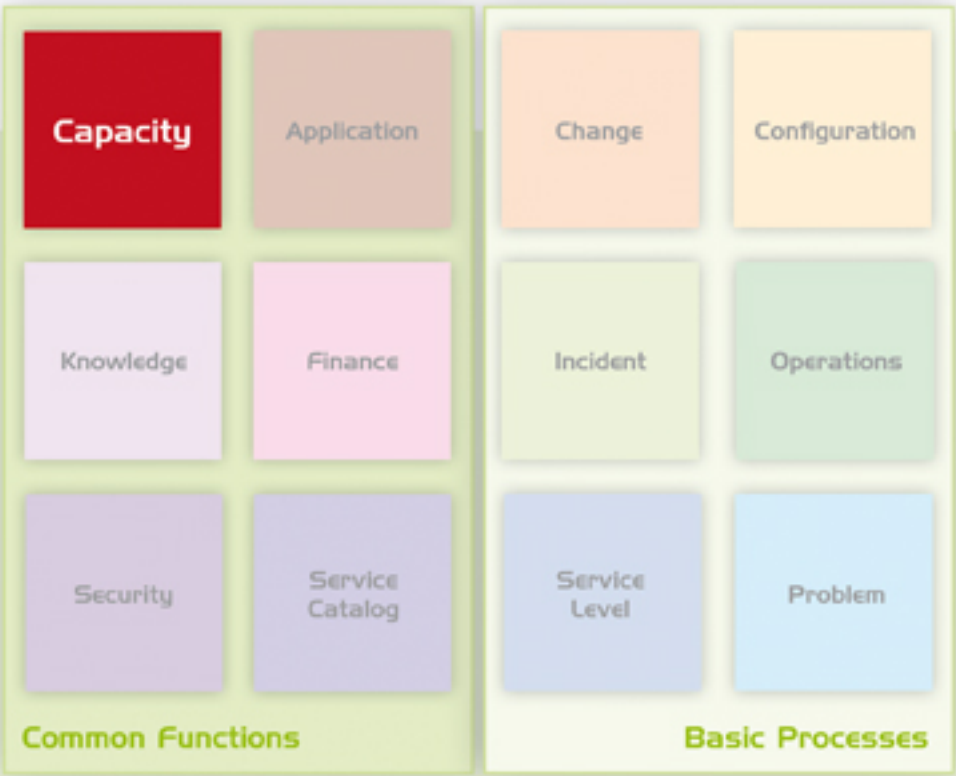
ITSM LIBRARY

Capacity Management

A Practitioner Guide



Other publications by Van Haren Publishing

Van Haren Publishing (VHP) specializes in titles on Best Practices, methods and standards within four domains:

- IT management,
- Architecture (Enterprise and IT),
- Business management and
- Project management

These publications are grouped in series: *ITSM Library, Best Practice and IT Management Topics*. VHP is also publisher on behalf of leading companies and institutions: The Open Group, IPMA-NL, PMI-NL, CA, Getronics, Pink Elephant.

Topics are (per domain):

IT (Service) Management / IT Governance
ASL
BiSL
CATS
CMMI
COBIT
ISO 17799
ISO 27001
ISO/IEC 20000
ISPL
IT Service CMM
ITIL® V2
ITIL® V3
ITSM
MOF
MSF

Architecture (Enterprise and IT)
Archimate®
GEA®
TOGAF™

Business Management
EFQM
ISA-95
ISO 9000
ISO 9001:2000
SixSigma
SOX
SqEME®

Project/Programme/ Risk Management
A4-Projectmanagement
ICB / NCB
MINCE®
M_o_R®
MSP
PMBoK®
PRINCE2®

For the latest information on VHP publications, visit our website: www.vanharen.net.

Capacity Management

A Practitioner Guide

Van Haren PUBLISHING

Colophon

Title:	Capacity Management - A Practitioner Guide
Series:	ITSM Library
Author:	Adam Grummitt (Metron, UK)
Publisher:	Van Haren Publishing, Zaltbommel, www.vanharen.net
Editors:	Jan van Bon (Inform-IT, chief editor) Annelies van der Veen (Inform-IT, editor)
ISBN:	978 90 8753 519 3
Copyright:	© Van Haren Publishing 2009
Edition:	First edition, first impression, July 2009
Design & layout	CO2 Premedia bv, Amersfoort-NL

For any further enquiries about Van Haren Publishing, please send an e-mail to:
info@vanharen.net

Foreword

When I was asked to write the Foreword for this book I was honored. I will attempt to explain why.

Adam Grummitt is a Gentleman and a Scholar, and an unsung hero in the field of capacity management in the wider IT service management space. Of course, to those who specialize in capacity management, he has long been a recognized expert (guru, luminary—select your favorite word) and someone whose advice and guidance many of us have been very fortunate to benefit from.

