., Palounek, L.R., Provetero, J. Skatrud, R.O., "Reliability, Availability, and Serviceability Design Considerations for the Supermarket and Retail Store Systems", IBM Systems Journal, Vol. 14, No. 1, pp. 81-95, 1975

5.0 Bibliography

Agajanian, A. H., "A Bibliography on System Performance Evaluation", Computer, PP. 63-74, November, 1975.

Agrawala, A. K., Mohr, J. M., Bryant, R. M., "An Approach to the Workload Characterization Problem", Computer, PP. 18-32, June, 1976.

Bard, Y., "Performance Analysis of Virtual Time-Sharing Systems", IBM Systems Journal, Vol. 14, No. 4, 1975.

Bard, Y., "Performance Criteria and Measurement of a Time-Sharing System", IBM Systems Journal, No. 3, PP. 193-216, 1971.

Bell, T. E., Boehm, B. W., Watson, R. A., "Framework and Initial Phases for Computer Performance Improvement", FJCC Proceedings, PP. 1141-1154, 1972.

Bell, T. E., Boehm, B. W., Watson, R. A., "How To Get Started on Performance Improvement", Computer Decisions, PP. 30-34, March, 1973.

Borovits, I., Ein-Dor, P., "Cost/Utilization: A Measure of System Performance", Communications of the ACM, Vol. 20, No. 3, March, 1977.

Boyce, J., Belhumeur, R., Raimer, P., Shute, T., "IBM System/370, Tracking and Resolving Problems and Co-Ordinating Changes", IBM Poughkeepsie Systems Center, Technical Bulletin #GG22-9000, June, 1975.

Bronner, LeeRoy, "Capacity Planning: An Introduction", IBM Washington Systems Center Technical Bulletin #GG22-9001, January, 1977.

Brotherton, D. E., "The Computer Capacity Curve - A Prerequisite for Computer Performance Evaluation and Improvement", Proceeding of the 2nd Annual ACM Sigmetrics Symposium, 1974.

Carlson, G., "Controlling Reruns", EDP Performance Review, Vol. 6, No. 4, April, 1978.

Computer Measurement Group (CMG), Proceedings of the CMG IX International Conference on Management and Evaluation of Computer Performance, San Francisco, California, December 5-8, 1978.

Computer Performance Evaluation Users Group (CPEUG), Proceeding of the Fourteenth Meeting, Boston, Massachusetts, October 24-27, 1978.

Dunlavey, R.F., "Workload Management", EDP Performance Review, Vol. 6, No. 4, May, 1978.

European Computing Conference, "Computer Performance Evaluation", Conference Proceedings, London, September, 1976.

Ferrari, D., "Workload Characterization and Selection in Computer Performance Measurement", Computer, PP. 18-24, July/August, 1972.

Gibson, C. F., Nolan, R. L., "Managing the Four States of EDP Growth", Harvard Business Review, PP. 76-88, January-February, 1974.

Heil, S.W., "One Approach to the Management of Computer Peformance Data", EDP Performance Review, Vol. 7, No. 1, Januay, 1979.

Hoffer, W. C., "An Automatic Scheduling System", Datamation, PP. 75-83, July, 1974.

Howard, P. C., Stevens, B. A., Carlson, G., "Evaluation and Comparison of Software Monitors", EDP Performance Review, Vol. 4, No. 2, February, 1976.

Howard, P.C., Stevens, B.A., Carlson, G., "Fourth Annual Survey of Performance-Related Software Packages", EDP Performance Review, Vol. 4, No. 12, December, 1976.

Howard, P.C., Stevens, B.A., Carlson, G., "A Case Study of Turnaround and Response Time Improvement", EDP Performance Review, Vol. 5, No. 2, February, 1977.

Howard, P.C., Stevens, B.A., Carlson, G., "Bibliography of 1976 Performance Literature", EDP Performance Review, Vol. 5, No. 3, March, 1977.

Howard, P. C., Stevens, B. A., Carlson, G., "How To Get Started in Performance Evaluation", EDP Performance Review, Vol. 5, No. 6, June, 1977.

