1985

A methodology to initiate computer hardware capacity planning /

Margaret M. McKinnon
Lehigh University

A METHODOLOGY TO INITIATE COMPUTER HARDWARE CAPACITY PLANNING

BY

MARGARET M. MCKINNON

A THESIS
PRESENTED TO THE GRADUATE COMMITTEE
OF LEHIGH UNIVERSITY
IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

LEHIGH UNIVERSITY
BETHLEHEM, PA
MAY 1, 1985

PREFACE

This thesis is accepted and approved in partial fulfillment of the requirements for the degree of Master of Science.

May 1, 1985

Date

Professor in Charge

Chairman of Department

CONTENTS

LIST OF ILLUSTRATIONS

1.0 ABSTRACT

This thesis proposes a practical approach for initiating capacity planning for large computer systems. A complete capacity planning effort should include planning for all hardware, applications software, systems software, physical facilities, data, personnel and budgets. This thesis only addresses capacity planning for the processors.

A major concern of information systems management is how to accomplish the computer hardware capacity planning required to determine and cost-effectively sustain service levels with growing demands upon computing resources and the performance levels of the resources. There have been many articles written supporting the need to do this kind of planning and the consequences of not doing this planning well. Several authors have proposed general methodologies for performing capacity planning. Software vendors and information systems consulting firms have software packages available to support a capacity planning effort. However, a detailed methodology, setting forth a step by step approach for doing capacity planning is needed.

Because the author's experience has been with large IBM systems, all examples will refer to these systems.

2.0 INTRODUCTION

2.1 WHAT IS CAPACITY PLANNING?

Computer hardware capacity plannning is the set of functions and procedures which ensure that an organization has sufficient computer resources to cost effectively provide a desired amount of service to clients. Planning and controlling workloads, services, equipment and cost are the heart of capacity planning in a complex data center. The computer services provided should support business objectives and corporate goals. Capacity planning is a commitment to service, both for current and future needs.

The most important of all the capacity planning objectives is providing service to clients. This service is measured in quantitative levels of availability, response time, turnaround time and throughput. Timely identification and procurement of the system resources needed to satisfy client service levels is another objective of capacity planning. Another objective is to improve capacity projection techniques through on-going analysis of the projected versus the actual resource requirements. Controlling and projecting costs and guiding system tuning efforts are other capacity planning activities which are necessary to gather data for good projections.

2.2 WHY IS CAPACITY PLANNING NEEDED?

Today, many business organizations are investing a greater share of profits in information processing in order to keep pace with changing economies. Rapid technological changes in the computing industry have contributed to the proliferation of computers throughout industry. The business functions supported by computer systems are becoming more critical in nature as management's awareness of computer possibilities increases. Applications are being developed and implemented more rapidly due to increased programmer productivity and use of proprietary software packages. The greatly increased capacity of computer hardware, complexity of modern application systems and operating systems, the extensive development of on-line applications and remote job entry systems have increased the amount and complexity of work processed through a single computer.

User acceptance of the computer has never been higher. Information Centers, where non-data processing professionals have access to computing facilities and education, are becoming more popular and their service requirements are growing at explosive rates as users develop and implement their own applications. Personal computers are being used by executives in all businesses. Office automation and distributed data processing are becoming widespread, as well as, robotics and Computer Aided Design/Computer Aided Manufacturing (CAD/CAM) in manufacturing industries.

When one couples the tremendous growth in computer usage with the fact that lead time for new hardware is often measured in years and expenditures in millions of dollars, the need for effective capacity planning becomes vital to an organization. There was a time when capacity planning was viewed rather simply. When there was too much work for an existing computer configuration, the vendor was called to upgrade or replace it. It may not be prudent to depend upon vendors to do the capacity planning for an organization or to wait until an existing system can no longer process the workload. As the information systems budget becomes a larger share of the corporate budget, more attention is being given to information systems, their management and their costs. Management Information Systems

(MIS) managers must come to grips with capacity planning to provide acceptable, consistent and cost conscious service.

2.3 CONSEQUENCES OF NOT DOING CAPACITY PLANNING

The risks in not planning resource growth accurately can be monumental, especially if resource need is underestimated. The MIS department's credibility among its users may be lost and take years to rebuild. On-line productivity can suffer in many areas, affecting clerical personnel, programmers, engineers and all other on-line users. User dissatisfaction because of poor service can result in excessive overtime to complete a job. Customer service may be impaired. Orders may be lost or delayed, slowing down the cash flow. New business opportunities may have to be postponed or even lost. Government-mandated financial reporting deadlines may be missed, resulting in legal action. Through capacity planning; however, these risks may be avoided if sufficient computer resources are available to provide the services needed (Lipner, 1982).

3.0 THE CAPACITY PLANNING METHODOLOGY

The methodology described in this thesis is for initiating computer hardware capacity planning for processors. The steps should initially be performed rapidly to gain a somewhat accurate assessment of the overall capacity of existing computer systems. By going too deeply the first time, the capacity planners may get hung up on details of a particular technique and will take much longer to have useful results. Overall accuracy is limited by the least accurate component. Multiple passes through the methodology will permit comparison of forecasts with actual usage. Attention can then be focused on those areas where the forecasts were either too high or too low.

To establish standard procedures, document tasks as they are completed. Place routine portions of the functions into daily production to provide some continuity to the planning process.

The capacity planning methodology is basically a three step, iterative process: 1) account for current usage; 2) forecast future usage; and 3) match resource requirements to a future configuration. Each of these steps will be described in this chapter.

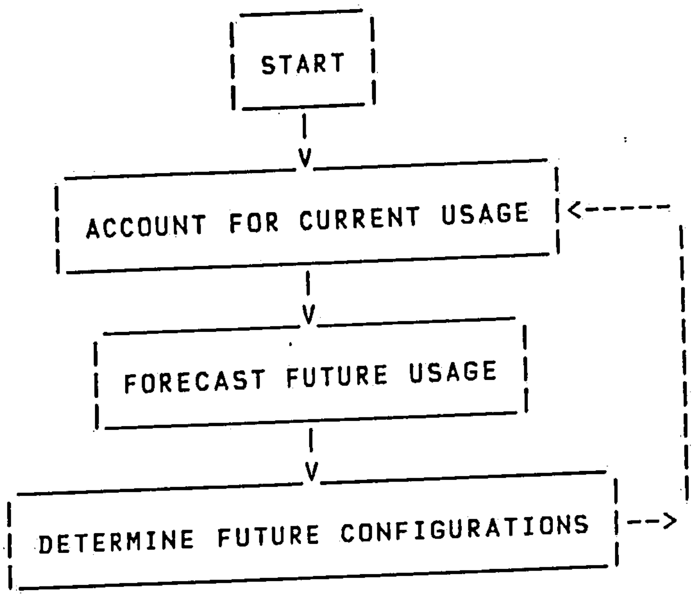

Figure 1. The Capacity Planning Methodology

3.1 ACCOUNT FOR CURRENT USAGE

3.1.1 UNDERSTAND EXISTING CAPACITY PLANNING EFFORTS

Just as each organization's computer systems grew from unique needs and circumstances, so will their capacity planning effort grow, especially if capacity planning builds upon planning and evaluation efforts that are already in place.