Together with Hans Dithmar and Brian King, the main contributors to the first ITIL 'module' on capacity management, the Metron team and Adam were (and are) the 'go to' subject matter experts. For over twenty years (to my knowledge) Adam has been quietly involved in capacity management, as practitioner, speaker (at various Computer Measurement Group (CMG), itSMF, Masterclasses and other events) and the company he helped to found (Metron) has specialized in capacity management.
His expertise, however, is much wider than capacity management, as you will soon realize when reading this book. Adam has finally done what he has promised: he and the supporting team of editors and reviewers have written the definitive book on capacity management.

Those familiar with the wider issues of capacity management will not find surprises (though they will definitely learn from the expert!), but the wider community will soon discover many important concepts and ideas that aren't available elsewhere that will make a real impact on their day to day operations.

In my view, capacity management is the most important of the ITIL functions and processes; this explains why the original module was well over 200 pages and was often described as a brick! This book is a fine and worthy successor – Adam and the many contributors and reviewers provide the detail and depth needed for practitioners to fully understand the topic.

This book ensures that some of the vast knowledge and experience from several of the world's leading experts is now available for all of us to use for a very long time.

And this is why I am honored to write this Foreword; one of my own personal heroes has, with his colleagues and reviewers, produced a work that will deservedly have its place on the bookshelf of any serious ITSM expert.

Brian Johnson

Author of ITIL V1 Capacity Management, co-founder of itSMF, Honorary Life Vice-President itSMF.

Acknowledgements

We like to thank the team of experts that have been involved in the development of this publication. Above all we acknowledge their never-ending enthusiasm and commitment.

First of all we wish to thank author Adam Grummit for gathering best practices on IT capacity management, using his own huge knowledge and experience, existing literature and information from peers. We also thank him for seriously considering all 700 issues of the reviewers and for his persistence in improving the quality of the manuscript. This has enabled us to develop a true best practice on IT capacity management.

We also wish to thank our international review team of experts who have contributed their huge experience and knowledge. They provided encouragement, criticism and useful new ideas, to ensure that the book reflects the very best practice. Their expert help has been invaluable.

The review team:

- Nick Bakker, Capacity Management expert at Getronics, NL
- Frank Bereznay, Capacity Planning Manager at Kaiser Permanente, USA
- Gabriele Biondo, Trust In People – DHL, NL
- Edouard Boris, Director of Service Engineering Europe, Yahoo, France
- Stephane Duperrex, Senior Associate – Capital Group Corporate International, Switzerland
- Irwin Friedman, independent Capacity Management consultant, USA
- Prof. Javier Garcia Arcal, Universidad Antonio de Nebrija, Manager of ITIL Best Practices at Oesia Networks, Spain
- Mike Hogg, ITIL Business Capacity Manager, NHS Account, CSC UK
- Kevin Holland, Head of Service Quality Improvement at NHS Connecting for Health, UK
- Neil Jabri, Availability, Capacity & IT Continuity Manager at Nestlé, Switzerland
- Ron Kaminski, Director of Capacity Planning & Performance at Harrah's Entertainment Inc., USA
- Tomoyuki Kawano , Managing Director of IIM, Japan
- Mike Ley, UK Computer Measurement Group (UKCMG)
- Gert Luyten, Senior manager Global Outsourcing Services at CSC, Belgium
- Tuomas Nurmela, Technology Manager at TietoEnator Processing & Network Oy, Finland
- Tony Oliver, Capacity and Availability Management at RWE IT, UK
- Bipin Paracha, Practice Head at Wipro Consulting, Wipro, India
- Gavin Pomfret, independent Capacity Management expert, UK
- Tony Verlaan, Capacity Management expert at Getronics, NL

We are also extremely grateful to the UKCMG and CMG and members of their boards for their contributions to the quality checking process. UKCMG is the UK chapter of the international Computer Measurement Group (CMG), see www.ukcmg.org.uk and www.cmg.org.

On the author

Adam Grummitt, MA (Cantab), C Eng, MIEE, CITP, MBCS

Adam has been playing with computers since graduating from Cambridge way back. Doing research in mass spectrometry[1] he used the first Digital PDP-8 in the UK and early IBM mainframes. He has since been an analyst, designer and programmer for end-users, software houses and a consultant. He has worked in a variety of computer applications, from scientific engineering and aircraft design to medicine[2], law, technical publishing and retail. He was a founding Director of Metron in 1986, which was an early partner in UKCMG, ITIL and itSMF UK. He is currently chairman of UKCMG and on an executive sub-committee of itSMF UK. He gives papers and workshops on capacity management practice at numerous international IT conferences and is a well-known speaker at many chapters of CMG.

