SERVICE LEVEL CAPACITY PLANNING: A TOUGH NUT TO CRACK
Carlton L. Arnold, General Telephone and Electronics

Abstract

Today's management is demanding that data processing decisions be based on the quality of service delivered to the customer. This demand has impacted the area of capacity planning to the point where many companies are recognizing that hardware plans must satisfy specified customer service levels. As a result, those companies are seeking a "capacity planning system" that is service level driven; however, it is the contention of this author that such a system does not exist today. Why doesn't one exist? This paper will identify some of the problems and constraints that have prevented the development of a complete capacity planning system that is service level driven.

INTRODUCTION

In general, capacity planning systems use one of two techniques to forecast hardware needs:

1. comparing component utilization levels to hardware guidelines, or
2. predicting customer service levels given a specified workload and hardware configuration.

The former is widely used and accepted by many companies as a technique for forecasting hardware needs. This technique considers customer service levels to be functionally related to component utilization levels. For example, when channel utilization reaches 25%, degradation in customer service may be anticipated. This technique assumes a particular customer service level when a component utilization level is established; therefore it is difficult to realize hardware needs if the service level is allowed to change.

The second technique is more desirable since it can provide the manager the opportunity to cost out (and, hence, budget) the hardware required to achieve certain customer service levels (e.g., terminal response times). This is possible because expected service levels are actually computed, and, therefore, can be compared to desired service levels. However, this latter technique has been difficult to incorporate into a company's capacity planning system because there does not exist a complete system which will forecast hardware needs based on service levels. There are companies that have purchased software packages which assist in predicting response times; however, these packages usually require a significant amount of user effort to provide input data and to interpret the results of the package. In either technique, the burden of data collection, data analysis, modeling strategy, and plan development remains with the user of the package.

More significant constraints are encountered when a company must develop a corporate hardware plan in a multiple data center environment. The hardware plans for each data center must be reasonably consistent to ensure the validity of the corporate plan. In order to address these problems, a complete capacity planning system which uses customer service levels as criteria for determining hardware upgrades is required. This paper will describe some of the problems and constraints that have precluded the development of such a system. This is done by examining four basic functions of any capacity planning system:

1) collect data
2) analyze data
3) predict future needs
4) report results

At GTE, we have developed a complete capacity planning system which satisfies these four functions. With the exception of collecting future workload requirements the system is automated to reduce the burden of effort required of the data center analyst. Figure 1 summarizes the four functions above as they appear in the GTE Capacity Planning System.

The system collects data from SMF for current workload and from computer users for future workloads. The users complete standard forms to size the particular application system that supports them. This sizing information is related to business activity so that the user can better understand the relative impact of a system. The implementation schedule for all new systems is also collected so that impact dates can be assigned to the anticipated workload. This data is now analyzed to determine the computer resource consumption and the anticipated growth rates from all the various workloads.

The analyzed data is then accumulated to define the total computer resource requirements for specified time periods (e.g., months or quarters). The capacity of existing hardware is calculated and, through graphing techniques, is compared to the load requirements. An example of this graph is shown in Figure 2. Using planning guidelines, upgrade points are identified and the capacity of candidate hardware is compared to the load lines. From the comparisons, the hardware necessary to satisfy the load requirements is determined.
COLLECT DATA

ANALYZE DATA

PREDICT FUTURE NEEDS

REPORT RESULTS

GTE CAPACITY PLANNING SYSTEM

SMF

CURRENT WORKLOAD

RESOURCE CONSUMPTION GROWTH RATES

DETERMINE LOAD REQUIREMENTS

DETERMINE CAPACITY

LOAD VERSUS CAPACITY

FUTURE WORKLOAD

IMPACT DATES
RESOURCE CONSUMPTION GROWTH RATES

USER

Figure 1
The most significant aspect of the system is that it provides consistent hardware plans, regardless of data center, which permits the development of a valid and realistic corporate hardware plan. However, the system does use hardware guidelines to determine future needs. Even though these guidelines have proven valid over several years, our goal is to make hardware upgrades directly dependent upon satisfying the customer's service level needs. The objective of this paper is to clearly define what is required to attain that goal.

