CAPACITY MANAGEMENT
A METHODOLOGY
Phase 2: Service Level Agreements

Adrian Heald
Capacity Reporting Services

Primitive Capacity Planning measures the "fullness" of each IT processing component and recommends action before they become "full". This ensures that there will be sufficient resource to carry out business functions but can lead to upgrading processing components that may not need upgrading or are not the most cost effective to upgrade. More sophisticated methodologies model user response and recommend action before these are impacted. This paper presents phase 2 of a 4 phase methodology that provides an orderly progression from "no capacity planning" through primitive "fullness" capacity plans to sophisticated response based capacity planning.

INTRODUCTION

Phase 1 of this methodology establishes data capture and collection systems. Data is gathered from both the business and technical arenas and stored in performance databases.

Using simple statistical analysis the association between resource utilisation and the business needs driving the utilisation is determined.

This association can be applied to projected business requirements thus establishing a "primitive" understanding of future demands on the IT resource. However this will cause recommendations for resource upgrades based solely on the "fullness" of the processing resource. No consideration has been given to the response required by the business units.

This paper presents phase 2 of this methodology. It establishes documents that detail the level of service, (response and throughput etc.) required by the business units to effectively carry out their function. This enables capacity management based upon response as well as providing a consistent level of service to the user population.

Phase 3 models the result of phase 1 to determine the impact on the service level agreements established in phase 2. Thus establishing a clear understanding of the impact of changing resource utilisation on the real work carried out by the business units.

Phase 4 establishes resource cost attribution which associates the cost of future resource acquisitions with the business units consuming them. This allows informed decisions regarding the worth of these acquisitions based upon the importance of the business function.

See CMGA'94 proceedings for "Capacity Management A Methodology Phase 1: Business Transactions and IT resource Requirements".

Figure 1 shows an overview of Phase1 and 2 of this capacity management methodology

Error! Not a valid link.

Figure 1 Capacity management methodology

A solid commitment from senior management is required before this phase can be implemented. The service level agreement must be considered a contract between the end user group and the IT area of the organization. If the level of service delivered does not meet that stated in the agreement then IT management must commit resources (both human and financial) to ensuring it does.

End user groups will soon lose interest in the process if the service level agreements are not honored by the IT units. This will result in capacity management that can only be based on the "fullness" approach which can results in unnecessary hardware acquisitions.

PHASE 2

IT units within an organization provide information services to the business units of the organization. A clear understanding of the level of service required by the various business units establishes a feedback mechanism necessary for capacity management. If there is little understanding of the response required by the business units how can the validity of a resource upgrade of acquisition be determined.

Running a CPU at 100% is valid if the end users are satisfied with the level of service delivered.

If the required level of service is not being met then the business needs of the organization are not being satisfied and IT is not performing the functions required of it. The required levels of service are detailed in service level agreements.

A service level agreement is a pact made between the management of IT and the actual user of the IT resources, the business units. The agreement clearly defines the level of service the user will receive in terms of response or turn around time. There are many other factors that should be included in service level agreements (such as help desk schedules, availability etc.), this methodology is only concerned with response and turnaround time and how that can be used to aid capacity management.

Service level agreements can also pin point areas of the processing environment that may benefit from tuning. Where a projection indicates that the level of service for a business unit will become inadequate at a certain point in time, steps can be taken to tune the effected areas and ensure the required level of service is maintained. This may delay the acquisition of additional resources thus providing a cost saving to the organization.

The process of establishing and maintaining service level agreements involves a number of steps:

1) Determine service level groups;
2) Determine the current level of service for each service level group;
3) Ascertain required levels of service for each service level group;
4) Monitor, reporting and tuning; and
5) Negotiate the service level agreements.

DETERMINE SERVICE LEVEL GROUPS

Once service level agreements are established they require close monitoring and quick action in the event of degrading responses. It is therefor necessary that the number of agreements be restricted to a manageable limit. This is achieved by classifying homogeneous work (similar users; transactions; or business units.) into groups and developing service level agreements for each group.

These groups can be made up of:

- users of the same transaction or transactions;
- users performing similar business functions;
- users in a similar geographical area;
- individual applications; or
- groups of applications.

☞ Determine and document service level groups.

When determining the composition of service level groups it is important to consider the management and maintenance of the final service level agreement. It would be difficult to maintain a service level agreement for a service level group that is made up of users or transactions of vastly different profiles.