Existing performance evaluation and resource planning efforts within the MIS department should be understood and perhaps built upon. There may be some concepts, guidelines, reports, and people that may aid and possibly make implementation of capacity planning easier if known during the development process. Keep in mind that capacity planning should support the MIS strategic plan and feed the capital appropriation process. A capacity planning system structured separately from existing planning activities may not survive the pressures and politics of the organization.

Software packages and education available from proprietary software vendors and MIS consultants should be investigated by the capacity planners before developing their own elaborate systems for performing capacity planning. Performance evaluation packages may be of special interest due to the variety of statistical information which they can provide.

3.1.2 ORDERLY VIEW OF CURRENT SYSTEM

The capacity planners need to know the existing computer configurations. Specifically, this means developing an inventory of the hardware, software and business applications currently available to users. The inventory should provide a consolidated list of the capabilities within the organization and the demands being placed on the organization.

Suggested hardware resources to be included in the capacity plan include: processors, terminals, storage devices, control units, communication lines, power and cooling systems, to name a few. A complete capacity planning program will eventually include all hardware resources; however, the initiation of a capacity planning effort should be kept as simple as possible. Focusing on the processors will simplify this process. Therefore, this thesis only concentrates on planning for processors.

Suggested data elements for the hardware inventory include:

- Location (especially for multiple data centers)
- Model numbers
- System ID's
- Rated capacity
- Operational capacity
- Operating system (MVS/XA, VM, etc.)
- Major applications (IMS, BATCH, TSO, CICS, etc.)
- Configuration relationships
- Backup possibilities
- Memory
- Channels
- Cost

Relationships between the hardware, software and business applications should be documented. Figure 2 gives an example which identifies several processors and the major workloads processed on each. A chart similar to this is very useful to the capacity planners.

SYSTEM ID	SYSA	SYSB	SYSC
LOCATION	BETHLEHEM	CLEVELAND	NEW YORK
PROCESSOR	IBM 3081-K	IBM 3081-D	IBM 3033-AP
MEMORY	32 meg	24 meg	16 meg
CHANNELS	24	24	16
===========	===========	===========	============
DEVL.BATCH		X	
PROD.BATCH	X		
PROF.BATCH			X
DEVL. TSO		X	
PROF. TSO			X
DEVL. IMS		X	
PROD. IMS	X		
DEVL. CICS		X	
PROD. CICS	X		
CAI			X

Note: X = Service is available

Figure 2. Services Available by Processor

All capacity analysis should be performed using the most current system configuration. Discrepancies in measurement tool output may be directly traced to configuration changes. For example, if an IBM 3081-D is upgraded to an IBM 3081-K, the capacity planner should see an immediate decrease in the percentage of utilization on the upgraded system. The 3081-K has been rated 50% faster than the 3081-D when looking at the number of instructions it is capable of processing per second. Without knowing that the upgrade had taken place, the decrease in utilization could be misinterpreted as a decrease in demand for the processing resources.

To keep the capacity planning inventories timely, an interface should be developed between capacity planning and change management so that capacity planning will be notified whenever changes are made to the system configuration that affect capacity. Please see Figure 3 for a list of some hardware and software changes which the capacity planners should know about.

HARDWARE CHANGES
===============

- Install New Processor
- Upgrade Processor
- Add/Upgrade DASD
- Add Channels
- Add Devices to Paging Subsystem
- Increase Memory

SOFTWARE CHANGES
===============

- Install New Operating System
- Change Control Parameters
- Install New Version of Major Software (IMS, TSO, CICS, etc.)
- Change Virtual Storage Definition
- Move/Reschedule Workloads

Figure 3. Changes That Can Affect Capacity

When defining capacity available for use, take into consideration workload mix, user service objectives and resource utilizations. User service objectives are an important factor in understanding a system's capacity. If the user service objectives cannot be reached on a particular system, regardless of how much potential capacity is utilized, that system has reached its operational capacity. The capacity that is available for use while allowing all service levels to be achieved is known as the operational capacity.

Operational capacity is usually about 10% less than the potential capacity of a system. Exceptions are heavy interactive systems, such as dedicated TSO or IMS systems, which have operational capacities about 20% less than the potential capacity of the system. System bottlenecks, such as insufficient memory or insufficient channels, cause the operational capacity to be lower yet.

3.1.3 CHARACTERIZE CURRENT WORKLOAD

The capacity planners must have a clearly defined workload before attempting to measure or forecast anything about an existing system. There are several ways by which workload can be defined (Yen, 1983).

Workload can be described as batch, on-line or time-sharing. These descriptions imply the hardware requirements involved and service levels to be expected. For example, IMS is IBM's Information Management System which allows users to access a computer-maintained data base through remote terminals. It is an on-line, transaction oriented environment. TSO is IBM's Time Sharing Option which provides conversational time sharing from remote stations allowing a number of users to execute programs concurrently and to interact with the programs during execution (Vocabulary, 1981). A data center should have service objectives which include response time goals for transactions of varying difficulty executed in these services. For example, response time for trivial TSO transactions may average less than one second; medium TSO transactions may average less than three seconds; and long TSO transactions may average less than ten seconds. The capacity planners need to know the service objectives for all services offered in order to determine hardware resources required to support these services.

Workloads can also be described according to their major resource consumption. A workload can be described as:

- CPU bound or I/O bound
- Large or small memory occupancy
- I/O activities and secondary memory requirements

A workload which contains simulations would be considered CPU bound. Simulations contain many calculations which consume more of the processor resource than a job that simply prints a report without performing calculations. A payroll check printing program would be considered I/O bound since it could tie up a printer for quite some time while printing large quantities of checks.

Workload can also be described according to the timing requirements of the job.

- Scheduled or non-scheduled
- Priority or non-priority
- Long or short elapsed times
- Time dependent or independent jobs

Workload can be categorized by its general nature or purpose: application development, professional computing, commercial services, on-line production, production batch and research, to name a few. Each has activity generated by a user community. Research and professional computing, whether engineering, business or systems development, has an ever growing workload. To determine the impact of these workloads, requires bounding them by constraints such as terminal availability, historical trends, staffing or budgeting constraints.

Service level agreements (SLA's) are being used by some computer operations departments to quantitatively define service objectives offered to clients. The SLA is a binding contract between the client and the data center which may contain any of the elements listed in Figure 4. If SLA's are used by an organization, they may provide much of the information needed by the capacity planners to characterize the current workload. If SLA's are not used by an organization, the capacity planners will have to develop service objectives for the various workloads or find out what objectives are being used by the group responsible for performance evaluation and systems tuning. Once the service objectives are known, the capacity planners can begin to develop profiles of the resources required to provide a particular level of service for each workload.

