Capacity Management: An Overview

Outlines the major areas of capacity management, including a list of the steps required to implement a capacity process or function within an IT organisation. We briefly describe how each activity (step in the process) should operate. The model is based on the assumption that a capacity plan is required in the near future.

CapProcess
Processes that work!

Philip Bailey
Director and Principal Consultant

Email: pbailey@capprocess.com
Phone: +61 4 13 487 701 (or 0413 487 701)

Table of Contents

Executive summary

The IT capacity management function has existed since the early 1980s. It grew out of the need for capacity management of the key capital equipment in a factory environment and telecommunications systems (erlang measures) from the early part of the 20th century. It was intended to manage capital acquisition costs (equipment/hardware), which have the biggest impact on budgets, both capital and operating expenses.

ITIL expanded the scope to include operational items (commodities – software licences, tapes, etc.).

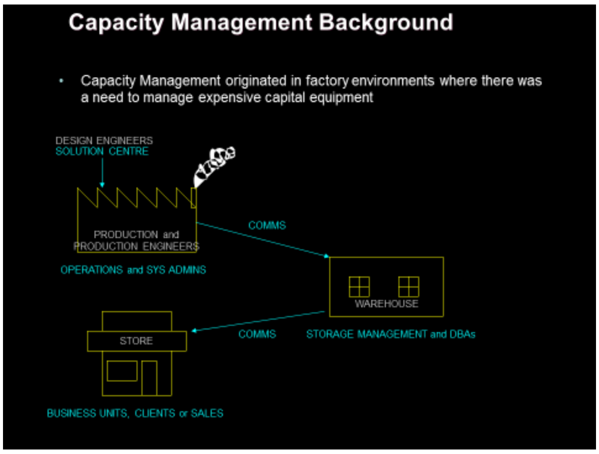
The capacity management knowledge domain is described in a way that provides a level of detail that enables organisations to obtain the benefits promised within ITIL.

A supported capacity management function and/or process can assist the organisation to maintain and improve its service delivery to customers as well as reduce costs.

Recommendations

Organisations should consider implementing the capacity management activities that eliminate or reduce the impact of poor IT delivery in a way that integrates with finance. This requires management commitment, selecting appropriate staff with the right capabilities and mind-set and providing the necessary tools so that the staff can perform their roles effectively and efficiently.

The recommendation is to create a set of processes for your capacity management activities that interconnect and flow so that your staff are productive and provide timely guidance to your organisation about its IT needs.



Capacity Management Background

- Capacity Management originated in factory environments where there was a need to manage expensive capital equipment

Conclusion

Good capacity management is noticeable in organisations from the lack of issues that affect the customer base.

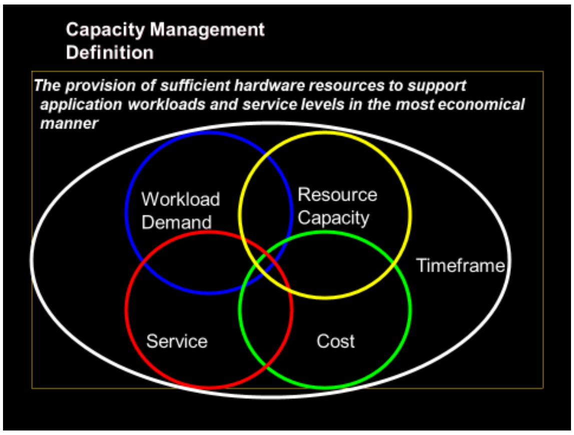
An effective capacity management process/function operates at a strategic level rather than at an operational level.

Introduction

The IT capacity management function has existed since the early 1980s. It grew out of the need for capacity management of the key capital equipment in a factory environment and telecommunications systems (erlang measures) from the early part of the 20th century. It was intended to manage capital acquisition costs (equipment/hardware), which have the biggest impact on budgets, both capital and operating expenses.

The problem for organisations that seek to implement a capacity management process is that little useful documentation (books and papers) exists explaining how to approach the task.

A process has a purpose of delivering defined or agreed outcomes.


Capacity Management Definition

The provision of sufficient hardware resources to support application workloads and service levels in the most economical manner

The capacity management ‘process’ as defined in ITIL consists of several processes. It includes short-term capacity management activities (often referred to in the IT industry as ‘performance management’, long-term capacity management activities (sometimes called ‘capacity planning’) and some aspects of ‘demand management’. Each of these process sets include a process for reporting current status, a process for tuning/balancing the workload and reporting on the effectiveness of the actions, and a process for ad hoc reports on the implications of planned changes to the environment (assessments or studies).

The capacity management knowledge domain is a set of integrated processes, serving such purposes as:

- manage the current situation
- manage the future (budgets and funding)
- provide advice (options, cost estimates)

The primary objective of the capacity management processes is to produce information that assists the decision making processes of an organisation (the IT equipment to buy and when, can an upgrade be deferred – for how long?). The capacity management reports (both performance management and capacity planning) should contain information that is important to the various organisational stakeholders. They should cover costs/budget and risks that can affect service while explaining the cost drivers (key applications).

In addition, capacity management is part of a functioning service delivery model and must evolve to meet the changing needs of the organisation (capacity management’s customer).

The implementation phases depend on your organisation’s priorities. The key is to implement enough of the process to address deficiencies in organisational performance and achieve some benefit early. Some work is performed in parallel with other steps and can include internal feedback loops – where the need for more information is identified and so on.

