# CAPACITY MANAGEMENT
# A METHODOLOGY
# Phase 1: Business Transactions and IT Resource Requirements

**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 1 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

Capacity Plans are simply a reflection of the impact of *business objectives* on *computer resources*. The Capacity Plan documents that impact, provides estimations of future resource requirements and establishes an understanding of the effect of delays on the acquisition of required resources.

The roles of Capacity Management can be defined as:

**Reporting**: provide timely, meaningful, easily understood reports to management. These reports increase management awareness of IT issues and provide detailed information on the impact of future organisational directions on the processing environment.

**Planning:** collate information gathered from technical and business arenas; perform detailed analysis on this information; and produce plans detailing current and future processing environment requirements.

**Tuning:** make recommendations for changes which result in improved performance of the processing environment and which facilitate attainment of corporate objectives.

This methodology utilises business transactions to ensure a correct understanding of the impact of future changes in direction of an organisation; the requirements for additional resources; and the business effects of delaying resource acquisitions. It is primarily concerned with Capacity Planning for the Major IT resource of Central Processing Unit (CPU), the principals however can be applied to other IT resources such as DASD, Network and in the client server environment.

The methodology consists of four phases:

**Phase 1:** Establishes data collection and the association of IT resource consumption with actual business transactions. From this association a primitive capacity plan can be produced which provides estimations of when IT resources become "full" based on the level of business activity. No formal consideration is given to transaction response.

**Phase 2:** Establishes Service Level Agreements (SLA's), which indicate the response required by the user population in order for them to function efficiently. If these Service Level Agreements are not being met, tuning, an upgrade or the removal of less important services may be necessary.

**Phase 3:** Models the results of phase 1 to determine the impact on the Service Level Agreements established in phase 2. This allows the production of a sophisticated capacity plan that recommends action before Service Level Agreements are impacted, therefore ensuring user efficiency is not degraded.

**Phase 4:** Establishes resource cost attribution which associates the cost of future resource acquisition with the business entities consuming them.

This paper presents an implementation strategy for Phase 1, subsequent papers will provide strategies for Phases 2 and 4. Phase 3 is best left to a proprietary product.

## DATA COLLECTION

The identification of sources of both business and technical information and the construction and maintenance of capacity management databases are key aspects in successful capacity management.

Summarising and analysing the collected data will produce information. This information will lead to an understanding of the various factors affecting the performance and capacity of the processing environment.

Data required for capacity management can be described as either:

**Classifying Data**
Data elements that describe what is being measured.

or

**Quantifying Data**
Data elements that are the actual measurements.

## Business Data

Capacity Management relies on an understanding of organisational business information. This information can be obtained from management activities such as, steering committees, executive meetings or change control systems. Identifying these areas and gaining access to their reporting mechanisms provides the Capacity Manager with a supply of information on the future directions of, and changes within, the organisation.

☞ Identify, and establish access to, information sources that may be useful for Capacity Management.

Sources for such information may include:

- regular gazettes;
- information memo's, newsletters;
- minutes of meetings;
- requests for service from user groups;
- general business trends;
- change control instructions;
- Corporate and/or Strategic plans; and
- Interviews with business representatives.

Determining the usefulness of the information at this stage is not absolutely necessary. When an understanding of the association between the various business and technical aspects is developed filtering can occur to remove the less valuable or redundant data.

Collection of business data is not restricted to within the organisation. General statistical data will be available from various sources (such as the Australian Bureau of Statistics).

The form of the actual data is largely determined by the functions of the organisation. In general, the data should be collated so as to reflect the reporting requirements of the organisation. This can be achieved by identifying business entities, within the organisation.

A business entity can consist of specific functional areas within the organisation, cost centres, or individuals. However, as capacity planning will be based on changes within these business entities and the changes will be determined mainly by interviews and surveys, a practical number of entities should be determined.

External business entities should also be identified and documented.

☞ Identify business entities, both internal and external.

For each business entity the following data is the minimum that should be collected.

### BUSINESS CLASSIFYING DATA

- A description of the business entity, including identification of its functions; and
- A description of any of the events identified in "QUANTIFYING DATA" along with expected implementation dates.

### BUSINESS QUANTIFYING DATA

- Any data relating to functions carried out by this business entity. The number of registered users for each application used within the business entity is the minimum data required here; and

- Any data relating to external factors likely to affect this business entities utilisation of the processing environment.

Some examples of business measures are:
- number of registered users of an application
- number of documents filed
- average number of pages per document.
- number of customers served
- number of widgets produced

This data will be used for association between measurements the business representatives understand and the technical measurements of the IT environment.