The author thanks all those in Metron, past and present, who have helped in the development of its capacity management material over the years and in reviewing early drafts of this book. In particular, Mike Garth, who first introduced the author (and many others) to formalized capacity management and who demonstrated powerfully just how effective it can be with a top practitioner.

The author would also like to acknowledge the support of all kin and kith over the years; above all his wife and, in the crafting of this book, his daughter's attempts to improve the style and grammar and his his son's efforts to improve the structure and logic (both somewhat in vain!).

1 Interpretation of mass spectographic ionisation efficiency curves by deconvolution methods using Fourier transforms assuming Maxwellian distributions, AEI MSI 902 proceedings 1968

2 Real time record management in general practice, Int. J. Bio-medical computing (8) 1977

Contents

Introduction

This book is intended to provide information complementary to that about capacity management as provided within ITIL V2 and V3. The book can be read either from start to finish or selected sections may be used for reference. In particular, various checklists and templates are provided as a starting point for readers to 'adopt and adapt'.

The intent of this book therefore is to augment the ITIL books by providing more specific detail on capacity management. Thus it assumes the reader has an awareness of the ITIL descriptions and so knows 'what to do' and wants to know a bit more about 'how to do it'.

There are numerous books (and white papers) available on the subject of capacity management. They tend to fall into three classes. There are the mathematically oriented ones that discuss the underlying principles of queuing theory and related statistical techniques. There are the business oriented ones that discuss the philosophy of governance and process management. Finally there are the pragmatic ones with discussions of practical experience with particular domains or the latest new 'hot topic'.

The papers on theory can very quickly develop into a mathematical treatise with equations and algorithms for the few. The papers on process definition can very quickly develop into an unending set of checklists of guidelines and project management that can give middle-management a bad name and lose the interest of the experienced practitioner. The papers on current experience tend to be very site and domain specific and age rapidly.

This book tries to give a reasonable appreciation of capacity management from all three perspectives. It takes an understanding of IT service management (ITSM) processes as a starting point and expands on the specific issues raised when implementing or improving capacity management practices. It introduces elementary arithmetical approaches to analyzing data.

The main objective of this book is to ensure that a realistic, practical and pragmatic approach is adopted, where timescales, effort, priority and money play as much a part as process dataflow diagrams. It is worth remembering that some experts summarize the Information Technology Infrastructure Library (ITIL) simply as 'documented common sense' and ISO/IEC 20000 as 'auditable common sense.' Of course the common sense in question may be based on many years of practical ITSM experience.

How to Use this Book

This book is intended to provide the reader with some general background and a little technical detail as regards the practice of IT capacity management. It is assumed that the reader has an understanding of IT infrastructure and may well have read the relevant parts of the IT Infrastructure Library (ITIL). No matter which version of ITIL (V2 or V3) has been read, the material amounts to some fifty pages of general description of the activities involved in capacity

management. If the reader has read the capacity management requirements identified in the ISO/IEC 20000 standard, they will appreciate that it is a three page summary of the expected deliverables of the practice. This book extends the description of the practice to around 200 pages and goes into a bit more practical depth, more in keeping with the excellent, original 200 page module in ITIL V1[3].

The target audience is anyone involved with capacity management, whether as a practitioner or as a manager or working in related areas and seeking a better understanding of it. It should be recommended reading for those in any ITIL activity as well as developers, testers, customers (ITIL's term for end-user managers paying for the IT service) and service level agreement creators. The book provides general descriptions of all the related activities and deliverables as well as many checklists going into some detail of tasks and data involved. Thus the book could be read from start to end but it is anticipated that for many readers the most value will lie in the checklists in the appendices, using them as draft templates for internal use. It is entirely within the spirit of ITIL, and most authorities on the subject, that the reader chooses at will whether to adopt or adapt any part of this book. Take it or leave it. That was true of the first version of ITIL and is still true today, despite many false prophets debating the precise interpretation of some of the contradictory descriptions within ITIL.

The book is structured into nine chapters.

The first chapter provides an introduction to capacity management practices in the context of IT service management and chapter two expands on its background.

Chapter three reviews the benefits gained by adopting the practice of capacity management.

Chapter four is the longest chapter and outlines the activities, inputs, processing and outputs involved in capacity management.

Chapter five summarizes different perspectives of capacity management.

Chapter six discusses implementation issues.

Chapter seven outlines the management of the capacity management practice and chapter eight its optimization.

Chapter nine considers the requirements and options for related tools.

The remaining chapters are essentially appendices, and contain a lot of useful information and checklists.