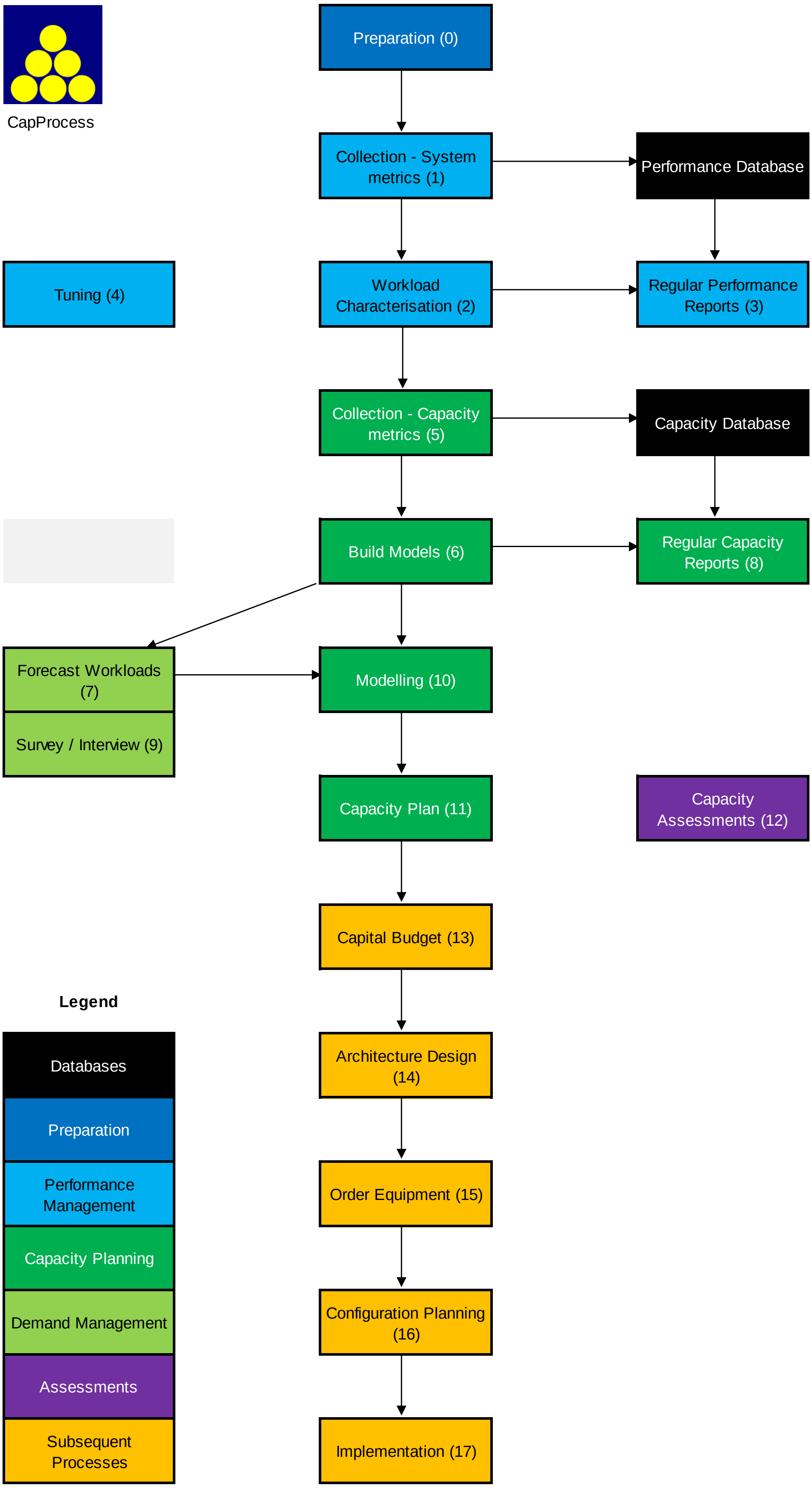


Figure 1 - Capacity Management Process - Simplified. Feedback loops are not shown.

The areas covered in Figure 1 - Capacity Management Process - include:

- preparation
- performance management
- capacity planning
- demand management
- assessments
- subsequent processes.

In practice, the areas operate in different time cycles:

- undertake – or at least consider undertaking – the preparation step each cycle through the systems (annually)
- performance management operates in the near term (daily, weekly and monthly)
- capacity planning operates in the mid-term (monthly, quarterly and annually)
- demand management also operates in the mid-term (monthly, quarterly and annually)
- assessments occur as required/requested, and
- subsequent processes occur annually and as required/requested.

Phase 0 – Preparation

Purpose

To gather information to understand the organisational environment and the best approach (priorities) to building a capacity management capability.

Definitions

This first step should define the various key concepts of capacity management. These definitions affect how data are processed and reported.

Examples include:

What is a day? Nine to five, 24 hours, 8.5 hours. How many days in a year? 250, 365? What is the scope of the capacity plan? Peaks, averages, peak peaks or average peaks. Are shifts important – such as business hours, overnight and weekend? For instance, should the peaks for these be tracked separately?

Background research

Understand the organisation and the environment it operates in and how it sees the world – and the challenges it faces. This usually requires obtaining a copy of the report to the stakeholders (for public companies its annual report to shareholders) and reading it.

Stakeholders

The key stakeholders for capacity management include the IT executive management, the business units, other process owners/manager, operations and support management and staff. There is a need to identify and engage with each group.

Library repository

A library or other repository is required to house your reports and source documents. A structure is required to enable finding items easily, as you will soon collect and produce a lot of material. Your documents should use some form of naming standard that includes the date. Make the reports accessible to others – although you may need a special subdirectory to house anything that is sensitive/secret.

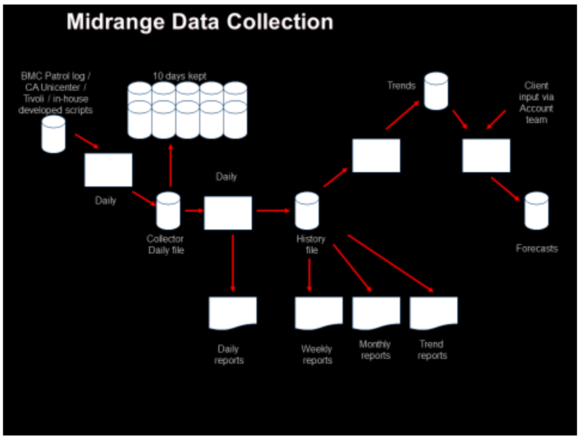
Phase 1 – Collecting System Metrics

Purpose

To systematically collect and store data so that they are available for analysis, model building and reporting.