E.g. two business units, one performing simple text data entry the other manipulating complex graphics, should not be included in the same service level group.

The composition of each service level group should be clearly documented and approved by senior IT management

☞ Obtain approval of the make up of each service level group

The information collection systems established during phase 1 will provide sufficient information to enable the development of service level groups.

There are a number of factors that must be determined before service level agreements can be established. For instance, how quickly can the IT units respond to degrading response. Most tuning changes require some component of the system to be shut down, if the production environment can only be shut down once every 2 months then for tuning there is a possible 8 week delay. Hardware acquisitions may require a budget cycle, possibly a 12 month delay. These factors must be considered while establishing service level agreements.

Once service level groups have been determined and documented the factors affecting the delivery of service can be determined. This can be done by establishing service level objectives.

ESTABLISH SERVICE LEVEL OBJECTIVES

A service level objective is a document that identifies the service level group and clearly defines the level of service the service level group is currently receiving. It is essentially the same as a service level agreement except it has not been negotiated with the users.

The document aims:

- to establish the service level agreement process in the organization;
- to get the responsible IT units thinking in terms of maintaining the level of service required; and
- to establish thresholds for the reporting mechanisms that will enable the timely resolution of response based problems.

During the establishment of service level objectives, correct monitoring and reporting techniques can be established. The triggers for tuning can be developed and an understanding of the impacts on resource acquisitions can be determined.

To establish service level objectives the end user response and turnaround times must be determined within each of the service level groups.

End user response time is the time taken from the instant when the user initiates a requests to the instant when the request is satisfied and the user has the required result. This includes the following components:

ELAPSED NETWORK TIME:
The elapsed time taken to transmit the request from the terminal to the CPU plus the elapsed time taken to transmit the result from the CPU to the terminal.

ELAPSED CPU TIME:
The elapsed time from the instant that the request was received by the CPU to the instant when the result was sent to the terminal, minus any time performing I/O synchronously to non network devices.

ELAPSED I/O TIME:
The elapsed time taken performing I/O synchronously to non network I/O devices.

Data should be collected for each of these response components within each service level group. Each of the response components can be further broken down, execution and queuing time for CPU etc. These should be collected if available, they will aid in the tuning process.

For batch turnaround time collect ELAPSED CPU TIME and ELAPSED I/O TIME.

☞ Collect the response and turnaround data for each service level group.

Data summarization

For each service level group the following statistics for response and turnaround time (broken down by network, I/O and CPU) should be calculated.

MINIMUM RESPONSE

AVERAGE RESPONSE

MAXIMUM RESPONSE

And the following percentiles
50 75 90 95 99

e.g. 50% completed in 0.5 seconds
95% completed in 2 seconds
99% completed in 1 minute

The preceding metrics are the ideal however if the monitoring software is only capable of producing average response times then that must suffice.

Statistics should be created for response as a whole and for the individual components that make up the response. This information should be added to the current on line performance database, then summarised and added to the trend database, if not already done so.

☞ Summarize the collected response data, add to the performance and trend databases.

Once an understanding of the current level of service is determined then a service level objective can be drafted for each service level group. Consultation with user groups should not be

entered into during the drafting of the service level objective. A service level objectives only purpose is as a vehicle to establish the internal mechanisms necessary for the correct management of the final service level agreement. The document should contain the following items and should reflect the level of service currently experienced not that required or expected by the service level group.

ONLINE RESPONSE

Expressed as a sliding scale in terms of the percentage of transactions completed within X seconds.

99% of transactions completing in less than 10 seconds.

95% of transactions completing in less than 5 seconds

90% of transactions completing in less than 1 second

BATCH TURNAROUND

Expressed in terms of elapsed time of batch jobs in a similar fashion to response.

99% of batch jobs completing in less than 3 minutes.

95% of batch jobs completing in less than 10 minutes.

90% of batch jobs completing in less than 20 minutes.

☞ Draft service level objectives for each service level group.

Service Level Objective format.

The following format can be used when drafting service level objectives and service level agreements.

SERVICE LEVEL GROUP A

Describes the make up of the service level group. The applications in use, the different business entities etc.

REQUIRED LEVEL OF SERVICE

BATCH

99% of jobs completing in less than 10 minutes.

95% of jobs completing in less than 5 minutes.