Appendix A provides the basic concepts for IT service management, and is the common philosophy for all books in the Practitioner Guide series. It is important that anyone not fully aware of the differences between processes and functions reads this Appendix to avoid conceptual

3 Authored by Brian Johnson

errors in the embedding of capacity management in their organization. ITIL and IT service management are most often related to process-based approaches, and capacity management can follow that approach. capacity management in itself, however, clearly is a function (or 'practice'), an organizational capability, using people, processes and technology to accomplish its goals. This Appendix explains the approach to make that work.

Appendix B lists the acronyms, models, frameworks and standards used and a personal glossary.

Appendix C contains checklists which should prove useful as a starting point for any site to review its own capacity management practice.

Appendix D contains a template for capacity plans and a simple, sample capacity plan.

Appendix E offers an alternative approach and wording of the six basic processes in IT service management.

At the end of the book you'll find a few useful reference sources and an index to keywords and their location within the book.

1 Context for capacity management

This chapter introduces the concepts of capacity management and how they are typically reflected in real life. It introduces some basic concepts and terminology of capacity management to pave the way for the rest of the book.

1.1 Setting the scene

Capacity management is a well-established practice that has been used in information and communications technology[4] (ICT, or IT) for several decades. In most large enterprises today, IT plays an increasingly significant part in the business. The focus is moving towards reducing both capital expenditure and operating costs and make more use of existing assets. Capacity management is a key component in helping organizations to optimize costs and also reduce carbon footprints by consolidation techniques. It can also impact on datacenter management at the level of contracts or even real estate management such as the construction of required new buildings.

IT solutions are implemented on mainframes, distributed systems on UNIX, Linux, Windows or thin clients or combinations thereof. Applications are developed that interact with the users in the business areas to help the enterprise succeed in its business goals. The applications may be developed in-house, off-shore or bought in as 'Commercial Off The Shelf' (COTS) packages. These applications are typically composed of interactive transactions to achieve particular results such as invoices, orders, database updates and so on. Developers tend to think in terms of the software development life cycle (SDLC) and applications with projects to develop them. The SDLC covers requirement definition, analysis, specification, coding, testing and pilot implementation. Once the application is in production, it then tends to be considered as ready for operations to take responsibility for it. It is worth remembering that many in IT are involved with development and testing and not much aware of ITSM or ITIL. In the same way, many involved with ITSM and ITIL are not much aware of the software development life cycle which has its own management philosophy and typically works under the control of projects rather than the implementation of infrastructure processes.

Operations people tend to think in terms of groups of applications used by particular groups of users and refer to them as services. The capacity managers have to ensure that the services run on suitable equipment so that the desired level of service is achieved. This is usually described in terms of availability (such as 7x24 at 99.99% uptime), continuity (such as a proven disaster recovery capability), performance and capacity (so many transactions per peak hour with a certain response time from end-to-end in terms of servers and networks) and maximum capacity (so much headroom for spare capacity to cater for special peaks in traffic). This is moving into the area of capacity management, which is usually taken to include all the activities to ensure the

4 Throughout this book, we use the term IT to indicate information and communication technology (ICT) as well as information technology.

performance of a service and the capacity of the infrastructure to support it. Thus it incorporates performance monitoring and analysis, capacity planning and related liaison with development, testing, service level management, business areas and more.

In essence, the challenge is to find the equipment necessary to meet the ever increasing demands of the business and then to ensure that the inevitable bottlenecks are minimized and the costs of providing the service are also minimized. This requires demand management in the sense of controlling and prioritizing requests as well as monitoring of performance and throughput to ensure an understanding of the relationship between the business application and the resource demands placed on the infrastructure.

All of the IT infrastructure is involved, from the space in the data center to the routers in the network. Over capacity and lightly used servers, especially on X86 platforms, are a focus for improvement to the benefit of both budgets and sustainable IT. Data center facility capacity management is a growing aspect as servers become higher density placing greater demands on air conditioning, as well as moves towards consolidation and green 'sustainability'. In practice, many practitioners will focus on the higher cost servers and the systems that support heterogeneous applications rather than a single workload. Thus a lot of this book will consider the impact of mixed workloads on larger servers. However, most of the practices described can be applied to distributed systems, networks, storage farms, data warehouses, or other IT solutions.

1.1.1 Lean, mean and green capacity management

In lean, tough times of need for competitive advantage, it is important to adopt a focused approach to core activities in all IT practices. In times of mean, frugal economic measures, it is essential to focus on those practices that are effective and yield practical deliverables. In enlightened times of sustainability, it is also an advantage to find solutions that appear to satisfy the criteria for 'greenness' – even if some of the benefits are debatable. In practice, the most pragmatic 'lean mean green IT service management solution' is to promote the same core activities that have been established over the years for effective capacity management.