Metrics

Collect the system metrics from the various devices/components in the environment from the source devices and house the metrics in an accessible repository or set of repositories. This repository may consist of multiple repositories, one for each technology, as the different key metrics are often difficult to structure into one database. The repository is sometimes called the performance database (PDB)[1]. The collection should include all data, such as the defined intervals (e.g. 5 mins), selected data or summarised data. They should include server, storage, network and other devices.



The various performance management stakeholders[2] (event monitoring, server administrators, database administrators, network staff, performance staff, capacity planning staff and applications staff) need to determine the key metrics to collect for reporting. These include counts, capacity, utilisations, throughput, queue counts, queue times and response times.

Ongoing monitoring

The event management process monitors the environment and undertakes some initial investigations and where appropriate raises an incident record. The data collection agents send data to both the event system and to the PDB. The performance management role (performance management staff or systems administrators) should have access to and use the event monitors regularly to see the events in the same manner as the event staff to help answer their questions.

[1] Performance database (PDB) for housing performance data (profiles and short-term trends) and a Capacity database (CDB) for housing capacity data (history and trend).

[2] these stakeholders differ from the stakeholders mentioned in Phase 0

Thresholds

Thresholds for the various aspects of capacity management should be defined with agreement from various stakeholders. The threshold setting process requires additional work to determine and refine the thresholds. Set them so that major issues are identified but do not overload the event monitoring staff (by having the staff investigate too many non-issues). An alert overload situation reaches a point where staff ignore alerts.

Set the thresholds differently for event, performance and capacity as the groups operate with different interval sizes – the shorter the interval the higher the threshold level.

Phase 2 – Workload Characterisation

Purpose

To build an understanding of the IT environment and how it reacts to various changes.

Organisational standards

Understand the environment and the naming standards. Enforcing the rules also helps. The name allows you to identify who owns it or uses it, or where it is located, making the task of summarising data much easier. Also, any chargeback system relies on standards to simplify its processing.

Workload characterisation

Group the work executed on the system into various views for analysis and reporting, using a systemic approach. These views can include application, location, size (small, medium and large) or some other grouping that makes sense.

Analysis

A key area of performance management is the monitoring of daily profiles and short term trends (versus event monitoring). The major reason for keeping figures is to assist in your understanding of the system and how it behaves under various loads. Such knowledge improves over a period of time from observing the values changing over time. Perform some preliminary data analysis to compute ratios, thresholds (where appropriate), abnormal behaviour and recurring or unusual patterns.

Rules of thumb

Rules of thumb (ROT) can simplify some of the work. You should create your own rules of thumb from an analysis of the system. ROTs analyse a limited time period. Circumstances change: new technology is introduced, the business changes, or the application or software is changed and what applied last month may now be incorrect. Build a system to regularly report on ROTs or review the ROTs at least every twelve months. You also need to think about any underlying assumptions.

Phase 3 – Regular Performance Reports

Purpose

To monitor and report on system behaviour to identify variations to investigate and understand. The audience is the internal IT staff – including capacity planners/analysts.

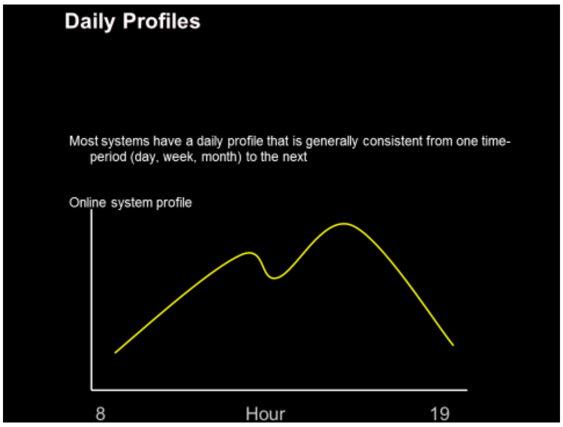
Scheduled reports

There are two types of reports – internal and external. Internal reports are required to support the staff in the performance of their roles – the reports should direct the focus onto areas that need investigation. External reports communicate with the stakeholders – successes, actions (approvals, funding, etc.)

Internal reports

Create a set of scheduled reports that identify issues and provide the necessary information for the review of the environment.

Produce regular reports for resource usage, device response time exceptions, etc. These reports can use the location codes from the standards manual, to group different items together for reporting purposes (creating different views of the data).



Daily Profiles

Most systems have a daily profile that is generally consistent from one time-period (day, week, month) to the next

Produce performance reports daily, weekly and monthly – and run the reports on demand to aid in investigating current issues[3]. The reports should show profiles and short term trends. Review and compare the profiles to prior data (how does yesterday compare to last week, one month ago or three months ago?).

Exception based key reports should list only items that trigger investigation activities, so create reports that only list the items to investigate.

Producing reports is a useful check to see that the data collection worked. Use of the data is the only way that you can guarantee that it is both captured and accurate. Determine and rectify the cause of any deficiency in the collected data.

External reports

Produce monthly or quarterly reports for IT management that discuss transaction volumes (preferably linked to business measures), successes – how many issues avoided, including any approvals required for workload moves, new equipment, etc.

[3] This may require regular data transfers and storage

Many managers these days prefer some form of dashboard / traffic light (Red Amber Green – RAG) summary of the situation. A simple table with commentary (within cell or in a separate comments column) is often sufficient.

Example:

Legend – use a legend to define the meaning of each coloured cell

R (red) – client/business unit action (decision to fund/approve)

A (amber) – vendor/IT action (further investigation required – cause / options / tune etc.)

G (green) – No action required – all good

Phase 4 – Tuning

Purpose

To better utilise existing equipment; for instance, balancing work across the environment to delay the need for additional equipment.

Capacity plan assumption

The capacity plan is usually built on the assumption that the system is tuned and that the tuning will continue, although the capacity planner can make some adjustments and allowances in their models if it is not tuned to an appropriate level. A lack of resources (or funds for upgrades) usually focuses the minds of managers and technicians on how to resolve the issue and ensures tuning activities occur.