☞ Collect and classify data for each identified business entity.

## Technical Data

The collection of technical data commences with the identification and establishment of monitoring software which can attribute resource consumption back to the various business entities identified previously.

The monitoring software must be capable of providing data in a form that can be stored in a database and be readily manipulated.

☞ Identify, establish and document relevant monitoring software.

Generally there are two types of data available, *Interval* and *Event.*

**Interval Data**

Interval data consists of records written throughout the life of a task which provide periodic data about the task, the records are provided at a set time interval. For the correct association between business and technical data it is essential that interval data be collected.

**Event Data**

Event data consists of records written when a certain event occurs, such as the completion of a batch job or on line transaction. In most cases event data can be converted to interval data for correct association with business data.

Most IT processing environments consisting of business entities, applications and processing environments lend themselves to a hierarchal structure of usage. That is, a *business entity* utilises many *applications* and many applications utilise a *processing environment*.

In the case of client server applications *business entities* utilise many *applications* and many applications utilise many *processing environments*.

For each combination of *business entity*, *application* and *processing environment* the following data should be collected.

### TECHNICAL CLASSIFYING DATA

The following list identifies the minimum classifying data required:

- Processing environment name;
- Business entity name;
- Application name;
- Start time of interval; and
- End time of interval.

### TECHNICAL QUANTIFYING DATA

The following list identifies the minimum quantifying data required:

- Number of users within business entities accessing the application;
- Processing resource consumed; and
- Number of transactions.

☞ Ensure correct technical classifying and quantifying data is being collected.

### APPLICATION OF THE CAPTURE RATIO

Most monitoring software is incapable of attributing all CPU utilisation to the actual consumer of the CPU. Generally the more detailed the breakdown of resource utilisation the less is seen and attributed by the monitors.

In a typical processing environment a "hardware monitor" monitors 100% of the resource consumed in the processing environment but not the consumers of the resource. A "subsystem software monitor" monitors the general workloads consuming the resource but less than 100% of the resource consumed can be attributed to these workloads. The "databases" or "transaction processors" monitor the individual users consuming resource but less than 100% of the resource consumed within the workload can be attributed to these users.

The data collected by the databases and transaction processors is used to attribute resource consumption to business entities and

applications. This clearly will not attribute all resource consumed by this processing environment. The application of a capture ratio will solve this problem.

Capture ratio's can be calculated using simple linear regression. See appendix "A" for a detailed explanation of the calculation and application of a capture ratio.

### APPLICATION OF THE PEAK TO AVERAGE RATIO.

The volume of data collected and stored must be tempered by the cost of the collection and storage of the data. Collecting data at the transaction or even program level is the ideal but the cost of storage and in some cases actual monitoring at this level is prohibitive. The aim of the capacity plan is to enable the acquisition of sufficient resource to cater for peak processing requirements.

A peak period may be as small as 15 minutes in a year but to predict which 15 minutes in the next year will be the yearly peak is impossible. A better period is the peak day.

Another factor that must be considered is the accuracy and reliability of the business planning information. Most organisations will be unable to accurately predict business transaction loads 1 or 2 years into the future. As the projection of future resource consumption is based on the business data, and this business data may have some in built inaccuracies, the development of data collection and storage systems that are accurate to many decimal places is not necessary.

It is better to collect data as daily averages (that is 1 record per business entity per application per day) and to understand the inherent inaccuracies, than to collect data for a smaller interval and potentially become the biggest consumer of resource on the processing environment. In any event data collection at the smaller interval will not improve the accuracy of the base planning data for the business entities.

If the peak period is less than one day then the application to the daily averages, of a peak to average ratio will ensure peak processing loads are reflected. See appendix "B" for an explanation of calculating a peak to average ratio.

## CAPACITY MANAGEMENT DATABASES

The data contained in the Capacity Management database is the heart of the entire system. This data will be utilised for planning, reporting and tuning purposes; and will contain detailed statistical descriptions of the processing and business environments.

The following simple database structure is the minimum required for capacity planning as described by this methodology.

### *ON LINE* Capacity Database

This database should contain sufficient data to allow analysis of the previous month, approximately 35 days data. Minimal summarising is performed on this database so as to provide the most detailed data available for analysis.

### *OFF LINE* Capacity Database

This database contains data from previous months i.e. one off line database per month. They provide the opportunity for detailed analysis of previous months. Minimum summarising is performed on these databases.

### *TREND* Database

The TREND database contains summarised data. This database reflects general trends in key aspects of the processing and business environment.

Data collection and manipulation routines should be automated where ever possible. This task will be mainly driven by the capabilities of the individual monitoring products and the operating environment in which the capacity management system runs.