The 'more' to be done usually means, these days, more applications in more services on more servers for more users of more critical business requirements. This means trying to automate as much analysis and reporting as possible to be applied to increasing numbers of machines, both real and virtual. The 'less' available usually means less available resources on all fronts. This includes all financial budgets, as well as reducing numbers of data centers and their staff, reducing spare capacity and headroom, consolidating servers, virtualizing machines with probably less specialist staff for all the work on both the infrastructure and all the related development projects.

However, the 'more' has to be related to what is actually achieved now (in a business sense) and the 'less' has typically to be assessed in financial terms for any 'overhead' costs involved in the provision of the infrastructure and services. This is also a time of increased outsourcing, off-shoring or otherwise off-loading some of the IT services. This may be done via a managed service provider (MSP) who may offer software as a service (SaaS). Whichever approach is adopted, the traditional needs within capacity management for baseline definition, workload characterization, business driver identification, application sizing, demand management, monitoring, analysis, forecasting and modeling are all involved.

In lean, mean times there is an increased desire to try to make the most of current investment, to identify any spare capacity and assess just how much more traffic can be accommodated without undue loss of service level. Virtualization of Windows servers, instead of merely grouping a number (like ten or twenty) of highly under-utilized servers to a single or mirror pair of larger servers, is moving towards more significant consolidation ratios like twenty or forty to one. With higher utilization levels, contention becomes a dominant consideration and performance degradation for virtual machines has to be assessed in the light of workload priorities, quotas and service level agreements. Also, as more significant services are virtualized, the overhead incurred and performance impact of an extra layer of software can become more evident.

Centralizing, virtualizing and consolidating machines give a company more opportunity to have an effective energy management policy by reusing the heat generated. The machines may occupy less physical space but may require more air conditioning and clear space around them, so the total green saving is debatable. The policy towards write-off or reuse of the old equipment will largely determine the green benefit. In theory, virtualization, consolidation, auto-provisioning, workload management and dynamic workload balancing (such as VMotion) allow companies to turn off machines at low demand periods. Potentially, combined with grid or cloud/sphere computing, these could offer the ultimate in greenness by only using power and machines when you really need them.

The net result is that there may or may not be fewer machines using more or less power. There may be better services on fewer machines, or contention may lead to degradation in the service. But in all cases, there is a need to find the costs and performance benefits of the current and proposed configurations to justify the levels of expenditure planned in the light of business demands. This balancing act is at the heart of capacity management and is as much a requirement in a well-managed IT environment as ever.

In order to optimize on the 'hardware costs versus capacity' and the 'user requirements versus service level' balances effectively, often it is necessary to have a quick technical audit of current capacity management practice (CMP). Many sites have fewer performance analysts and capacity planners than in the past, yet looking after more servers for services that are ever more business critical. The capacity management team (CMT) is often stretched in different directions by the competing demands for the IT expertise that is necessarily resident within the team. There are fire-fighting demands for optimization, tuning, debugging and detailed 'project work' (usually arising from development demands, test labs or pilot trials). These all compete with effort required to achieve the ITIL description of good infrastructure practice.

Many sites have an attitude to CMP that is derived from a long history of a datacenter glasshouse – silos and ivory towers tend to be the key words. But sites with large investments in major UNIX super-servers, or even so many hundreds of smaller UNIX servers and thousands of Windows servers, are rediscovering the IT planning infrastructure ideas that have served the mainframe so well for so long. The focus and metrics are of course different. The amount of analyst time per application, service or server is much reduced. But the need to balance 'supply versus demand' and 'capacity versus cost' remains the same.

Thus the lessons are clear. In current times, the need is to make the most of the resources already in place, both in terms of computer hardware, software, licenses and support staff and expertise. All the activities undertaken by the CMT need to be reviewed. Long periods set aside to maintain some esoteric reporting regime for a long-stable application could be dropped. Coding corrections to some complex Excel solution developed some years ago by a previous analyst for a particular solution could be dropped in favor of a solution that is now available within some proprietary tool already in place. Excessive reports to an intranet with lots of tables of figures that are out-of-date, on irrelevant metrics or even inactive servers and without any exception reporting could be reviewed.

Thus capacity management remains at the hub of IT service management (ITSM) practices, in summer and winter, economic expansion or recession. It provides the performance metrics and their interpretation to ensure that the IT service is meeting expectations, whether explicit in a service level agreement (SLA) or implicit just by identifying potential relative degradation if nothing is done.