Performance management

Performance management should work within the constraints of the equipment provided (as defined in the capacity plan and organisational budget) the assumption is that sufficient equipment is available. If this assumption is wrong/inaccurate – and the equipment is not sufficient –the performance team should communicate the issue to both the capacity planning team and IT management to address the issue.

Tuning only occurs until such time as the agreed service (or service level objective) is achieved and guaranteed for the short term (usually a few months).

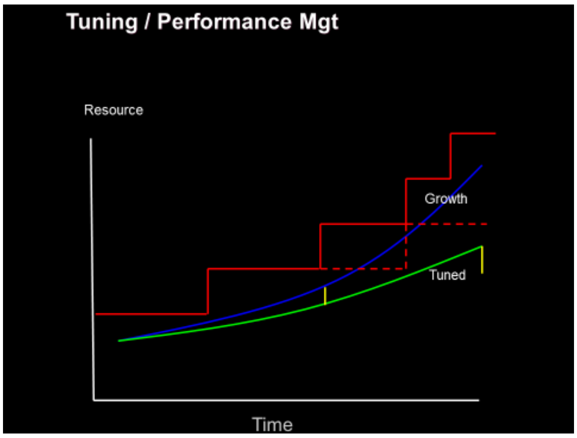
Disk space management

Managing the organisation’s disk space is an important and critical function. These days the largest cost of running a data centre is the disk storage costs and the fastest-growing cost is the ‘environmental’ costs (power and air-conditioning/cooling).

Two areas to consider are (1) performance and (2) space. The objective of space management is to make efficient use of the space while not impacting the service provided to the business. With most organisations using a small percentage of their disk space (often 40–60%), plenty of scope exists to save money.

Some file types (e.g. databases) require some free space within the file. The amount of free space assigned may be inappropriate. Sufficient space is required to allow for planned reorganisations during long weekend periods (Easter, Christmas etc.) with a safety margin.

Also, take the amount of data behind a disk controller and the input/output per second (IOPS) rates into account to ensure that the disk controller is not overloaded. Monitor these IOPS metrics and ensure the vendor recommendations are followed until sufficient knowledge and skills are developed in-house.



System tuning

New equipment has the capability to self-manage such as balance workloads, but it requires the appropriate settings for the monitored parameters and regular reviews.

Often, plenty of capacity is available after an upgrade and there is no urgency to tune. As the work on the system grows, the increase in work can rapidly consume the spare capacity. Without tuning, the system will require an upgrade before it is really needed. A self-perpetuating cycle of upgrade – fills up – upgrade – fills up – upgrade is all too common.

Application system tuning

Many opportunities to tune the application systems exist but often other priorities intervene.

The reasons are, firstly, that the programming staff who design and write the applications may lack the training or experience to create efficient applications. Secondly, other priorities – rolling out new applications or changes to satisfy the business need to gain or keep market share. Thirdly, an assumption that the cost of hardware (servers [CPU and memory] and storage devices) continue to fall making hardware an inexpensive aspect of running an IT department. Whilst this is true to some extent, note the following points:

1. The unit costs of hardware continue to fall but the costs of running the operation fall at a much slower rate as software, staff and facilities costs, either increase over time or fall at a slower rate than hardware (Moore’s law).
2. The benefits gained from the reduction in cost can only occur if the applications are efficient. If the hardware costs fall by 50% in a time period and the new applications are twice as resource intensive[4] as the original, higher running costs will result.

[4] New applications often contain more functionality, increasing the resources required per business transaction.

Phase 5 – Collecting Capacity Metrics

Purpose

To collect and maintain sufficient data to build trend forecasts and to understand the changing environment and the drivers of that change.

Extraction

Extract the data from the performance database (PDB) and summarise them into meaningful capacity metrics. Summarise the performance data into hourly intervals (e.g. 12 x 5 mins) and extract the peak hours and store these in the capacity database (CDB) to become the capacity data for future trending. The data should cover servers, storage, networks and other devices.

The first capacity plan (forecast trends) produced might be built using less than a year's worth of data, if there is an urgent need – usually the reason why the capacity management function is being created is to justify the budget for a new critical project for the organisation. In such a case, a few months of data can be collected and processed in a few weeks, but only if the raw data are stored on some medium in an accessible format. Statisticians recommend a history twice as big as the forecast period (e.g. two years of history to forecast out one year).

Monitoring

The capacity planning staff should review the performance reports (daily, weekly and monthly) regularly. They do not need to access the online monitors (performance management staff do – see earlier). Situate the capacity planning and performance management teams next to each other so that they can discuss any relevant issues. Alternatively, the teams should hold regular information-sharing meetings where observations can be discussed and they can agree on any actions required.

Metrics

The various stakeholders (event monitoring, server administrators, database administrators, network staff, performance staff, capacity planning staff and applications staff) need to determine the key metrics to collect for reporting. These metrics should include counts, capacity measures, utilisations, throughputs and response times.

Thresholds

The thresholds used for the capacity reports should be consistent with the event and performance management thresholds, but take into account the different thresholds due to different interval sizes, e.g. 5 min = 60%, 15 min = 75%, 1 hour = 90%.

Phase 6 – Build Models

Purpose

To create a prediction mechanism (the model) forming the basis of the annual capacity plan (or ad hoc assessments).

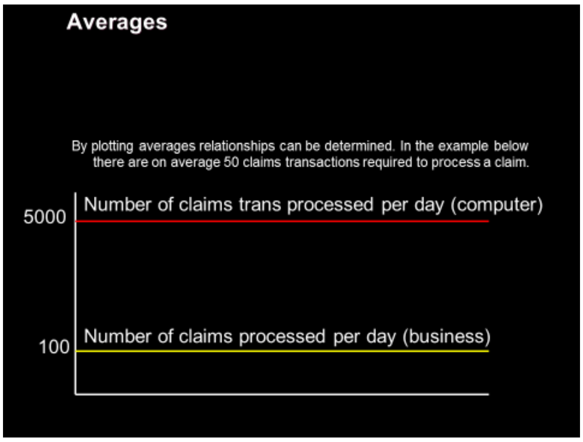
Understand the environment

Some understanding of how the environment functions is required when building a model of it, because a model is a representation of some key aspects of the environment. Prioritise the modelling based on the questions that require answers. Undertake some preliminary data analysis – ratios/ROTs, thresholds, abnormal, patterns. This work builds on the outcomes of the previous steps.