- Identification of the contracting parties
- Description of work to be processed, including type, volume, mix and time of arrival
- Service levels to be provided, including response time, turnaround time, deadlines, accuracy, and availability (for both normal periods and contingencies)
- Performance reporting procedure specifying frequency and types of reports to be provided to users and data center management
- Descriptions of support services, such as problem determination or consulting
- Rates for services provided, if a chargeback system is used
- Penalties for non-compliance with agreement
- Provisions for modifying the agreement
- Expiration date

Figure 4. Service Level Agreement Elements (Witzel, 1983)

3.1.4 MEASURE PERFORMANCE & RESOURCE CONSUMPTION

The performance of the system must be analyzed to attempt to understand how the workloads and the system components interact. Service levels provided should be compared against service level commitments. Just as there were several ways to define workload, there are many ways to quantify workload (Yen, 1983).

The simplest and easiest workload measurement is the number count: Number of jobs or steps, number of transactions, number of users logged on, number of terminals used, etc. Hardware monitoring tools and system accounting packages can provide this data. Tracking these workload counts gives a good indication of increasing or decreasing workload volumes.

System accounting packages can also provide usage measurements for specific resources, such as memory or CPU seconds used. External resources used by a job, such as tape drives, disk drives, special print forms, etc. can also be reported. Some external measurements can also be obtained simply from job descriptions on a job card.

What to measure and what unit of measure to use are important and basic questions which the capacity planners must answer. Unfortunately, a unit of measure that can be used for processors is not as straight forward as using megabytes when measuring storage capacity. Some vendors have compared the capacities of their machines using a relative performance rating which assumes that the capacity of a particular machine is rated as one and the capacity of all other processors are compared to it. For example, the IBM 370/158-3 is commonly used as the base model. On this scale, the following processors are rated thus:

• IBM 370/158-3	1
• IBM 3033-U	5.6
• IBM 3081-D	10
• IBM 3081-K	14
• IBM 3084-Q	28

Some other vendors use the IBM 3033-U as a base, since large mainframes have much greater capacity than the IBM 370/158-3, as can be seen in the example above.

Some installations have developed computer resource units which are calculated from very complicated formulas combining financial information and several components of utilization information (such as memory occupancy, service units consumed, storage devices accessed, tape drives utilized, time of day, and transactions). This method is too complicated for an initial capacity planning effort because it requires years of historical information and a financial model to determine and maintain the coefficients used in the calculations.

An initial capacity management effort should rely upon percentage of CPU utilization and CPU minutes consumed or service units consumed by various workloads. Comparisons of the workload characteristics with the utilization information and response time information can be used to establish basic guidelines which can be used later in the methodology for forecasting processor demand.

To gain a better understanding of system performance and resource consumption, the capacity planners may want to perform studies similar to the following for each workload:

- Evaluate response time and its effect on user groups
- Correlate transaction rate with number of active users
- Track transaction rate versus time of day
- Evaluate queue time versus response time

The ratio of active users to total users can be used for calculating expected values and for making projections about the future number of active users and the resources they will need. This will be covered later in the methodology.

There are many hardware and software monitors available to aid the capacity planners in gathering system information. The capacity planners must become familiar with these tools and decide how and when they can be used most effectively. Amount of overhead involved in using these tools needs to be determined and considered during selection. The following are examples of the types of information available from system performance monitors (MVS, 1983):

- Processor activity
- Channel path activity
- Device activity
- Paging activity
- Workload/transaction activity
- Address space activity
- Page/swap data set activity
- Enqueue & reserve activity
- Domain activity
- Real storage/processor activity

System monitors can only partially account for the CPU time used by an application or workload. For this reason it has become important to distribute the remaining unaccounted CPU time among the active workloads. The capture ratio analysis is a method that enables us to break down the overall CPU utilization into the utilization attributable to each individual workload. It also enable us to estimate the total service time from the measured service time, reported in the RMF statistics.

The capture ratio is the ratio of measured service time to total service time. Regression analysis can be used when determining capture ratios. IBM suggests some Rule of Thumb capture ratios, but recommends that each installation verify these numbers or adjust them according to their own data (Armstrong, 1982).

• TSO	0.60
• TEST BATCH	0.80
• PRODUCTION BATCH	0.90
• IMS	0.80
• CICS	0.95

Capture ratios, regardless of how they are determined, should be tracked to develop consistent, reasonable numbers. Peak hour data should be used when determining capture ratios. Changes to the capture ratios will occur whenever the workload mix changes or if the system utilization changes by about 10%. Because measurement facilities are not currently available to validate the estimates derived from these techniques, a standard method has not been adopted by the industry. The concentration has been on consistency within an installation and repeatability of the estimated CPU times (Felix-Simpson, 1984).

Performance standards should be defined. Heavy users, system bottlenecks and any observed utilization trends should be identified and investigated. The capacity planners may want to recommend that some jobs which consume many resources be studied and perhaps tuned. Tuning individual jobs that have high resource usage can increase overall system performance. A realistically tuned system is basic to the capacity planning effort since tuning can affect the operational capacity of a system.

The capacity planners must understand the relationships between workload, resource utilization and service, as they relate to overall system capacity. Capacity is a function of the time of day, week or month. Scheduling of workload bears heavily upon understanding a system's capacity. The system utilization of on-line workloads fluctuates throughout the day causing peaks and valleys in utilization levels. The capacity planners should identify these peak periods and valleys.

Scheduling low priority batch work in the valleys can optimize the peaks. Indicators should be determined which can measure the peak and valley curve, such as number of concurrent active TSO users or the percentage of processor utilization.

The capacity planners need a schedule of peak utilization periods, identifying which workloads (batch, time-sharing, on-line data base/data communications systems) and critical applications run on which computer systems and when they are active. Expected values for utilizations can be calculated and compared to actual utilizations. The capacity planners should identify the job mix by shift for each computer system (i.e., first shift: Batch 5%, Production IMS 45%, Development IMS 30%, TSO 15%, etc.). Also, any priority system used for executing jobs should be defined.

The capacity planners must decide how the raw measurement data can be reduced, sorted and managed to provide input to the next step of the methodology which is forecasting future usage.

It may be useful to prepare a capacity review report on a monthly or quarterly basis, similar to that shown in Figure 5. A report of this nature identifies each processor and readily shows how busy each processor is and what its workload mix consists of. Shifts in workload mix are also apparent. The capacity planner may want to include information about capture ratios, average number of batch jobs per day, average TSO users per hour, or average IMS transactions per day which will also indicate shifts in the workload.

	1Q84	2Q84	3Q84
CONFIGURATION			
CPU1	3081-K32	3081-K32	3081-K48
CPU2	3081-D24	3081-K32	3081-K32
CPU3	3033-U16	3033-U16	3033-U16
OPERATIONAL CAPACITY (% BUSY)			
CPU1	80	80	80
CPU2	90	90	90
CPU3	75	75	75
AVG UTILIZATION (% BUSY) (DAY SHIFT ONLY)			
CPU1	72	75	68
CPU2	75	60	63
CPU3	65	68	71
WORKLOAD DISTRIBUTION (%)			
CPU1			
devl tso	60	61	65
test ims	15	17	16
test batch	25	22	19
CPU2			
live ims	75	71	73
prod batch	25	29	27
CPU3			
prof tso	75	77	76
prof batch	20	19	20
cai	5	4	4

Figure 5. General Capacity Review

3.1.5 SHOW CURRENT UTILIZATION

Charts should be plotted showing peak utilization over time, as in Figure 6. Additional lines can be added to the charts in each step of the methodology to show future demand and proposed configuration capacity.