1.2 Introduction to capacity management

This section is a brief summary introduction to the discipline of capacity management and establishes a common foundation for what follows. The headlines in the figures indicate some of the key aspects which will be expanded within this book. They are shown here in brief to indicate to new readers the sorts of issues involved. They are not comprehensive lists – these emerge later and in the appendices. Experienced practitioners will be more interested in later discussions on these topics. Those who are ITIL certified may well find this summary too simplified, but it is intended to establish a common foundation for what follows.

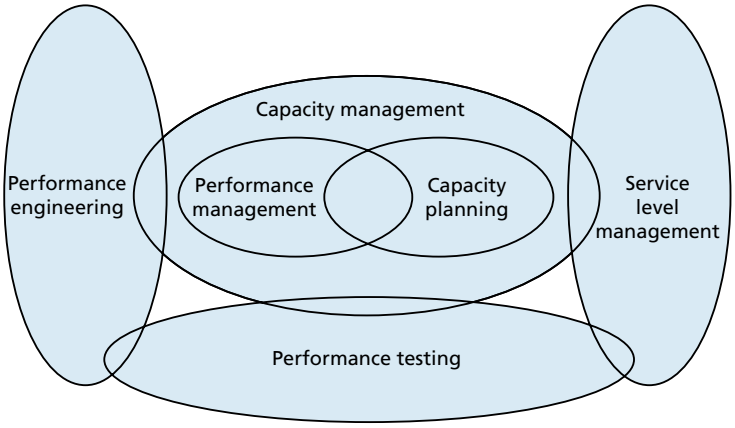


Figure 1.1 Capacity management and some related practices

Capacity management can be thought of as a combination of performance analysis and capacity planning with close links to related activities in performance engineering, performance testing and service level management (SLM), particularly in the area of service level agreements (SLAs). The

practices shown in figure 1.1 indicate a degree of overlap and hence a need for defined interfaces and data flows across the boundaries. Like all good boundaries, there is a need for proper control of passage in both directions with agreed procedures for cooperation on both sides.

The prime objective of capacity management *is the provision of a consistent, acceptable service level at a known and controlled cost. This requires the control of two essential balances: supply versus demand and resources versus cost.*

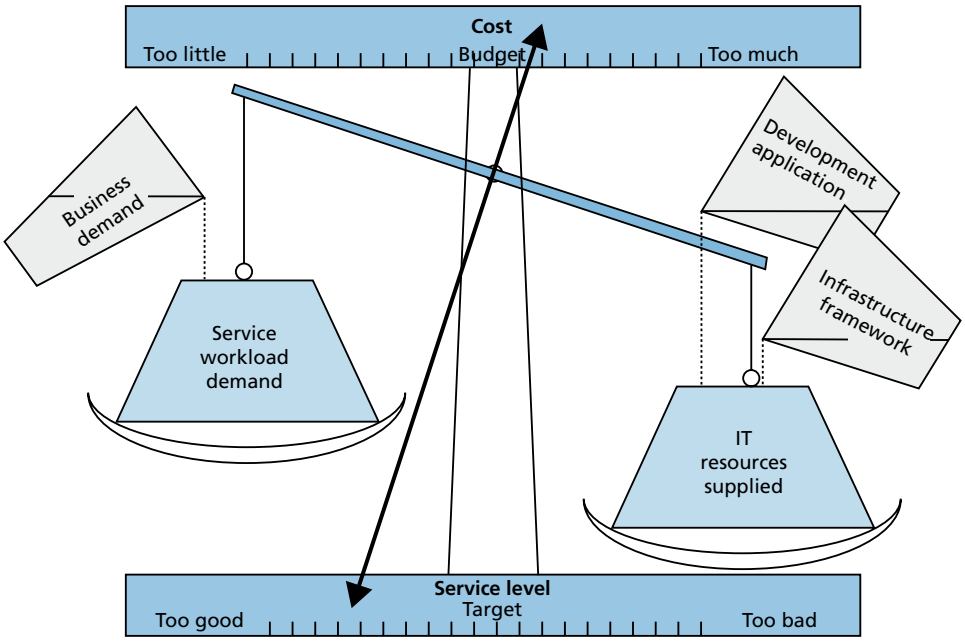


Figure 1.2 Capacity management balances

This balancing act is a continual practice as indicated by the continual changes due to the 'trickles' shown from demand management, application development and infrastructure framework requirements.

In order to maintain this practice, it is necessary to assess things in three views: what has happened in the past, what is happening now and what is likely to happen in the future. This is achieved by three main activities within capacity management: past performance trending, current performance analysis and future performance forecasting. A number of related activities are linked, such as shown in figure 1.3

The list of related activities and where they reside within departments will vary somewhat in different organizations, but again there is an inevitable overlap that will need control to ensure effective control of information.

The prime objective above can be extended and shown beside the list of key deliverables as in figure 1.4.