☞ Establish databases and data collection routines.

## MONITORING AND REPORTING

The establishment of the Capacity Management databases allows the commencement of regular reporting to management.

☞ Establish weekly and/or monthly reports to provide simple utilisation and response information.

There are several important points to consider when producing these reports.

- They should be written to the audience they are intended for. Reports for senior management should not contain detailed technical data.
- They should be appealing and not cluttered.
- Regular reporting has the tendency to show the same thing in each report. While this provides a "warm" feeling, changes in the reported data can often be missed and therefor should be highlighted.
- They should be regular. Establish a publishing schedule.

## THE FIRST CAPACITY PLAN

The development of a Capacity Plan is accomplished by dividing the process into the logical steps set out below:

1. The development of Business System Profiles.

2. The development of Application System Profiles.

3. The identification of changes (business and technical) that could impact the computing resource. These changes should be categorised as being definite changes or possible changes.

4. The identification of the relationship between the Business System Profiles and Application System Profiles.

5. The identification and incorporation of growth that is unrelated to any known changes occurring.

6. The estimation of future resource requirements.

The following depicts the structure of the finished Capacity Planning Document.

**EXECUTIVE SUMMARY**

**BUSINESS SYSTEM PROFILES**

**APPLICATION SYSTEM PROFILES**

**SCHEDULE OF OUTCOMES**

## The Executive Summary

The executive summary is produced separately from the main part of the Capacity Plan but should always be included in the completed document. The production of the executive summary is best left until after the main body of the document has been completed.

### EXECUTIVE SUMMARY FORMAT

The executive summary should contain the following sections.

**OVERVIEW**

Describes what the document is and list any expected expenditure required to support the predicted resource consumption.

**METHODOLOGY**

A brief explanation of the methodology used to produce the capacity plan.

**COMPARISON WITH PREVIOUS PLAN**

Identifies and explains any discrepancies found between the previous predictions and what actually happened. This section should be supported with graphs showing the previous predictions compared to the actual consumption.

**PLANNING INFORMATION**

Details of where the information used to make the predictions came from.

**TABLE OF SIGNIFICANT EVENTS**

A table detailing the significant events likely to affect resource consumption and the time frame of these events.

**EXTERNAL USAGE**

Describes external organisations which utilise the processing resources. This section should be supported with graphs which compare internal and external utilisation.

**SERVICE LEVEL AGREEMENTS**

Describes the current service level agreements, detailing points in the future when they will not be met. A brief summary of the impact of this should be included. (THIS WILL ONLY BE POSSIBLE AFTER THE COMPLETION OF PHASE 3 OF THIS METHODOLOGY)

**RECOMMENDATIONS**

Details of any recommendations including the specifics of any resource acquisitions required.

## Business System Profiles

The establishment of a business system profile for each identified business entity is the first stage in determining the association between IT resource consumption and the business functions that the IT system supports.

The profile for a given entity will consist largely of a narrative describing the nature of the

business entity and the way in which the various computing systems fulfil its requirements.

Business measures need to be identified. These are the performance and volume indicators which best reflect the workload generated by the business and which are used as the fundamental input to the plan. Factors which affect these business measures also need to be identified. It is important to remember that these measures need to be correlated with actual resource usage.

☞ Identify business measures

The primary source of information relating to long term significant change in the usage of computer resource should be Corporate and Strategic management plans. It is not necessary to be conversant with the entire plans, however it is necessary to be able to identify those items that relate to the data processing workload. The number of users to be supported, the groups they fall into, any changes to the functionality of applications or Corporate entities, and any additional applications should all be identified.

### BUSINESS SYSTEM PROFILE FORMAT

The following is a general layout of the business profile.

**BUSINESS ENTITY NAME**

**Entity description**

Describes the Business entity. It should contain a brief description of the entities activities, and a cross reference to computing systems used to support these activities.

**Business Measures**

Identifies any business measures available for this entity.

**EXTERNAL ENTITIES**

Where possible a description of external business entities should be included in the Capacity Plan. This information can be gathered from discussions with these users or application support groups.

☞ Establish business system profiles

## Application System Profiles

A description of each application system should be included in the plan. The information needed to build these descriptions can be obtained from application design documentation, user guides, technical system overviews and application development staff. All new applications that will be implemented during the period covered by the Capacity Plan must be included.

### APPLICATION SYSTEM PROFILE FORMAT

An application system description should have the following general format.

**APPLICATION SYSTEM NAME**

**System Description**

This should be a concise description of the system, including a high level description of the various functions performed.