ONLINE

95% of transactions completing in less than 5 seconds

Figure 2 Service level objective format

With the establishment of these objectives a target is set. All effort must now be focused on maintaining the current level of service for each service level group. Close monitoring of the levels of service delivered is now necessary.

MONITORING

Comparison should be made on a regular basis between the service delivered and the service level objective. Where the service level did not meet the objective further investigation should be carried out to ascertain why. This comparison should be automated providing notification before the level of service falls to the point the service level objective is not being met. Remember the aim is to ensure that the level of service delivered continues to be the same as that specified in the service level objective. Statistical Process Control (SPC) can be of use here.

Statistical Process Control

To use statistical process control in this context it is necessary to first define some terms.

POPULATION All transactions, past, present and future.

SAMPLE A subset of the population.

Statistical process control is based on the fact that the mean of a sample is an estimation of the mean of the population the sample belongs to.

If we determine the historical mean response and call this the population mean or baseline mean then if no change occurs to actual response, future daily mean responses values should be approximately that of the baseline mean.

We can assume the response data approximates a normal distribution (precision is not necessary as an early indication of changing response is all that is required). Figure 2 shows a normal distribution curve.

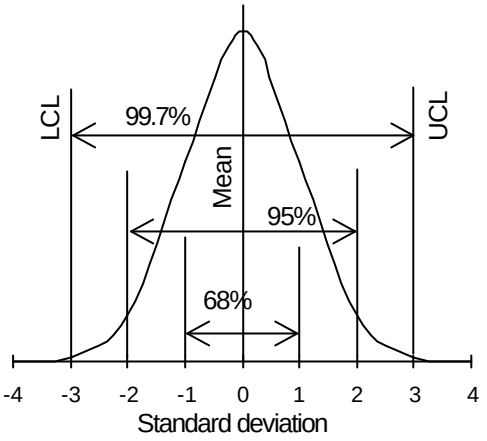

Figure 3 Normal distribution

The 68 - 95 - 99.7 statistical rule for normal distributions states that 68 percent of daily averages should fall between plus and minus one standard deviation about the mean, 95 percent of daily averages should fall between plus and minus two standard deviations about the mean and 99.7 percent of daily averages should fall between plus and minus 3 standard deviations about the mean.

The point +3 times the standard deviations above the mean is known as the upper control limit.

From this rule a number of probabilities can be calculated.

- One point outside the 3 times standard deviation level has a probability of 0.003.

  That is 0.15 percent of all daily response measurements should fall above the upper control limit and is the result of normal fluctuation. In other words about 1 day in every 2 years.

- A run of nine points on one side of the center line has a probability of 0.002.

  that is 0.2 percent of all runs of 9 daily response measurements should fall above or below the average and is the result of normal fluctuation.

- Two out of three points beyond the 2 times standard deviation level on the same side of the center line has a probability of 0.002

These probabilities form the basis for rules that are applied to the collected response data to determine if response is degrading.

Care should be taken to determine rules that indicate in the environment being monitored degrading response. If each daily mean response continues to obey the rules the response is said to be in control (i.e. No change).

E.g. the probability that the daily average response falls above the average on two consecutive days and the apparent increase is a result of normal fluctuation and not an underling response increase is 0.25 or about 25 percent. That is 25 percent of consecutive two day runs will be above the average. This would most probably not indicate an increase in response. Three consecutive days yields 12.5 percent, and possibly indicates degrading response., four consecutive days yields 6.25 percent and almost certainly indicates degrading response.

Statistical process control aids the management of service level agreements by providing an early indication of response degradation. From the previous example, if the average daily response for four consecutive days is even marginally above the baseline average the there is a good possibility that response has increased and further investigation could be carried out. Such small increases in average response can be detected using this method and will probably not cause the service level agreement to fail.

ESTABLISHING CONTROL CHARTS

A control chart enables the easy comparison of the daily average response values with the base line response.

We can use the data used to determine current level of service for each service level group as the base line data for the construction of a control chart. The mean and standard deviation of the data are calculated. A chart can be constructed that shows the mean, the upper and the lower control limits (+/- 3 times the standard deviation).

☞ Determine baseline response and standard deviation for each service level group.

Figure 3 shows a sample control chart.