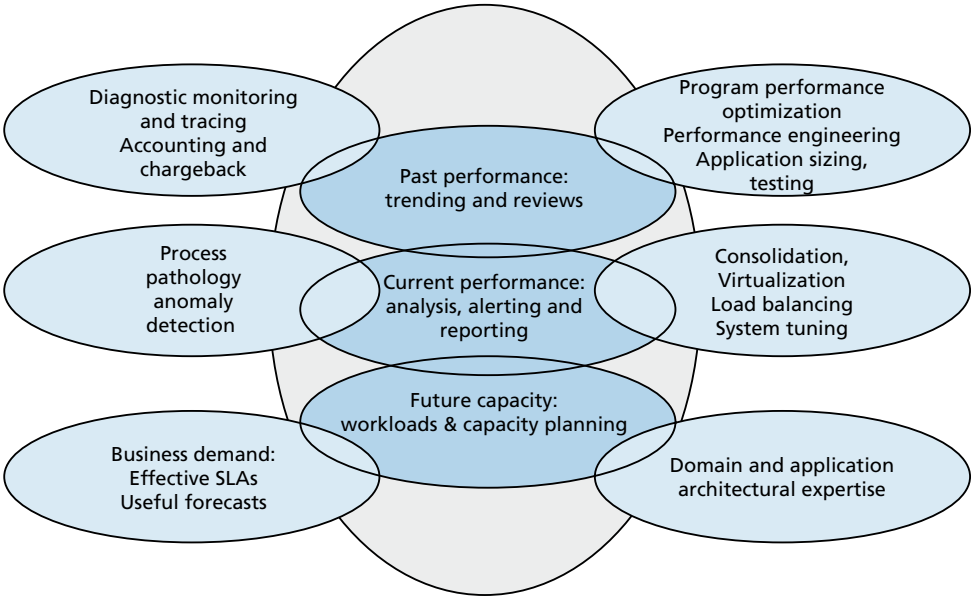


Figure 1.3 Capacity management and related activities

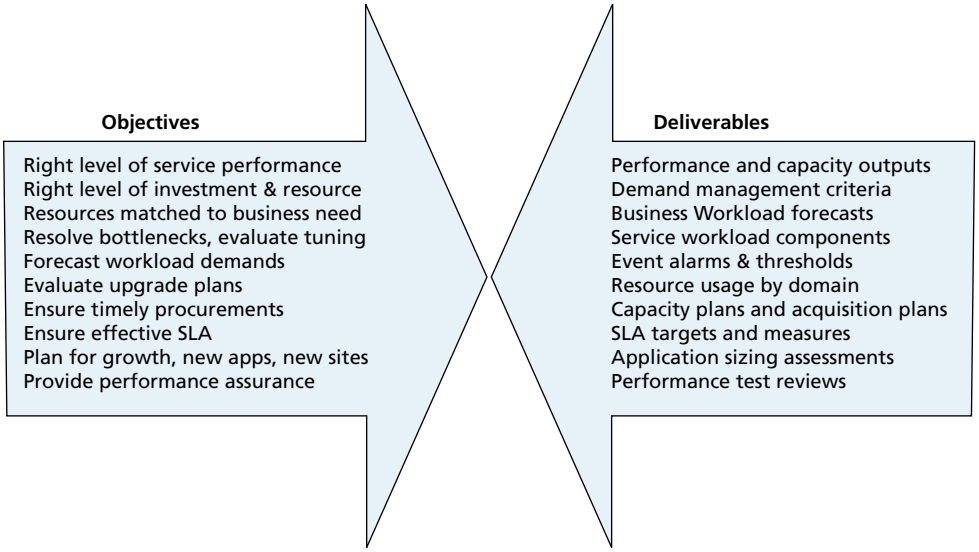


Figure 1.4 Capacity management objectives and deliverables

The objectives and deliverables shown in figure 1.4 do not match line for line, but there is some degree of correlation between the two. These lists are expanded and discussed in later chapters. The practice of capacity management is essentially at three levels, covering the business view, the service view and resource or component view. These are known as business capacity management (BCM), service capacity management (SCM) and resource/component capacity management (R/CCM). The performance and capacity inputs and outputs are many and various.

The inputs are from a wide variety of sources, ranging from business plans and IT budgets to operational statistics. The data captured and collected is maintained in a capacity database (CDB) and used by the capacity management information system (CMIS). The outputs range from performance reports and capacity plans to guidelines for SLAs and event management thresholds.