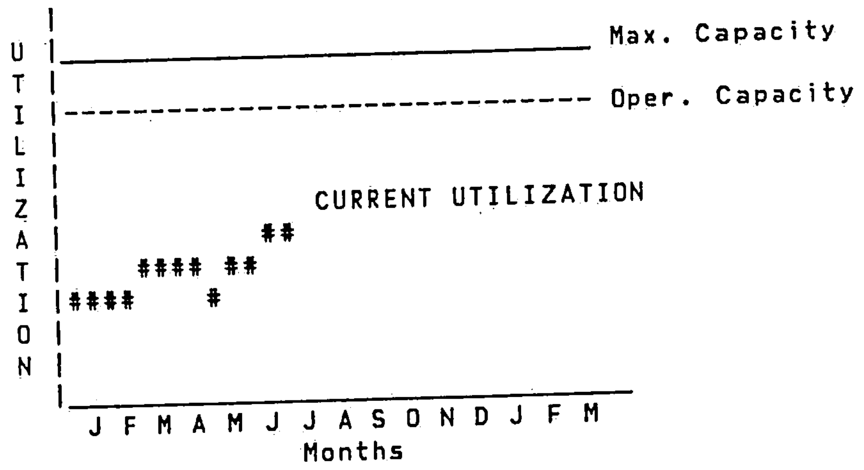

Figure 6. Graphical Presentation of Current Utilization

3.2 FORECAST FUTURE USAGE

3.2.1 PROJECT FUTURE WORKLOAD

The main sources of future workload information are the end-users. Users usually know what they will be doing in the next few months and are usually quite willing to help the data center plan. The capacity planners should ask questions that identify and quantify expected business growth. The users may be able to provide actual data reflecting anticipated volumes and frequencies of production work (jobs or transactions), CPU processing, data storage and output volumes (reports or messages).

User forecasts can be made in either computer terms (i.e. number of terminals, CPU hours, transactions/second) or in natural forecast units (i.e. items sold, number of new accounts, number of customers served). It is practical to get natural forecast units from users because those are units of work which the user understands. However, the capacity planners must then convert the natural forecasts into computer terms, then spread the demand over the available resources. This may not be easy to do unless historical records which correlate the natural forecasts with computer resource consumption are available.

Some growth areas that can be included in the planning process are:

- Existing systems growth
- New development impact
- Testing/development usage
- Changes to service levels
- Variances to planned figures
- Expected technology changes

One of the most difficult problems in capacity planning is predicting requirements for new applications. The most readily used means is comparing the proposed application to an existing application where certain performance parameters are known. The capacity planners try to interpolate between other workloads that are similar. Common sense and judgement must be exercised when doing this. If a workload will consume a significant percentage of the system resources, both an upper and lower bound should be estimated.

The capacity planners can assist the users in making better projections by providing them with feedback reports comparing what was estimated versus what was actually used so that the users may improve their predictions. Feedback should be provided on a regular basis, perhaps quarterly.

MIS management should be familiar with company goals and should provide direction to the capacity planners so that MIS can best support the corporate goals. Vendors, familiar with national growth trends and other data centers or computing organizations that have already experienced growth in planned areas, may be able to provide valuable insights about workload growth.

Contingency allowances should be included in all forecasts, especially when initiating a capacity planning effort. In a steady predictable business, the contingency allowance is probably a small number. In a business with highly varying business opportunity, a large contingency allowance may result in a better return on investment. It depends on how the business is to be run. When initiating a capacity planning effort, use at least 10% contingency on all workload estimates.

3.2.2 PREDICT EXPECTED PERFORMANCE

With credible forecasts of demand for service and new hardware/software technologies, the capacity planners can lay out a number of planning scenarios and attempt to determine whether system performance will be acceptable to clients. Performance criteria for major resources should take into consideration processing power, storage, communications requirements and user performance.

Several techniques for predicting performance of future processor configurations will be presented in this section. Figure 7 contains a diagram which shows how the various techniques compare when looking at complexity, cost, and time commitments required to use each technique successfully. All techniques should use peak hour data when making projections. If a configuration can provide desired service during a peak hour, there should be no problem in providing that same service during non-peak hours.

```
    /|                                                        |\
   / |________________________________________________________| \
  /                                                               \
 /  LO          COMPLEXITY, COST, MAN HOURS                  HI    \
 \                                                                 /
  \  ________________________________________________________     /
   \ |                                                        |  /
    \|                                                        | /

        |   LINEAR     |  ANALYTIC      |             |
 RULES| PROJECTION   |  MODELING      | DISCRETE    |  FULL
  OF  |--------------|----------------|SIMULATION|BENCHMARK
 THUMB|TIME   |REGRES|SINGLE|CENTRAL|             |
      |SERIES|       |SERVER|SERVER |             |
```

Figure 7. Spectrum of Performance Analysis Techniques (Bronner, 1979)

With simple approaches to capacity planning, a wealth of understanding and insight is possible concerning the complex operation of a computer installation. When the current environment is sufficiently understood and confidence is established in critical modeling parameters, then an installation may want to use queueing analysis or certain automated predictive tools which will enhance projections. However, it is possible to do meaningful analysis and forecasting without referencing a queueing relationship or discrete simulator (Bronner, 1977). If a system is not understood, complexity of a modeling technique cannot compensate. It is impossible to model what is not understood.

When initiating a capacity planning effort, the methods at the lower end of the spectrum in Figure 5 should be used. As experience, understanding, and confidence are gained with these methods, some of the more complex methods may be tried.

3.2.2.1 RULES OF THUMB

"Rules of Thumb" planning is the least complex and least expensive of the techniques. Rules of thumb are guidelines which relate utilization levels of components to overall system capacity. "Rules of Thumb" is a good entry level approach that can be followed by more sophisticated methods when analysis of an organization's resource performance yields better information about its resources.

Before using rules of thumb, an organization must define utilization limits based upon the level needed to satisfy committed service objectives. Utilization levels can be measured versus response time and correlations can be made. Utilization levels and known arrival rates can be used to calculate processing times. Then these times can be used to set utilization limits that will guarantee acceptable service levels.

The advantages of using "Rules of Thumb" are low cost and ease of use. Large amounts of capacity are not needed to predict capacity. Nor is a large commitment of staff required.

A disadvantage is that one cannot predict how growth will affect any individual component. Network and direct access storage device performance cannot be predicted well with this method.

3.2.2.2 LINEAR PROJECTIONS (TREND ANALYSIS)

This technique assumes that future resource requirements can be forecasted using a linear projection based upon past resource consumption. The principle advantages of this technique are that it doesn't require input from users and it is easily automated. This technique may not be appropriate for volume dependent or cyclic applications based upon time of year and for growth rates that are known to be non-linear. Also, there is no assurance that current trends will continue (Sarna, 1979). However, very credible capacity planning has been done by MIS organizations using simple guidelines, monitoring their systems on a continuing basis and using linear projections for future requirements (Bronner, 1979).

Another disadvantage of using this method is that historical data must be available. If an organization does not have historical data, this method cannot be used.