Projection/regression (build models)

The use of natural forecasting units (NFUs) or key volume indicators (KVIs) is one way of forecasting demand for resources. Retain historical statistics for both business and computer resources and analyse the relationships.

Using business/service metrics as the basis of forecasting is in fact a more sophisticated version of the rule of thumb (ROT) approach.



Averages

By plotting averages relationships can be determined. In the example below there are on average 50 claims transactions required to process a claim.

Calibrate/validate/model building

Validate the model to increase the likelihood that it will give satisfactory estimates. Use prior period data to validate the model – use the first few months of the data to forecast more recent actual values and compare and investigate any significant differences.

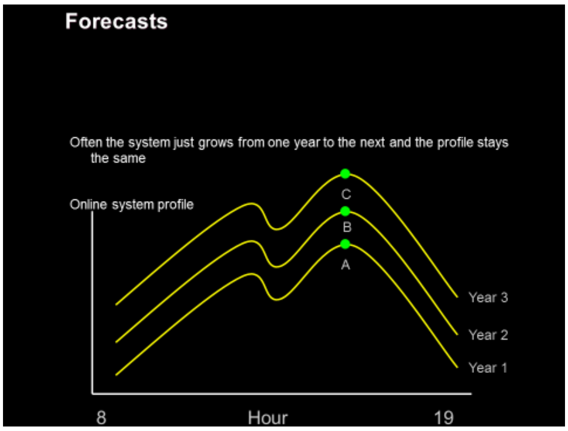
Phase 7 – Forecast Workloads (Part 1)

Purpose

To build forecasts for each key workload so that the most accurate model (within the designated timeframe) can be built.

Workloads

The workloads form the basis of the load on the equipment in the environment. Each workload has its own profile and grows at its own rate. Forecast each workload out for the period modelled. That period is the time horizon of the organisation’s capital budget. The intention is that the workload group forecasts map to the workload groups that drive the resource usage in the capacity planning models. A disconnect between these two activities makes the capacity modelling difficult, if not impossible.



Forecasts

Often the system just grows from one year to the next and the profile stays the same

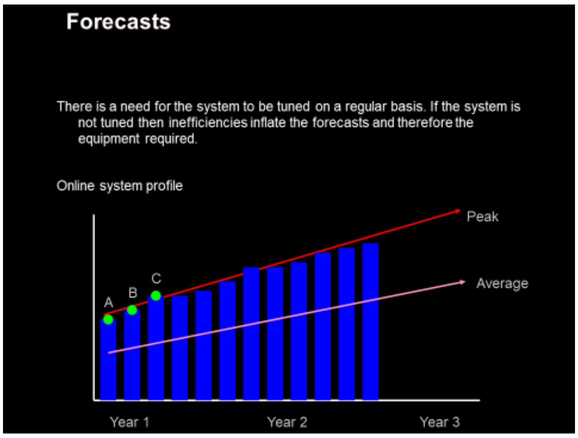
Ease of workload forecasting is related to the architectural design of the environment and how the various applications utilise the underlying resources. Better separation in the use of the equipment leads to better measurements, better models and reports that lead to better management decisions.

The demand management process or function should perform much of this work. Capacity planning generates the forecasts to provide charts and tables of historical values for the various workloads to demand management. These data enable demand management to provide historical trends to the business to assist their projections/estimates. Demand management can also forecast the workload and business volumes – business transaction rates, customers, accounts, etc.

Demand management should interview the business about future plans and projects; and obtain reports of business volumes. Map these measures to the workloads as agreed with the performance and capacity teams. Demand, performance and capacity teams need to agree on how the mapping is to occur, its frequency and who performs the work.

Projection/regression (build models)

Producing a capacity plan requires some workload data for processing. The PDB (Performance Database) feeding the CDB (Capacity Database) and into forecasts is the preferred approach. The workload data are projected out into the future to the limit of the capacity plan (one year, two years, or five years, or in alignment with the capital budget time horizon). The forecasts for the first year or two often use some form of regression (linear, quadratic etc.).



Forecasts

There is a need for the system to be tuned on a regular basis. If the system is not tuned then inefficiencies inflate the forecasts and therefore the equipment required.

Demand management and business unit forecasting can be easier if the capacity planning system provides the history of the resource usage (and other relevant collected metrics) to assist the business to make better forecasts.

Phase 8 – Regular Capacity Reports

Purpose

To communicate recommendations to the key decision makers and other stakeholders.

Reports for stakeholders

The capacity planner should produce reports for the stakeholders. The reports should show the growth in peak hours over time and project out into the future – it should match the time horizon of the organisation's capital budget.

Standard and ad hoc reports

The capacity process should produce regular reports and graphs to monitor the key measures, but the capacity team/process should also do some reporting on an ad-hoc basis.

Show the resource usage of each workload in the reports. The capacity staff need to produce reports about resource capacity and utilisation, identifying how much time is left until the next upgrade is required.

The reporting frequency depends on the business policy – spending/upgrade frequency, business change frequency, etc.

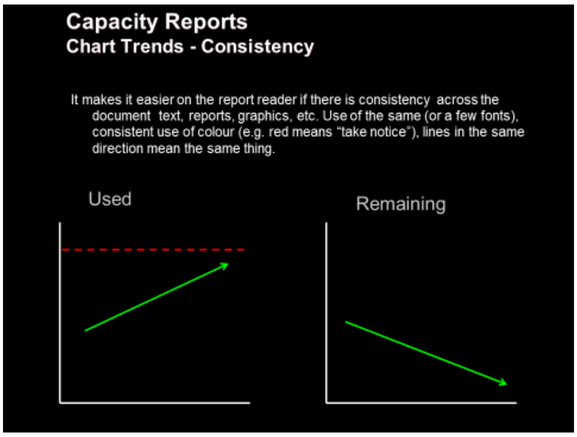
Scheduled reports

There are two types of reports – internal and external. Internal reports are required to support the staff in undertaking their roles – the reports should highlight issues. External reports communicate with the stakeholders – successes, actions (approvals, funding, etc.)