**Initial Justification**

This is a statement as to the primary benefits which are associated with the operation of the application system. This is of value in the preparation of supporting information for the inclusion in expenditure proposals which may be generated through the expansion of or addition to this application system.

**Business Configuration**

The business configuration section identifies business entities that utilise this application. The associated business transaction volumes, number of registered and active users and the quantity of stored data (in business records not megabytes). This section cross references to the business system profile section

**Technical Configuration**

This section sets out the actual utilisation and an indication as to the relationship between technical transactions and the business measures described above. See Appendix "C" for further details on the association of business and technical measures.

**Future Status**

This describes the changes in the system, either planned growth or new initiatives, and should be set out under headings which succinctly describe the nature of the change. The heading should include the time frame of the change to the nearest quarter. There should be one "FUTURE STATUS" section for each significant change. The information required for this section can be determined through interviews with appropriate application development staff and user representatives.

The Future Status section should contain an indication of which area of resource consumption will be affected by the change, and present some indication of the degree of change.

**Summary**

This section contains the net effect of all future status sections, normally in tabular or graphic form.

**NON-DEPARTMENTAL COMPUTER SYSTEMS**

A brief description of computer systems used by the organisation which run on other organisations environments should be included for completeness only, no resource usage or predicted growth need be specified.

☞ Establish application system profiles.

### PLANNING FOR NEW APPLICATIONS

As information for new applications may be scarce, educated guesses about their utilisation must be made. These guesses will become more accurate as the application passes through the various implementation phases. Only once the application is fully implemented in production will it's true utilisation be known.

**CONCEPTION PHASE**

Look for an application of similar function already implemented, and use the utilisation of this as the estimate. The application profile for this similar application will include resource consumed per transaction. Application system analysts will provide estimated number of active users and transactions issued. Simple multiplication will provide an educated guess of the resource consumption of the new application.

**DEVELOPMENT PHASE**

Measure the transactions of the application under development, taking care to eliminate any outliers. Multiply this with the estimated number of transactions, this then replaces the guess made during the conception phase. Development transactions may not however reflect the type of transaction that will be run in a production environment.

**TESTING PHASE**

Measure the transactions of the application under test, eliminating any outliers. This value replaces the development phase value. Data obtained from the test environment should be a lot closer to that expected in a production environment.

**PRODUCTION IMPLEMENTATION PHASE**

Where possible the production implementation of a new application should be phased. An initial trial group consisting of randomly selected users from the registered user population should be first to gain access to the application. This trial group will reflect the utilisation of the whole population of registered users. From the trial period obtain the following data.

- Ratio of the number of trial users to the number of active users.
- Resource consumed per active user (See Appendix "C").

Applying the ratio of trail users to active users to the total number of registered users will indicate the number of active users of the application when fully implemented. Multiplying the estimated number of active users by the resource consumed per active users will indicate the average utilisation of this application when fully implemented.

## Schedule Of Outcomes

The "schedule of outcomes" section consists of a table containing the summary information from each application system, with an indication of the net effect of all future changes. This is the impact on the available IT resource of all known changes to the business and technical arenas.

Most organisations will not be able to accurately identify all changes that will occur during the period covered by the Capacity Plan. The amount of change identified will be largely dependant upon the business of the organisation itself. The Capacity Plan must however cater for all outcomes, therefor some estimation of "unknown change" must be made. See Appendix "D" for details of estimating unknown change. The impact of "unknown change" should be identified in the schedule of outcomes.

Any future resource requirements should be noted in the Schedule of outcomes.

## CONCLUSION

The collection of business and technical data; the establishment of capacity planning databases; the commencement of regular reporting to management; and the establishment of a standard capacity planning document has enabled the development of a primitive capacity plan. This plan however will make recommendations for resource upgrades based solely of the "fullness"

of the processing resource. No consideration has been given to the required response of the business.

Phase 2 of this methodology establishes Service Level Agreements (SLA's) which determine and track the response required to support the business functions.

Phase 3 models response based on projected workloads to determine when Service Level Agreements will be impacted.

It may not always be possible to follow this methodology precisely but if it's general direction is followed a useful, simple to manage Capacity Management system will result.

## APPENDIX A - APPLICATION OF CAPTURE RATIOS

The following SAS code fragment will determine capture ratios:

```
PROC REG;
        MODEL TOTALCPU = STCCPU TSOCPU
CICSCPU BATCHCPU DBASECPU;
```

Where:

| | | |
|---|---|---|
| TOTALCPU | = | Total CPU utilisation (as measured by the "hardware monitors") |
| STCCPU | = | CPU consumed by Started Tasks |
| TSOCPU | = | CPU consumed by TSO users |
| CICSCPU | = | CPU consumed by CICS regions |
| BATCHCPU | = | CPU consumed by BATCH jobs |
| DBASECPU | = | CPU consumed by database engines (as measured by the "subsystem monitors") |

The parameter on the left of the "=" sign is the dependant variable, the parameters on the right are the independent variables. That is TOTALCPU is dependant upon STCCPU, TSOCPU etc.