If another, more complex method is used to project expected performance, linear projection may be used to compare the projected results with past growth and performance rates. It is also interesting to compare calculated growth projections against national growth rates which can be provided by some vendors. For example, if your direct access storage growth rate for next year shows a 150% increase over what your data center is currently using and the national annual growth rate is 60%, you may want to investigate why your organization's projected growth is so much higher. You may also want to check your calculations before investigating anything.

3.2.2.3 ANALYTIC MODELING

Analytic modeling is a technique based upon the mathematics of queueing theory. The use of analytic models can help when projecting computer performance to point of saturation, estimating performance of data base systems, estimating transaction response for applications or forecasting processor and disk storage usage.

Boole and Babbage's Capacity Management Facility (CMF) includes an interactive analytic modeling system for predicting service levels and equipment requirements for a data center. A model of an existing computer system can be created from system data extracted from a monitor included in the package. The models can be used in planning scenarios to analyze the effects on performance and resource utilization of anticipated workload growth and proposed hardware configuration changes.

Queueing models have some advantages over other techniques. With no programming required, the input data obtainable from standard measurement reports (IBM's Systems Management Facility (SMF), IBM's Resource Measurement Facility (RMF), accounting data, hardware monitors, etc.) is used to describe workloads and configurations. Models can calculate performance information. No knowledge of queueing theory is required to use a model once it has been developed (Lipner, 1982). Any systems engineer, systems programmer or capacity planner could analyze various workload scenarios with a developed model.

The accuracy of predicted performance is as accurate as the input data, consisting of descriptions of configurations and workloads, both current and future. Also, queueing models process faster than simulations or benchmarks, which makes them suitable for interactive use.

Models can be rough or refined. Specific parts can be refined for in-depth analysis. The decision to look at some areas in greater depth is up to the capacity planners.

Analytic models can look at the performance capabilities of a wide variety of equipment and are not confined to any vendor. Performance of various alternatives can be reported from many viewpoints: throughputs, response time, utilizations, queue lengths, memory usage, etc., enabling the capacity planners to investigate potential performance problems and avoid them if possible.

3.2.2.4 DISCRETE SIMULATION

Simulation is a method of imitating how a given workload will perform on a particular computer system, using a physical representation of the system and keeping statistics on representations of jobs run through it (Stevens, 1981).

Performance characteristics of resources such as CPU's, terminals, communication-lines and many others are built into the simulation programs. This information must be kept up to date, which makes simulations difficult and expensive to maintain in dynamic environments.

A good example of a discrete simulator is SNAP/SHOT (System Network Analysis Program/Systems Host Overview Technique), a proprietary service of IBM. SNAP/SHOT uses simulation programs to investigate how jobs will consume computer resources.

Simulations are awkward to use for a wide range of planning scenarios and questions. They usually require data not readily available, especially for new applications. Persons requesting results from simulations must have some idea of the desired answer and use the technique to prove or disprove it. Simulations require specific sets of conditions and much set up time. Complexity of the simulation increases with each new variable, requiring more data and time for analysis.

3.2.2.5 BENCHMARKING

Benchmarking is the most expensive and complex of the techniques. Benchmarking involves testing with actual resources under simulated conditions. Success depends upon careful planning of the expected results and the simulated environment. A very good understanding of a system's operation is needed to make benchmarking accurate.

A drawback of benchmarking is that only programs and hardware already functioning can be studied. Benchmarking can't be used for new applications, or where future hardware isn't available. It is awkward to use for a wide range of planning scenarios and questions.

This method should be used sparingly compared to others and should not even be considered in an initial capacity planning effort. In general, only installations with a very good understanding of their MIS operation can adequately use the results of benchmarking as their only capacity planning tool (Bronner, 1979).

3.2.3 SHOW FUTURE UTILIZATION

A demand line can be added to the chart that was drawn in the last step of the methodology. Please see Figure 8. This line shows the anticipated growth in resources needed. It should look like an extension of the current utilization line. If the demand line crosses the operational capacity line, this indicates that additional resources will be needed at that time.

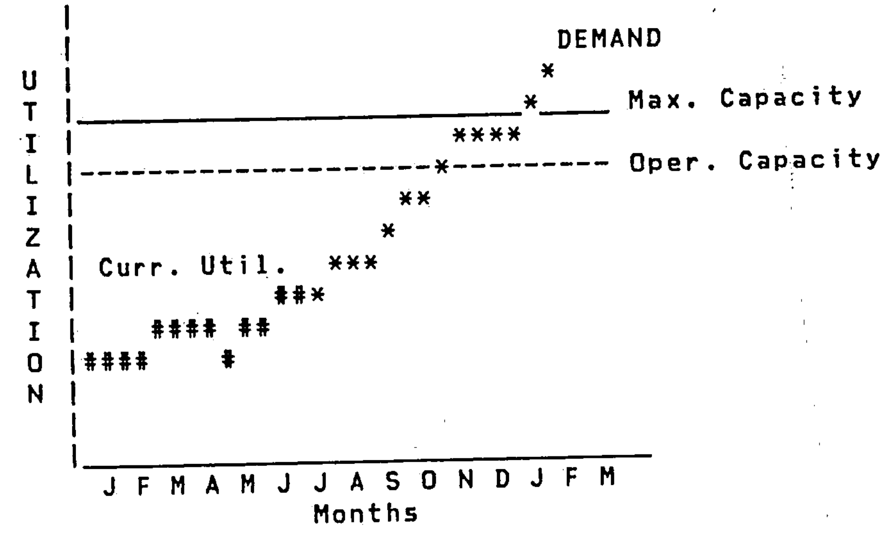

Figure 8. Graphical Presentation of Future Utilization

3.3 EVALUATE FUTURE CONFIGURATIONS

After the current environments and workloads have been defined and measured, and the future workloads have been projected and estimates of their resource needs have been made; the capacity planners must recommend future system configurations that will cost-effectively provide the levels of service and performance that have been proposed.

The same techniques and tools described above in "Predicting Expected Performance" can be used for evaluating future configurations. In this part of the methodology, the capacity planners try to measure proposed capacity, spot potential problem areas, model the problems, assess workload impacts and model alternative solutions. Analysis of costs versus benefits, availability of solutions and whether or not the solutions fit into the installation's plans must also occur.

The capacity planners base their evaluation on several criteria. Proposed configuration plans must provide specific levels of service and support the aims of the enterprise, as well as balancing cost and complexity with benefits gained. The capacity planners assign priorities to recommendations based upon timing, cost, whether the proposed configuration is mandatory by law, regulation or parent organization directive and whether the change will be beneficial to the parent organization or key user. Throughout the planning process, the capacity planners should remember that consistency of service is desired. Once service is established, it should be maintained at the same level.

The following are some guidelines for the capacity planners in this phase of the methodology.

- Do not plan for unannounced vendor products. Include only hardware and software currently available in the market place. Since capacity planning is an ongoing, iterative process, the plan can be updated whenever new products are announced which your organization can benefit from.

- Plan to achieve a few important goals.

- Avoid over-commitment of resources. Recommend solutions which will adequately provide services needed.