I. FUNCTION ONE: COLLECT DATA

Collecting data involves identifying, retrieving, and storing capacity planning data necessary for the analysis and projection functions. Primarily, this data consists of historical computer usage data but will also include data which describes hardware, system software, and future workload requirements. The specific data items collected and the frequency of collection is dependent upon the methods used in the analysis and projection functions. Data collection problems and constraints which prevent the development of a complete capacity planning system that is service level driven are the following.

A. The tracking of service level data is difficult, if not impossible. That is, the technology required to measure the service delivered to a customer is lacking. This assumes, of course, that service levels have been defined in meaningful terms.

B. The number and variety of measurement tools have led to a confusing, conflicting and contradicting use of data provided by the tools. The simple measure of "CPU time" can be provided by numerous measurement tools, each with its own unique definition of CPU time.
C. The collection of data which describes future workloads (i.e., new application systems) is an arduous task for the capacity analyst. There exists no standard method for sizing future applications which is compatible with predicting overall hardware requirements. A company will employ simulation to size future workloads and then naively hope that the results can be made part of the hardware plan. The results of sizing efforts involving future workloads must be interfaced to the capacity planning system, preferably with an automated method.

The opportunity to resolve some of the problems is being provided by statistical packages such as the Statistical Analysis System (SAS). This is being done in several companies through the use of a performance measurement data base maintained with SAS code. However, the full impact of a tool such as SAS has not been realized in this function. Such a tool may not only satisfy this function but may also provide an interface to the analysis and projection functions. This could greatly enhance the consistency in collecting data in order to compare planned hardware needs to actual needs.

II. FUNCTION TWO: ANALYZE DATA

The objectives in analyzing data are:

1) to develop a workload profile which adequately characterizes the the various processing workloads,
2) to determine the most representative processing environments on which to base future projections, and
3) to provide the statistical inputs required in the projection function.

The analysis to achieve these objectives occurs either periodically to establish certain static conditions or continuously to remain sensitive to changes which may have significant impact. In other words, the analysis function attempts to provide statistical significance to events that have occurred or will occur in a data processing environment.

The current problems and constraints within this function are the following:

A. There does not exist a standard automated method to characterize workloads. For example, the methods currently used to develop workload characterizations are usually dependent upon the input requirements of a particular prediction tool or technique. Or, in some cases, workload characteristics methods are developed without a prediction technique in mind; thus, the characterization effort may be informative and interesting but fruitless.

B. Probably the greatest constraint within this function is the complex analysis required when determining the representativeness of workloads. The capacity analyst has the responsibility of identifying representative workloads on which to base future hardware needs. This is a non-trivial task when one considers that computer workloads are time dependent. An example of this is the common two hump distribution of real-time workload occurring on a prime shift. The most difficult question facing the capacity analyst is, "What is a representative workload?" A complete capacity planning question facing the capacity analyst question.

As in the first function, a tool such as SAS may provide some answers to these problems. The ability to statistically analyze large amounts of data in a straightforward manner may provide workload characterization and representation that has not been possible in the past. The real gain may be the development or refinement of standard statistical techniques (e.g. cluster analysis) that will automatically provide workload characterization and representative workload profiles for any given workload. Once this can be done, a technique can be developed which will provide an interface between this function and the next.
III. FUNCTION THREE: PREDICT FUTURE NEEDS

This function uses the data collected and analyzed to predict the computer hardware required to satisfy specified criteria. As mentioned earlier, this criteria may consist of component utilization levels or customer service levels. The actual method used to predict must also provide output that will result in consistent hardware plans when interpreted by capacity analysts at different data centers within a company. At the same time, the method must provide a projection capability at a corporate level while maintaining sensitivity to customer service levels. For these reasons, predicting future needs is the most complex function. Furthermore, it is in this function that most of the research and development on new capacity planning tools is being concentrated.