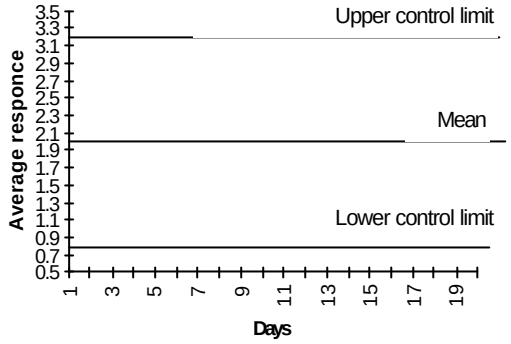

Figure 4 Sample control chart

By plotting the average daily response over the control chart early notification of changes in response can be shown.

A control chart should be constructed for each service level group allowing close monitoring of each service level agreement.

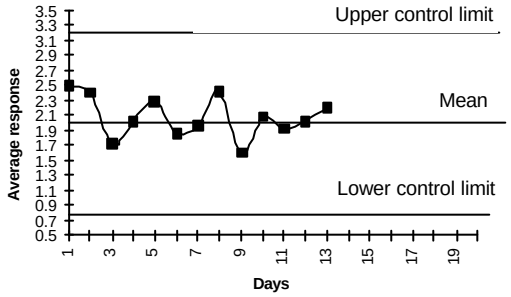

Figure 5 Control chart tracking daily average response

> ☞ Construct a control chart for each service level group.

Where possible chart construction should be automated along with automatic detection and notification of daily response values that cause one of the probability rules to be broken. This event should trigger further investigation into the cause of the apparent response increase.

> ☞ Implement automatic notification of response degradation.

CHART MAINTENANCE

If the response data is unstable some of the more precise probability rules may trip when response is not increasing. In these cases further investigation must be carried out to determine a more suitable set of rules.

The control charts established will continue to be useful for tracking changes in end user response until there is an underlying change in the baseline. That is, a change that causes the average response or the standard deviation to change. These changes can be a result of hardware acquisitions, increased user activity, change in software etc.

If the change is degrading response, all effort must be made to restore response to the established baseline. However if this is not possible or response has improved then a new baseline for the control charts must be calculated.

> ☞ Calculate new baseline whenever the baseline changes and cannot be restored.

REPORTING

Information covering the service level agreements should be included as part of the regular reporting schedule established in PHASE 1 of this methodology.

A table of the following format should be included in the monthly report with explanations of any periods when service level objectives were not met.

SLO	Objectives	Delivered	Delta
Group A			
<10 sec	99	98	-1
<5 sec	95	94	-1
<1 sec	90	90	0
Group B			
<10 sec	95	98	3
<5 sec	90	96	6
Summary	94.40	94.60	7

Table 1 Service level objective reporting

> ☞ Include service level data in the regular report cycle.

In the preceding table we see that service level objectives were not met for some individual groups but overall they were. An explanation as to the cause of the poor results must be included in the report. This case indicates possible options for tuning by moving processing resources from GROUP B to GROUP A. If done carefully both groups may be able to meet their service level objectives without further expenditure on additional processing resources.

When it is determined that a change is necessary to maintain a service level objective the time taken to implement the change should be documented. These times for each type of change (Hardware acquisition, tuning etc.) form the rules of thumb that must be applied when determining the lead time necessary when ordering changes.

E.g. A site takes two months to upgrade a network line. Any notification of degrading response must consider the lead time of two months necessary to rectify the problem.

☞ Document the time taken to implement changes.

Once the reporting has been established key IT units must receive a copy and their responsibility for the maintenance of the service level objectives detailed. These groups should include System software groups; Network groups; Database groups; Application development groups; and Operations.

On completion of 12 consecutive months of service level objectives monitoring and reporting the establishment of service level agreements can commence.

ESTABLISH SERVICE LEVEL AGREEMENTS (SLA)

The establishment of service level agreements commences by negotiating with representatives from each of the service level groups. The required level of service for each group is ascertained and expressed in the same terms as the service level objectives.

☞ Ascertain the required level of service from each service level group.

The service level agreement should reflect the stated requirement, not the actual current level of service delivered. This may provide some scope for the particular service level group to grow and also allows the possibility of redeploying processing resources to other service level groups.

The data gathered during the negotiations with the service level groups should be documented in the service level agreement. This document takes the same form as the service level objective.

☞ Document the service level agreement.