- Present several alternatives to management, with advantages and disadvantages of each alternative documented. A capacity plan that proposes only the "one best solution" and ignores other alternatives may be doomed to failure. Let management make the final decision about what will be funded.

Additional lines should be added to previous charts to show the proposed increase in operational capacity and the time at which the proposed increase is needed for each alternative, as in Figure 9.

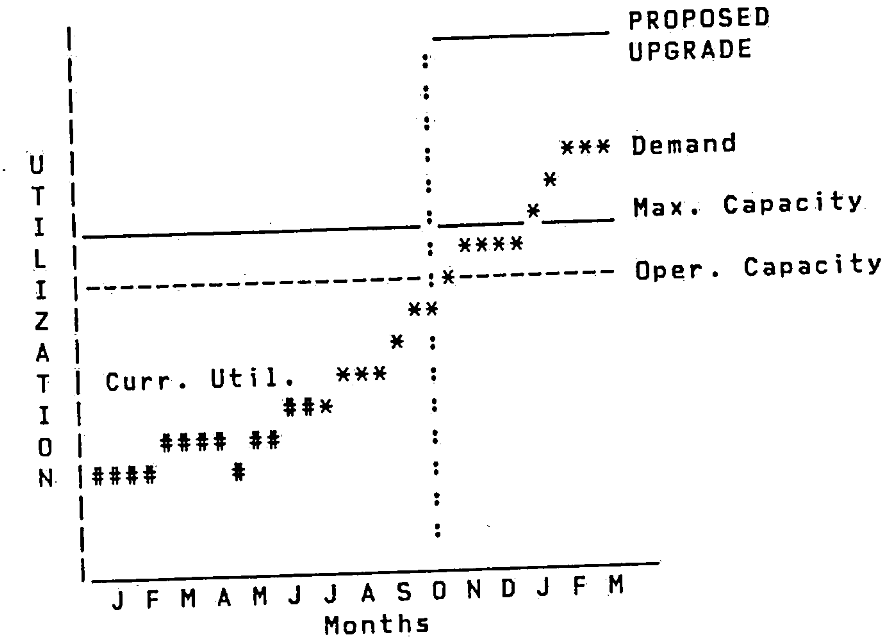

Figure 9. Graphical Presentation of Proposed Upgrade

4.0 THE CAPACITY PLAN

The actual, published Capacity Plan should be a sensible, achievable and coordinated plan for the total MIS function. It should contain configuration descriptions, trends and assumptions that have been identified and forecasted versus actual utilization validations. It should include charts showing the capacities of each systems environment, current utilization, future demand, and proposed increases. It is good to show at least one year's worth of historical data (if available) and at least eighteen month's worth of future projections. Advantages and disadvantages of each proposed alternative should be documented. The number of reports included in the plan should be kept to a minimum. A graphical format will facilitate easy review.

Requests for additional resources should be presented to management in terms of business problems arising from the lack of sufficient capacity to support the organization's business and goals (Finehirsh, 1983). References should be made to the strategic MIS plan, showing that the capacity plan complements and supports it.

The plan should track actual resource attainments over the planning period and make adjustments to forecasts and estimates for any variances that occurred. All variances should be analyzed, performance improvements should be reported and if necessary, replanning should occur. Adjust user forecasts based upon prior accuracy. If there is no follow up, it is likely that little of the plan will ever be completed.

To realize the maximum benefits of the tracking effort, the installation should do the following (Finehirsh, 1983):

- Define the measured variables most viable for tracking.

- Implement and maintain a data file or data base to store the tracked information.

- Define performance objectives to be used when comparing actual versus expected performance and workload activity.

Alternative methods for meeting the forecasted requirements should be presented to management. Some alternatives to be considered are: increase the number of shifts, change systems plans, change workload scheduling or workload mixes and offer decreased services. Many must be considered so that the most appropriate alternative may be chosen.

The capacity plan should be updated on a regular basis, perhaps quarterly, as well as whenever events require additional planning. Feedback reports should be provided on a regular basis to all areas affected by the plan, and it should be published for all interested managers and operating personnel to see.

Dissemination of the plan to all concerned managers and operating personnel becomes an interesting challenge, especially if it is to be a true working document. A simple hardcover notebook that can be easily updated is very useful. The plan can also be placed on-line to make it more accessible.

Communicating the results of a capacity planning effort to management can be the most difficult part of the entire endeavor. A purely technical presentation may overwhelm management and either compel them to accept a solution, or force them into a defensive posture, causing them to reject the solution. A combined technical and business presentation has the strength of showing management both sides of the issue in an intelligible manner. An interactive graphics presentation has been used very effectively for capacity presentations to management (Bell, 1983). All presentations should encourage questions, just in case more detail is wanted by management than was anticipated when preparing the presentation.

The Capacity Plan must be economical to avoid wasting dollars on unnecessary upgrades. It must be prepared without consuming undue resources. Most importantly, the plan must provide for adequate performance to clients at all times, even as workloads grow or shift.

5.0 STAFFING CONSIDERATIONS

The number of persons required to do capacity planning depends upon the size and complexity of the MIS department, the skill mix of the capacity planners and the risk of being wrong. If a shop has two to three processors, ten major applications, 10-15 user groups and a two-year upgrade cycle, probably two to three people should be involved in capacity planning, with at least one devoted full time (Allen, 1984). A larger shop would require more capacity planners, a smaller shop would require fewer.

Successful initiation of capacity planning requires selecting the right persons as well as positioning the function properly within the MIS organization. The capacity planners require analytic and problem solving skills, knowledge of processing hardware and software, performance measurement and evaluation knowledge, and tuning experience. A working knowledge of the operations, systems and application areas would be useful since the principle inputs to the capacity planning process come from these areas. Knowledge of the accounting and budgeting procedures used by the organization and some project management experience are also useful.

In addition to technical skills, the capacity planner needs a combination of business skills, communication skills and political skills. Communications skills, both written and verbal, are essential. The capacity planners must interface with the user community and inform all management levels about capacity requirements. The capacity planners must be able to make technical data understood in terms which personnel in the business areas will understand and be able to make a recommendation in the form of a business case for the organization (Finehirsh, 1983).

Finally, since initiating capacity planning, a new function in the organization, is similar to starting a new business, the capacity planner should also have the skills and attitudes of an entrepreneur (Allen, 1984). Figure 10 shows eight skills associated with entrepreneurs, which could be useful to the capacity planners.

INNER CONTROL	Exercising control over life situations rather than letting them be determined primarily by chance, fate or other people.
INNOVATION	Applying ideas, borrowed or original, to situations in which they have not been used.
DECISIONMAKING	Generating appropriate solutions to situations and carrying them out.
HUMAN RELATIONS	Acting in ways that reflect an understanding of one's own and others' needs, values, and goals.
PLANNING & GOAL SETTING	Designing and carrying out courses of action for the future.
REALITY PERCEPTION	Seeing people, things, or situations as they are rather than as distorted by imagination, emotions or faulty assumptions.
USING FEEDBACK	Collecting and using information for the purpose of confirming or changing decisions, perceptions plans or goals.
RISK TAKING	Taking informed action in uncertain situations.