Howard, P.C., Stevens, B.A., Carlson, G., "Performance Management Information Systems: State of the Art", EDP Performance Review, Vol. 5, No. 7, July, 1977.

Howard, P.C., Stevens, B.A., Carlson, G., "Bibliography of 1977 Performance Literature", EDP Performance Review, Vol. 6, No. 3, March, 1978.

Howard, P.C., Stevens, B.A., Carlson, G., Dunlavey, R., "A Data Processing Annual Report", EDP Performance Review, Vol. 6, No. 10, October, 1978.

Hunt, E., Diehr, G., Garnatz, D., "Who are the Users" An Analysis of Computer Use in A University Computer Center", Proceedings of the SJCC, 1971.

Hunter, J., "Bridling Data Processing Costs", Computer Decisions, PP. 44-54, June, 1977.

IBM Corporation, Installation Management Series, "Managing the Data Processing Organization", Manual #GE19-5208, October, 1976.

Jenkings, J. M., Howard, P. C., "Measuring System Capacity", EDP Performance Review, Vol. 5, No. 4, April, 1977.

King, G. M., "Graphic Throughput Analysis of Mixed Workloads", IBM Washington Systems Center Technical Bulletin #GG22-9017, February, 1978.

Kiviat, P. J., Morris, M. F., "Getting Started In Computer Performance Evaluation", Computer Measurement Group Transactions, No. 10, PP. 3.2-3.9, December, 1975.

Malick, Paul, "Systems Performance/Measurements - A Quantitative Base for Management of Computer Systems", Proceedings of the NCC, 1974.

McFarlan, F. W., "Problems in Planning the Information System", Harvard Business Review, PP. 75-89, March-April, 1971.

Morris, J. A., "Performance Constraints in Computer Systems", EDP Performance Review, Vol. 4, No. 8, August, 1976.

Noe, J. D., "Acquiring and Using a Hardware Monitor", Datamation, PP. 89-95, April, 1974.

Nutt, G. J., "Computer System Monitors", Computer, PP. 51-61, November, 1975.

Parker, G.C., Sequra, E. L., "How to Get a Better Forecast", Harvard Business Review, PP. 99-109, March-April, 1971.

Peeples, D.E., "Measure for Productivity", Datamation, PP. 222-230, May, 1978.

Performance of Computer Installations, Proceedings of the International Conference on the Performance of Computer Installations (ICPCI 78), Gardone Rivera, Lake Garda, Italy, June 22-23, 1978.

Phister, M., "Data Processing Technology and Economics", Santa Monica Publishing Company, Santa Monica, California, 1976.

Price Waterhouse & Co., "Management Controls for Data Processing", IBM Installation Management Manual, #GF20-0006, 1976.

Share European Association (SEAS), Proceedings Spring Technical Meeting, Berne, Switzerland, April 3-7, 1978.

Share Project, "Computer Measurement and Evaluation", Edited by Janet Wixson, Share Inc., Chicago, Ill., Vol. III, December, 1973 - March, 1975.

Stanley, W. I., Hertel, H. F., "Statistics Gathering and Simulation for the Apollo Real-Time Operating System", IBM Systems Journal, No. 2, PP. 83-102, 1968.

Stanley, W. I., "Measurement of System Operational Statistics", IBM Systems Journal, No. 4, PP. 299-308, 1969.

Stevens, B.A., "Audit and Control of Performance in Data Processing", EDP Performance Review, Vol. 6, No. 1, January, 1978.

Walston, C. E., Felix, C. P., "A Method of Programming Measurement and Estimation", IBM Systems Journal, Vol. 16, No. 1, 1977.

READER'S COMMENTS

Title: Capacity Planning Implementaton
Washington Systems Center
Technical Bulletin GG22-9015-00

Please state your occupation: ______________________________

Comments:

Please mail to: LeeRoy Bronner
IBM Corporation
18100 Frederick Pike
Gaithersburg, Md. 20760