The completed service level agreement should be signed by both parties and must be considered a contract between the IT management and the users of the IT resource. All possible effort should be made to achieve and maintain the level of service stated in the service level agreement. Failure to do so results in a lack of confidence in IT support and eventually the processes established to define service level agreements will fail.

TUNING DRIVEN BY SERVICE LEVEL AGREEMENTS

Where it is determined that service level agreements are about to fail for a particular service level group then the first step in resolving the failing service level agreement is tuning.

☞ Tune to maintain service level agreement

The tuning should be directed at improving the response for the affected service level group, this may result in the degradation of response for another service level group which is acceptable provided that the service level agreement of the degraded service level group continues to be met.

By examining the various components of the service level group's response (i.e. CPU time, I/O time and network time) a determination as to the best area to target tuning efforts can be made.

In some cases tuning will not improve the response sufficiently to meet the service level agreements, here an upgrade to one of the response components must be considered. In most cases upgrading network components provides the greatest benefit, followed by upgrading the I/O sub-system and lastly the central processing unit. This order is also the cheapest to the most expensive and in relation to impact on the users runs in impacting the least number of users to impacting the most number of users.

In general the following list details the options available to improve the level of service delivered.

1) general tuning;

2) redeployment of processing resources from another service level group, provided the service level agreement for the other group continues to be met;

3) upgrading existing processing resources, such as network lines etc.; and/or

4) acquiring additional hardware

These options should be examined in the order specified above, in most cases this is from cheapest to most expensive.

In the case of option 2 above, some circumstances will arise that can be satisfied by moving processing resources from service level group "A" to service level group "B", even if this results in the service level agreements for group "A" failing to be met. These situations occur when the business function of group "B" is considered more important than those of group "A". These decisions should however be made by senior management in consultation with the affected groups.

PHASE 3

Phase 3 of this methodology utilizes one of the modeling tools available such as Best1. To determine future response, based upon the current workload and the projection of future workload gathered as part of phase 1.

This future expected response is compared with the service level agreements developed as part of phase 2. If the service level agreements will be impacted then steps can be taken to improve the situation, such as tuning, resource upgrade or resource acquisition.

The details of Phase 3 are best left for the vendors/manufacturers of the chosen modeling tool.

☞ Model Future expected workloads to determine effect on end user response.

CONCLUSION

The term "Capacity Planning" is a misnomer we do not plan capacity we "Manage Capacity".

Capacity Management is the art of monitoring, reporting and balancing the business needs of the organization with the capabilities of the processing resources. There can be no separation of the technical aspects from the business aspects of capacity management.

In any organization the IT resources are there to support the business function. It is therefore the business function that drives IT resource consumption. For this reason this methodology focuses on the business functions.

By collecting data from the various business areas. Storing this data along with various technical metrics in well defined performance databases, and associating the business data with the corresponding resource consumption data, an understanding of future workloads can be obtained.

This information is however not sufficient for cost effective capacity management. Upgrading IT resources because they are full will often prove a costly exercise. An understanding of the effect of full processing resources on the end user response will provide the necessary mechanism to ensure expenditure on IT resources is in accordance with the business functions driving them.

Designing service level groups; setting service level objectives; establishing the internal mechanisms necessary to ensure maintenance of the service level objectives; and finally negotiating service level agreements establishes the knowledge base necessary for the correct interpretation of models of future workloads against end user response.

Only after the impact of future resource consumption on end user response has been fully understood can a capacity planning document hope to portray the most cost effective solution for the organisations future IT requirements.

The final phase in this capacity management methodology (Phase 4) provides mechanisms for the establishment of a rudimentary cost attribution system. This system associates the IT resource consumed by a business unit with the actual cost of maintaining that IT resource. Including any future expected resource acquisitions. The system does not have all the bells and whistles of an off the shelf package but correspondingly does not have the high cost of purchase or maintenance associated with most of these packages.

REFERENCES

Introduction to the Practice of Statistics
ISBN 0-7167-1989-4
(David S. Moore & George P. McCabe)
W.H. FREEMAN AND COMPANY

An Introduction to probability and statistics
ISBN 0 85564 110 X

(Beryl Hume)
University of Western Australia Press

Capacity Management A Methodology Phase 1: Business Transactions And IT Resource Requirements
(Adrian Heald)
CMGA ’94 Proceedings