To put into perspective the basic problem within this function, it is necessary to review the use of component utilization levels in determining hardware upgrades. In particular, the GTE Capacity Planning System will be briefly described.

As mentioned earlier, our capacity planning system correlates component utilization with customer service. This gives an indication of potential degradation of customer service if the conditions are similar to those when the correlation was made. That is, a guideline is established for component utilization which is related to customer service under certain conditions. To develop guidelines for all possible conditions in a computer environment is impossible. For example, the variables included in establishing a guideline under the current GTE system include the following:

- **Hardware component** (CPU, Tape, DASD, etc.)
- **Model type of each component** (370/168, 3033, AP, etc.)
- **System Control Program** (DOS, OS, MVS, etc.)
- **Subsystem control programs** (CICS and IMS)
- **Workload Mix** (on-line and batch application systems)
- **Shift** (day, evening, or night)
- **Type of processing** (TSO interactive, batch, etc.)
- **Amount of workload** (CPU hours, tape drive hours, etc.)

In addition, all of these variables influence one another; i.e., a change in one variable will most likely change other variables. Depending upon the magnitude of the change, these other variables may affect one another and cause additional changes. Because of the number of variables, the possible variance in their values, and their interdependence, the forecasting of customer service levels presents an extremely complex problem. It is here that guidelines resolve some of the complexity while introducing uncertain results if all possible combinations of variables are not considered. A more direct method would be to calculate the expected customer service for any set of variables through the use of an algorithm. The interdependence or relationship of all the variables would be included within the algorithm so that, as values change, their influence upon one another could be predicted. It is the development and validation of this algorithm into a model which poses a significant constraint in obtaining a capacity planning system driven by customer service levels.

Several companies have developed or bought models which forecast customer service levels. However, the problems within this function are caused by the inadequacy of the other three functions. That is, the analyst must determine the input parameters (Functions One and Two) to a model and must be able to evaluate the model's results (Function Four) to determine the need for hardware.

In other words, the method used to predict future needs is actually a constraining factor because of its non-automated demands placed on the analyst. These demands are the following:

1. collection and analysis of data,
2. development of workload characterizations,
3. determination of representative workloads, and
4. consistent evaluation of results.
All of these efforts must be compatible to the prediction method.

Therefore, the basic need in this function is the development of standard automated procedures to satisfy the other three functions in a manner which would be compatible with the prediction method. This is the underlying constraint of developing a complete capacity planning system which is service level driven.

This compatibility may only be attainable through modifications of current tools and/or the introduction of new prediction techniques. With a tool such as SAS, it may be possible to provide interfaces between each function that will improve compatibility by introducing statistical significance to the inputs and outputs of a prediction tool. However, current attention devoted only to this function must be expanded to consider all four functions as one integrated system.

IV. FUNCTION FOUR: REPORT RESULTS

This function of the capacity planning process serves to communicate the results obtained in the other three functions. The major output of this function is the hardware plan which reflects a schedule of hardware upgrades. Similar to Functions One and Two, this function must currently be accomplished through manual means by the analyst. Also, as mentioned earlier, hardware plans from various data centers in a multiple data center environment must be consistent. The manual effort required and the possibility of inconsistency highlights the problems within this function.

SUMMARY

One of the most significant problems which constrains the development of a complete capacity planning system is that most, if not all, attention is given to one function -- the development and use of a technique to predict future needs. Because of this, an algorithm has not been developed to be used in a model which is compatible to all four functions of a capacity planning system. Emphasis must be placed on the total system (i.e., Functions 1 - 4) with each function mutually compatible. At the same time, the criteria for hardware upgrades must be based upon satisfying customer service levels.

At GTE we are using an automated corporate capacity planning system which emphasizes all four functions. Currently, we are involved in projects which attempt to solve some of these problems. We are committed to the task of making customer service levels an integral part of capacity planning and we are optimistic that industry, as a whole, is making a similar commitment.
SAS TRAINING AND SUPPORT

Chairman:
Sally Carson
The RAND Corporation