Internal reports

Create a set of scheduled reports that provide the charts for assessments and plans.

Produce the reports monthly – but be able to run the reports on demand to support the capacity assessment process. Producing reports is a useful check to see that the data transfer from the PDB was successful.



Capacity Reports
Chart Trends - Consistency

It makes it easier on the report reader if there is consistency across the document text, reports, graphics, etc. Use of the same (or a few fonts), consistent use of colour (e.g. red means “take notice”), lines in the same direction mean the same thing.

External reports

Production of the monthly or quarterly reports for IT management that discuss transaction volumes (preferably linked to business measures), successes – how many issues avoided, including any approvals required for workload moves, new equipment, etc.

These reports can also use the

Many managers these days prefer some form of dashboard / traffic light (Red Amber Green – RAG) summary discussed in Phase 3 – Regular Performance Reports.

The monthly meetings to discuss the reports are an opportunity to cover a range of topics:

- From capacity planning to managers, questions about organisational changes and their implications
- from managers to capacity planning, questions about their view on various scenarios, business plans and options.

Tracking (actuals versus predicted)

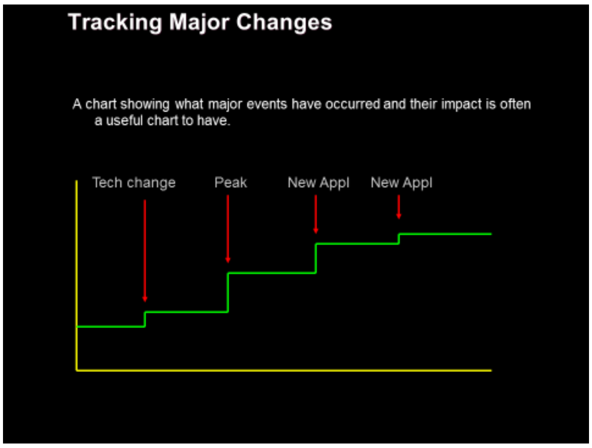
The tracking activity is not executed for the first capacity plan, but it should once the first capacity plan is produced.

The capacity plan should contain a section that compares the projection from the last capacity plan to the current one and the actual values all on the one graph. The comparison is used to show the various stakeholders how accurate the predictions are, or to explain the reason for the differences to maintain credibility.

A monthly version of this tracking (Actuals versus Predicted) report can be produced to monitor the accuracy of the last capacity plan rather than waiting until next year’s capacity plan.

Produce these reports from the regular monthly reporting system and incorporate them into the annual capacity plan section that tracks the actual monthly values against the forecasts. Review the reports each month to assess whether the capacity plan forecasts continue to remain accurate.

The three areas to cover are (1) resources, (2) services and (3) business (including transaction volumes, utilisations, queue lengths and response times for each area). Use these comparisons to explain why any differences exist. Is the number of customers below or above expected? Is the server CPU utilisation higher or lower than expected? Are the estimated response times above or below the predictions? Will the next upgrade be sooner or later than predicted in the capacity plan?



Tracking Major Changes

A chart showing what major events have occurred and their impact is often a useful chart to have.

Such a report can also can indicate drivers of change in the environment and items to be investigated and if required, corrected.

Also, it may occasionally be useful to produce a report comparing all of the previous forecasts to visualise any improvement in the forecasting (a narrowing of the forecast lines for best-worst forecasts – or a more aligned trend line for single forecast projections).

Phase 9 – Survey/Interview

Purpose

To collect business and other inputs to improve the accuracy of the capacity forecast model.

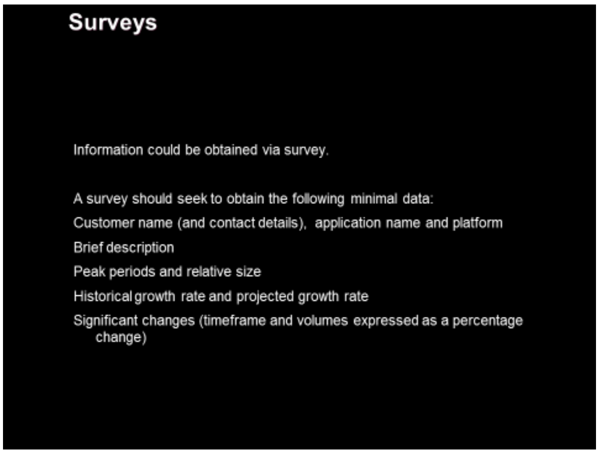
Business units / demand

The demand function should engage with the business, otherwise the capacity planner must talk to the business unit managers (or their delegates). Use business terms and not IT terms. The capacity planner must interpret the discussion and later convert any business data into workload information.

Survey/interviews (IT)

One of the roles of the capacity manager and demand manager perform is as translator between the business and IT.

Obtain any required information through interviews or surveys. Modify the projections made from the historical data using input from the interviews with the IT staff such as the application teams as well as the testing team. If the testing team regularly tests the applications and have been for some years then they can have some useful insights into the applications.



Surveys

Information could be obtained via survey.

A survey should seek to obtain the following minimal data:
Customer name (and contact details), application name and platform
Brief description
Peak periods and relative size
Historical growth rate and projected growth rate
Significant changes (timeframe and volumes expressed as a percentage change)

After the first rounds of surveys, it may be useful to bring historical records to interviews. These would show how previous forecasts compared to actual changes. This information can help improve forecasts, though some managers may find this approach confronting.

Surveys can cover more departments/areas/project teams and reduce the amount of time required to gather data but a lot of design effort is required. Some science exists for good survey design. If significant effort is not put in to the survey design the data gathered may not align and be of limited value.

The most useful interviews are often conducted at the interviewee’s desk in an enclosed office. Conduct the interview in a meeting room where an open and honest discussion can occur. Open plan offices make this difficult as conversions are not private. At least meet at their desk first, because little snippets/pieces of useful information are often obtained. Often you will find a graph on the wall in someone's office or cubicle that is useful to your work and unless you visit their desk or office you may never know of its existence.