From the resultant output the R-SQUARE and PARAMETER ESTIMATE values are extracted.

Key points in the validation of the regression model are:

- The INTERCEPT value must be small compared with the maximum CPU utilisation value. The INTERCEPT value shows the amount of resource not attributable to any of the independent variables. If this shows a large value then perhaps all workloads are not included in the independent variable list;
- Each PARAMETER ESTIMATE must be greater than 1. The parameter estimate shows the affect on the dependant variable (TOTALCPU) of an increase of one unit in the independent variable (STCCPU, TSOCPU etc.). That is, if the parameter estimate for BATCH is 1.25 then for each second of CPU recorded in the subsystem records for BATCH 1.25 seconds of CPU were actually consumed in the processing environment hardware; and
- The R-SQUARE value should be greater than 0.95. The R-SQUARE value expressed as a percentage shows the percentage of the variation in the dependant variable (TOTALCPU) which is explained by variations in the independent variables (STCCPU, TSOCPU etc.).

The inverse of the PARAMETER ESTIMATE is the capture ratio. That is if the parameter estimate for BATCH is 1.25, 1/1.25 = 0.80 that is 80% of CPU consumed by BATCH is captured. In other words 1 second of CPU accumulated in the subsystem record (BATCHCPU) for BATCH results in 1.25 seconds of CPU being accumulated in the hardware record (TOTALCPU).

## APPENDIX B - APPLICATION OF THE PEAK TO AVERAGE RATIO

In order to reflect the peak processing load of a system and to ensure a minimum volume of data is collected the use of peak to average ratios may be used.

If data is collected as daily averages and the minimum period of resource constraint is the prime working period say 09:00 to 17:00 then the ratio of the utilisation in the historical peak prime period to the average utilisation must be calculated.

By plotting the daily averages over the past year the peak daily average can be determined. By examining the off line monthly database for this period the average utilisation during the prime period 09:00 - 17:00 can be ascertained. Divide this peak value by the average utilisation of the month that the peak period falls in and the result is the peak to average ratio. Any projections based on daily averages should be multiplied by the peak to average ratio value in order to reflect the peak processing load.

## APPENDIX C - ASSOCIATING BUSINESS MEASURES AND RESOURCE CONSUMPTION

How much CPU resource does it take to perform a particular business function? This question can be answered by analysing the association between the collected business data and the CPU consumption for the particular business entity. This is done using simple linear regression. As for regression in determining capture ratio's the model must be valid.

e.g. For a particular application and business entity CPU MIPS and the number of widgets produced have been collected over a period of time on an interval basis. Regression analysis provides the following:

| | | |
|---|---|---|
| R-SQUARE | = | 0.7308 |
| INTERCEP | = | 0.1445 |
| WIDGETS | = | 0.0028 |

The R-SQUARE value shows that 73% of the variation in CPU utilisation is explained by the variation in the number of Widgets produced. This value is high enough to consider this model good, in any event if widgets produced is the only business measure available the model must suffice.

The INTERCEP value shows the amount of CPU consumed when there are no widgets produced and is in effect the overhead of this system.

The WIDGETS value shows that for every Widget produced 0.0028 MIPS will be consumed.

## APPENDIX D - PLANNING FOR UNKNOWN CHANGE

The ability of an organisation to accurately predict change will largely be dependant upon the management ethos of the organisation.

e.g. An automobile manufacturer will have a very good idea of the number and type of vehicles that will be produced over the next five years and therefor the amount of unknown change should be small. Contrasted to a government department who will not know which political party will be setting their direction after the next election.

The amount of unknown change can be estimated by examining the previous capacity plan. If no plans exist then a best guess must be made.

From the previous capacity plan at each quarter year calculate the percentage difference between the actual utilisation and the predicted utilisation.

For each quarter year where a difference exists, determine why. This information will be included in the "Executive Summary". If the difference is due to enhancements in the capacity planning process this discrepancy will not occur in the new plan and this value can be discarded.

Separate all remaining values into positive and negative and average these two groups. The resultant percentages are simple estimations of how much unknown change exists within the organisation, and can be applied to the predicted utilisation each quarter year.

Some causes of unknown change can be eliminated by educating the providers of business information. As they become familiar with capacity management requirements, their information will become more accurate.