Figure 10. Entrepreneurial Traits (Allen, 1984)

Many organizations place the capacity planning function under the responsibility of the manager of Technical Support where the capacity planners work closely with those responsible for performance evaluation and tuning. This seldom works well. Technical Support traditionally keeps the systems running and focuses on short term goals, instead of the much more global issues of capacity planning (Thorn, 1983). Ideally, the capacity planning function should be independent from daily operational problems.

To be most effective, the capacity planners should report to a level of management above the managers of Operations, Applications Development and Programming, and Technical Systems Support. A good location for the capacity planning function is as staff to the MIS director, as shown in Figure 11. This is especially true when there are multiple data centers under the MIS director, each with its own Technical Support, Operations and Application Development groups. This enables the capacity planners to have insight and access to information concerning the overall operation and future growth plans of the MIS installation, visibility within the organization and isolation from daily problems (Bronner, 1979).

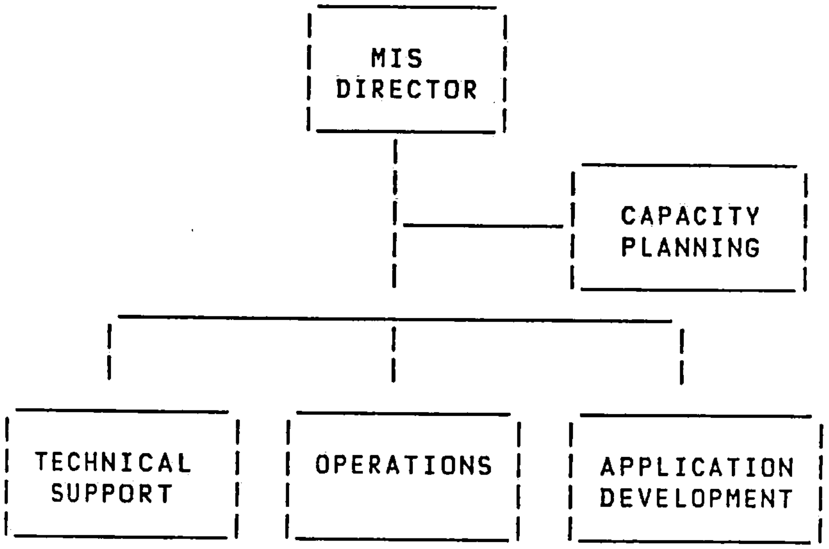

Figure 11. Suggested Organizational Structure

Functions within the MIS department with which hardware capacity planning must interface include (Bronner, 1979):

- Budget Planning - Converts individual plans into financial terms and identifies how funds will be acquired and allocated. The capacity planners must inform the budget planners of future acquisition plans so that budgetary constraints can be identified as soon as possible.

- Change Management - Concerned with the control and scheduling of changes on a regular basis to minimize disruptions to the computing environment. The capacity planners need to know when changes are made to the computer configurations and processing software so that their plans and models are current.

- Problem Management - Responsible for identifying hardware, software and operational problems in the systems and providing effective means to track these problems and ensure their resolution. The capacity planners should know when problems are related to resource shortages.

- Operations - Functions as a service organization, carrying out the instructions of various user departments; also schedules the user workload. The capacity planners interface with operations whenever rescheduling user workloads can improve service to users.

- Network Management - Responsible for design, testing and support of the telecommunications networks. The capacity planners must know the current network configuration when modeling computer systems.

- Performance Management - Responsible for defining performance objectives, requirements or specifications for operation and measuring how effectively the requirements are met. The capacity planners work with this group whenever characterizing workloads and determining service objectives. Capacity planning and this group use much of the same data.

- Data Base Management - Concerned with management of all automated data required by the organization. The capacity planners must plan for the amount of storage space required by the data bases.

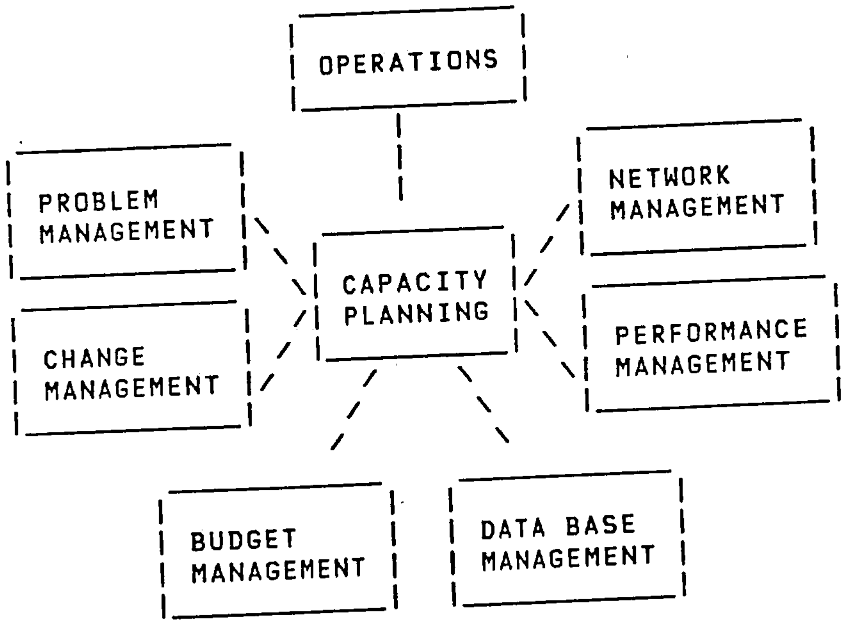

Figure 12. Capacity Planning Interfaces

Finally, to be successful, the capacity planning effort needs the blessing, attention and respect of top management, and management's active commitment to the establishment and implementation of the plan.

6.0 SUMMARY

The capacity planning methodology should be a combination of performance management ideas and measurement technology. Basically a three step process, the capacity plan involves: 1) accounting for current usage; 2) forecasting future usage; and 3) matching resource requirements to a future configuration.

Input from the client departments is vital. The clients initiate the workload. Providing service to clients, cost effectively, is a very important objective of capacity planning.

To account for current usage, the capacity planners must begin with a well-defined view of the current system. This can be achieved through the use of inventories of hardware, software, applications and people. The current workload must be defined and measured, as well as consumption of the computing resources.

The next step is to project the future workload. Input to this step comes from the clients, MIS and corporate management and vendors. Short and long term workload plans by application area must be developed. The capacity planners must identify the workload and determine the resources needed to satisfy the service requirements that the clients are projecting for the future. Growth and/or retraction of resource usage should also be noted in the workload plans. Future service levels anticipated must be defined. When initiating capacity planning, it is recommended that the planners use the Rules of Thumb and linear projection techniques, instead of attempting simulations or benchmarks.

Then the capacity planners must forecast performance of the projected workload on various computer configurations. There are many techniques for evaluating the ability of future configurations to provide the required services. Again, when initiating a capacity planning effort, the capacity planners are encouraged to use Rules of Thumb and linear projection more heavily than simulations.

It is not necessary to have an elaborate system to start benefitting from capacity planning. Development of a complete capacity planning function is an evolution that takes years to complete due to both the complexity of the system and the learning curve for the planners and the user community.