Phase 7 – Forecast Workloads (Part 2)

Purpose

To apply the inputs from the business (via demand management) to the trend forecasts to create a more accurate model.

Workload forecasts

After completing the interviews, assess the information and adjust the workload projections to allow for major projects. Assume that the information is credible because it can only be confirmed through use and time. The growth curve already incorporates the additional small systems in its history; as it is history extrapolated. The capacity planner has to decide if a project is significant enough to affect the trend line.

The next step is to make adjustments to these base historical trends based on the results of the surveys/interviews for any application, business or external change significant enough to affect the projection. New applications often cause significant adjustments in the forecast, but other reasons may exist, e.g. a significant marketing effort, or a change to a high-use program in an online environment.

Other sources of input

The annual report to shareholders/stakeholders, internal memorandums, minutes of meetings and business magazines are other sources.

Phase 10 – Modelling

Purpose

To provide the basis for the capacity plan (or capacity assessment).

Technology/vendor knowledge

When modelling, knowledge of available (and near future) technology is important. Focus on the various vendors that your organisation deals with on a regular basis and read their announcements. Monitor new technologies and other vendors because solutions to your organisation's problems might be found there.

Refresh, leases and software licences

A refresh should trigger the evaluation of alternatives in the capacity plan. Also, leased equipment requires an assessment of the available options in preparation for when the lease ends. Preparation is preferable to being caught unprepared.

When reviewing workload configurations in an organisation with several large servers, awareness of the licenced software is required because expensive software is often not licenced for all servers.

Other inputs

Other information and questions may arise from the interaction and engagement with the other processes, areas and functions, such as:

- Availability management
- IT service continuity management (ITSCM)
- policy and strategy
- configurations
- applications
- architecture teams
- Standard Performance Evaluation Corporation (SPEC) or other relative ratings.

Service levels – queuing models

Another input is the organisation's service levels. Use SLAs as one of the inputs into the computer system queuing theory model (if one is used) to describe the service target. Use the assumed service levels when no service levels exist. The assumed service level should reflect the complaint threshold – in other words, when do the key business unit managers escalate to the operations manager/CIO (by-passing the service desk) to complain or demand action to fix an issue affecting their staff's productivity?

Queuing theory models require forecasts of inputs such as:

- transaction volumes, users, jobs
- resource usage growth for each workload unit (transactions, users, jobs)

Projection/regression (build trend models)

The equipment required is determined using some form of model. The model may be as simple as using the historical resource usage trends or it can include adjustments for planned projects and changes in the environment. This method can be satisfactory, but it ignores the effect of queuing on the response time.

Do not forget that capacity planning is not a technical function; it is a management function[5], where close enough is often good enough. Perform computations to the precision required to make a decision.

Modelling/options/workload configuration

Models to assess the planned demand and related resources required to support it are required. This requires the changes and additions in the capacity plan to be documented before providing a feed into the financial capital budget process. Often the daily profiles for the moving and target server workloads differ. More detailed analysis may be required. The impact is not a matter of adding the two peaks together, but the entire profiles of both. Writing a program to select the relevant data to create a graph or table of numbers showing the impact is relatively easy.

Sensitivity analysis

Making adjustments to the model to assess the effects of uncertainties is called sensitivity analysis. Identify when the resource breaks (number of users or number of transactions) using sensitivity analysis. It can be critical/important when, for instance, the organisation is trying to extend the use of the equipment for a few more months.

What ifs

The more equipment (or sites) the organisation has, the more options it has available to move and balance the work.

Phase 11 – Capacity Plan

Purpose

To recommend the amount of funds to include in the organisation’s capital budget and when they will be needed, so that the required equipment purchases are funded – also to justify the budget figures.

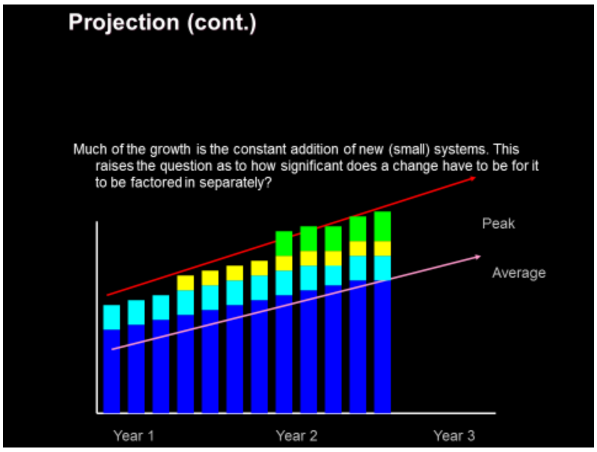
Major output/product

The key deliverable of the capacity planning process is a capacity plan. The production of the capacity plan is an annual cyclical process.

[5] Performance management is more technical and requires more precision than capacity planning

Equipment required

Document the equipment required in the form of a report (the capacity plan). It should include discussion and analysis (present insights) of the current topics that are attracting management's attention. Maintain regular contact with the key influencers in your organisation[6] to keep track of the current key topics. The capital budget should include the difference between the current and required equipment.



Projection (cont.)

Much of the growth is the constant addition of new (small) systems. This raises the question as to how significant does a change have to be for it to be factored in separately?

For capacity planning purposes the actual design is not important as long as the figure provided for the capital budget covers the actual costs of the modelled solution. The solution as described in the capacity plan is hopefully close enough to the eventual solution (or at least similar) but to achieve it means closer engagement with the architects and engineers to test ideas.

Actual detailed planning and design work occurs when the expenditure is approved, often some time later, by the organisation (Board, CIO, Assistant General Manager (AGM) etc.). This is when it is determined whether the capacity manager has performed their role by the accuracy of the budget figures. Additional funds will be required if the budget estimate is well below the real cost – usually from the IT budget – and the blame will fall on the capacity manager for the error/failure.