If management does not recognize the need for capacity planning, efforts to develop a plan will not be taken seriously by the organization and the plan will not be given the attention required to produce an effective methodology.

Capacity planning is an iterative process that requires cooperation with and continuous feedback both to and from all affected parties. Unless developing the plan is a cooperative effort amongst all areas affected by it, success will be doubtful.

The scope of the plan should be conservative. An initial capacity plan should not try to include all types of resources. The planners should concentrate on the processors and storage devices. Other resources can be included after expertise in capacity planning has been acquired. Processors are the most critical system resource to plan for since they are the most expensive and have the longest lead times for ordering. Further, the planning horizon should be limited such that it does not exceed two years for the initial plan.

A capacity planning methodology must be as independent as possible of any specific vendor's products. All products should be viewed as inputs to the capacity planning process.

Careful planning is important when implementing a capacity planning process so that disruptions to the planning function due to product changes or vendor changes can be minimized.

Capacity planning must be an on-going activity. The computer industry is a dynamic environment where things may not go exactly as planned, requiring constant monitoring. Once the plan is developed, it will require updating at regular intervals, perhaps quarterly, and as needed as a result of significant events. Routine monitoring of performance parameters by the capacity planners will be beneficial to the planning effort. Tracking performance parameters on a continuing basis allows one to gain certain system insights not always possible from a blitz type of data gathering effort.

Cost-effective service should always be emphasized, not state of the art hardware.

BIBLIOGRAPHY

Allen, L. E. and C. B. Kaplan, "The Capacity Planner as Entrepreneur", Journal of Capacity Management, Vol. 2, No. 3, 1984, pp. 191-206.

Armstrong, R. M., "Capacity Planning and Performance Management Methodology", IBM Washington Systems Center Technical Bulletin, GG22-9288-00, August, 1982.

Artis, H. Pat and Mario M. Morino, "Capacity Planning Experiences Based on MICS", Proceedings of SHARE 57, Chicago, IL, August 23-28, 1981.

Bell, Graham E., "What Happens If...? An Interactive Approach to Capacity Presentations", Proceedings of the Fifth Annual International Conference on Computer Capacity Management, New Orleans, LA, April 18-21, 1983.

Berg, Philip J., "Improving System Performance Measurement and Tuning", Software News, April, 1982.

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

Bronner, LeeRoy, "Capacity Planning Implementation", IBM Washington Systems Center Technical Bulletin, #GG22-9015, January, 1979.

Bronner, LeeRoy, "Overview of the Capacity Planning Process for Production Data Processing", IBM Systems Journal, Vol. 19, No. 1, 1980, pp. 4-27.

Brown, Bob, "Capacity Planning for a CMS Intensive Environment", Proceedings of SHARE 57, Chicago, IL, August 23-28, 1981.

Canning, Richard G., "Quantitative Methods for Capacity Planning", EDP Analyzer, Vol. 18, No. 7, July, 1980.

Dooley, Ann, "Need for Capacity Planning Stressed", Computerworld, March 2, 1981, p. 2.

Febish, George J., "Coping With the Explosive DP Growth", Proceedings of the Fifth Annual International Conference on

Computer Capacity Management, New Orleans, LA, April 18-21, 1983.

Feil, R. J. and B. A. Ketchledge, "MVS Performance Prediction Using Mechanically-Generated Queueing Models", Proceedings of the Computer Performance Evaluation Users Group 16th Meeting, Orlando, Florida, October 20-23, 1980, pp. 139-156.

Felix-Simpson, Sue, "Using CPU Measurement Data for Capacity Planning: How to Calculate Capture Ratios", Morino Associates, Inc., 1984.

Finehirsh, Sidney D., Thomas S. Moran, J. William Mullen, "Report of the Capacity Management and Planning Task Force", Proceedings of the Fifth Annual International Conference on Computer Capacity Management, New Orleans, LA, April 18-21, 1983.

Gill, Philip J., "Capacity Management Programs Bring Efficient DP Shop Service", Information Systems News, September 20, 1982, p. S5.

"How to Buy Enough Computers", Business Week, December 8, 1980.

Lipner, Leonard D., "Capacity Planning Simplified", Computer Decisions, September, 1979.

Lipner, Leonard D., "The Overloaded CPU", ICP Interface, Vol. 7, No. 1, Spring, 1982, pp. 30-36.

Long, Larry E., "Design and Strategy for Corporate Information Services: MIS Long-Range Planning", Prentice-Hall, Inc., 1982.

Major, J. B.,"Processor, I/O Path and DASD Configuration Capacity", IBM Systems Journal, Vol. 20, No. 1, 1981, p. 63.

"A Management System for the Information Business, Volume I, Management Overview", GE20-0662-1, IBM Corporation, 1981.

"A Management System for the Information Business, Volume IV, Managing the I/S Resource", Draft Version, IBM Corporation, 1981.

Morino Associates, Inc., "MVS Performance Management and Capacity Planning Survey Report", 1983.

Morino Associates, Inc., "TSO Capacity Planning", Draft Version, May 11, 1981.

"MVS/Extended Architecture Resource Measurement Facility (RMF) Reference & User's Guide", LC28-1138-1, IBM Corporation, August, 1983.

Sarna, David E. Y., "Forecasting Computer Resource Utilization Using Key Volume Indicators", Proceedings of the AFIPS 1979 National Conference, Vol. 48, New York, New York, June 4-7, 1979.

Shaw, Dennis, "Benchmarking and Its Alternatives", Proceedings of the Computer Performance Evaluation Users Group 18th Meeting, Washington, D.C., October 25-28, 1982.

Stevens, Barry, "Capacity Management and DP Planning: A Basic Approach", Computerworld, Vol. 15, No. 36, September 7, 1981, p. 58.

Strauss, Melvin J., "Computer Capacity - A Production Control Approach", Van Nostrand Reinhold Company, New York, 1981.

Thorn, Norman R., "The Steps Toward Getting Started in Capacity Management", Proceedings of the Fifth Annual International Conference on Computer Capacity Management, New Orleans, LA, April 18-21, 1983.

Vincent, David R., "Performance Standards for Capacity Planning", Computerworld Extra, Vol.15, No. 35a, September 1, 1981, pp. 39-48.

"Vocabulary for Data Processing, Telecommunications, and Office Systems", GC20-1699-6, IBM Corporation, July, 1981.

Witzel, Christine N., "Service-Level Agreements: A Management Tool for Technical Staff", Journal of Capacity Management, Vol. 1, No. 4, 1983, p. 344.

Yen, Elizabeth H., "The Importance of Workload Definition in Capacity Planning and Performance Evaluation", Proceedings of the Fifth Annual International Conference on Computer Capacity Management, New Orleans, LA, April 18-21, 1983.

VITA

Margaret Mary McKinnon was born to Mary and Thomas Hastings on September 15, 1955 in Cleveland, Ohio. She received her Bachelor of Science degree in Computer Engineering from Case Western Reserve University, Cleveland, Ohio in 1977 and accepted a management training position at Bethlehem Steel Corporation, Bethlehem, Pennsylvania. She received her Master of Science degree in Industrial Engineering at Lehigh University, Bethlehem, Pennsylvania, in 1985.