Justification

The mapping of workloads to their respective resource usage growth allows the capacity plan to indicate the impact of projects on the planned expenditure. Include such information in the capacity plan (and presentation).

Report creation/delivery (presentation)

The executive summary section of the capacity plan should contain a short summary (preferably in table format) that can be incorporated into the organisation's capital budget process.

The capacity plan should convince its readers of the need for a decision or action (purchase equipment, consolidate workloads, remove equipment, etc.) – or non-action if no additional equipment is required (a rare situation).

Present the results to the IT senior management team, to explain the need for major equipment upgrades and why it is so. Such background information assists the IT managers in their discussions with the business and finance.

The IT senior management team may focus on either the technical aspects (resource utilisations etc.) or the business aspects (customers, business transactions etc.), depending on their background and aspirations. The report

[6] In fact, if you follow the approach, you should become a key influencer yourself

and presentation should provide the information in the manner/style the stakeholders prefer.

The presentation gives the invited audience the opportunity to ask questions to better understand the various sections of the capacity plan, as well as the rationale and reasoning that forms the basis of the recommendations. Note: Identify from the feedback the concerns of the various individuals and identify where the report can be improved. Do not take feedback personally – take notes and use them to improve the report for next time (or update and issue a complete revision if the report is a disaster).

The capacity plan presentations are also an opportunity to discuss a range of topics:

- From capacity planning to the managers: questions about organisational changes and their implications
- from managers to capacity planning: questions about their view on various scenarios, business plans and options.

Phase 12 – Capacity Assessments

Purpose

To assess the implications of changes in the environment – either new resources, new applications or changes in business processes.

Capacity mini-plans

The various types of capacity assessments are capacity mini-plans with a focus on either technology, a single business area/decision or an application rather than the whole environment.

New project cost justifications (CP)

As organisations focus management attention on costs, anticipate that the capacity management function will become involved in the cost justification of new projects. If the capacity planner discovers that the current installed technology cannot support the requirements of the application the capacity planner has to inform the relevant parties of the issue.

Application reviews/performance assessments (PM)

Estimating resource requirements for new applications is difficult because estimates for both the transaction volumes and resources per transaction lack the desired precision.

Phase 13 – Subsequent Processes

Purpose

To receive the capacity plan and to prepare and implement it.

Additional processes

The process activities described in Phases 0 through 12 are part of the capacity management process domain or part of demand management. The steps or processes described here (Phase 13) are outside the capacity process[7]. In large organisations, these other processes are the responsibility of other owners. Initiating spending on an acquisition would only occur if the capacity manager was pre-authorised to spend money on upgrades; then, these activities would form part of the capacity management processes. But the authority is rarely delegated to the capacity planner in large organisations – the CFO is often responsible for the budget process including obtaining approvals, especially for significant expenditure. The CFO also often oversees the authorisation to spend (draw funds from the capital budget allocations) and may also manage the supply chain management (purchasing / acquisition) team.

Capital budget

Even though the capacity plan process is performed annually it is an ongoing cyclical process that should feed into the organisation's capital budget process (often annually, occasionally quarterly). Reporting the expected expenses by quarter or month allows the organisation's financial people to manage the organisational money better – providing the knowledge of how much and for how long the finance team can invest the spare funds in the short term overnight funds money market and for how long.

Architecture design

Design drives cost and that can be outside the scope and authority of capacity management. While customer requirements drive application design, application design drives the cost of delivery.

The architecture team/s in organisations act as the co-ordinators of various pieces of information (finance, technical, capacity, availability etc.) and use the information to arrive at a workable solution.

Order equipment

When the need to acquire equipment is triggered the supply chain management or a similar group is engaged to place orders with the suppliers or select the winning proposal and conduct price negotiations.

Configuration planning

Often an engineering team implements the solution designed by the architects, usually after validating the design. The design may even require modifications to make the solution work.

Implementation

Implementation is often assigned to a project manager assigned from a pool of project managers. The project manager engages with key stakeholders and develops a schedule to deliver the project on time and within budget.

Facilities/property/environmental upgrades

Newer technology equipment (servers and peripherals) are heavier and require more cooling, restricting where the equipment can be located, including placement within the building. Another department such as property or facilities may be responsible, rather than IT

[7] ITIL considers these to be part of the capacity management process

Post-implementation review

Conduct a post-implementation review (PIR) after a project is completed and identify anything requiring improvement (bad) or to repeat (good) in future projects. Involve all areas and review their areas of responsibility to identify process improvements. The performance and capacity functions should review the change from their point of view. The results of the various assessments (performance, demand and capacity) should feed into any PIRs.

Conclusion

Implementation of capacity management can occur in several different ways.

When implementing capacity management, start with a skeleton set of activities that address the existing issues and build on them, adding new steps and refining the existing tasks at each step/activity in each of the processes.

The building of a functioning capacity management process/function operating with a strategic view rather than an operational view is key to adding value to the organisation.

Create a set of processes for your capacity management activities that interconnect and flow to improve productivity as well as providing timely guidance to your organisation about its IT needs.

Philip Bailey

Philip Bailey is an expert in the field of IT capacity management. Phil has implemented, guided and improved the capacity management processes in several organisations (financial and outsourcing). In addition to 25 years of experience, he brings first-rate conceptual / process skills, a creative approach and proven leadership to the role as consultant.

Phil has established a capacity planning methodology at various organisations. He has also trained several capacity practitioners over the years. He is an ITIL V3 Expert and ISO 20000 Consultant certified. He has written articles and spoken extensively about the methodology he developed relating to the IT capacity management process and its underlying activities.

Phil has a habit for acquiring and reading non-fiction books.

CapProcess

Processes that work!

CapProcess is a company that provides expert advice in the capacity management process domain. Expertise that includes how to improve capacity management activities so that IT service improves and costs are effectively managed.

Our purpose is to transfer our years of experience to your staff are more self-sufficient and provide the support that your customers expect.

CapProcess is based in Sydney, Australia.

To find out more about how CapProcess expertise can help your organisation contact us on +61 4 13 487 701 or send an email to info@capprocess.com or visit our website at www.capprocess.com