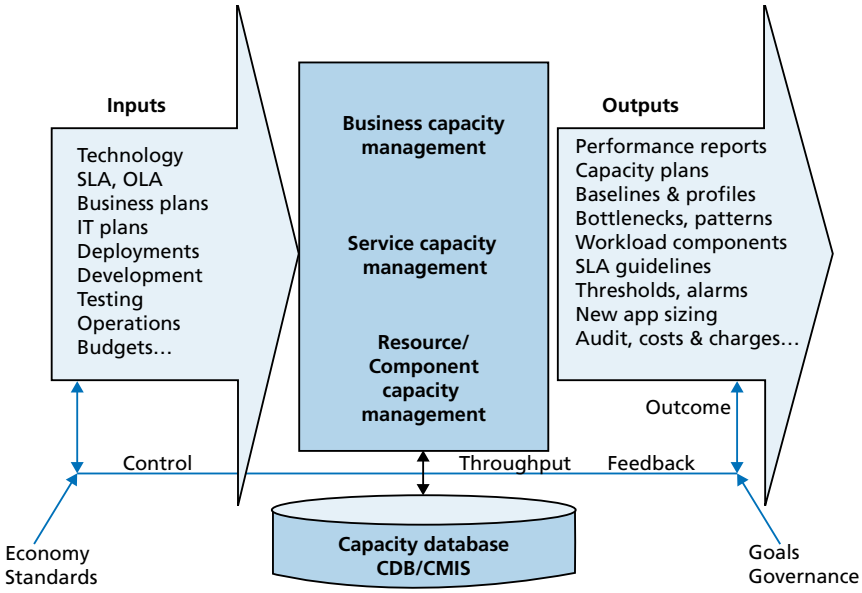


Figure 1.5 Capacity management inputs and outputs

Note that where there is not a useful SLA in terms of performance requirements, it is often useful to coin an internal operational level agreement (OLA) which determines an approximate overall performance requirement, such as 'not more than twice as bad as the current level'.

The sub-practices of business, service and resource or component capacity management involve a number of activities. These are actioned in concert but at different levels of perspective, granularity and time. The level of control to be achieved depends on the organization and the business need for IT. IT resources may be controlled on a minute, hourly or daily basis, IT services reviewed on an hourly, daily or weekly basis and business demands assessed on a monthly, quarterly or annual basis.

The CDB contains large amounts of data from a wide array of sources.

Note that quality assurance (QA) and testing of applications (typically in a test lab) should provide useful input to capacity management. Also the liaison with business should be aimed at deriving relevant metrics for the work to be done, plans for its likely changes and key performance indicators (KPIs) to assess its success. relational database management systems (RDBMS) such as Oracle and SQL Server are major components of many applications. Application instrumentation ideally yields relevant transaction statistics, maybe in the format suitable for application response

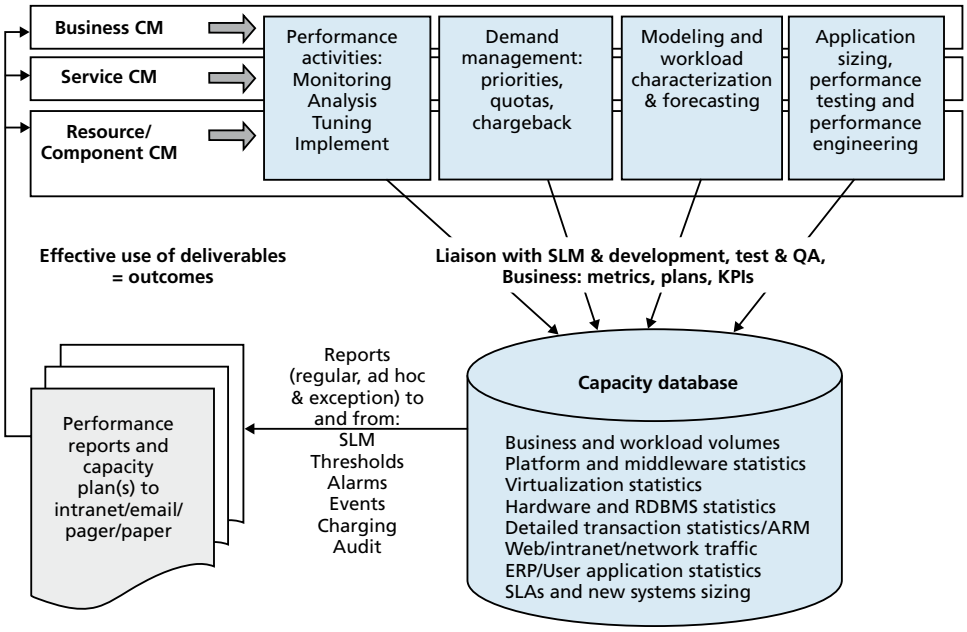


Figure 1.6 Capacity management activities

measurement (ARM) application program interface calls. enterprise resource planning (ERP) solutions such as SAP, Baan, Oracle Financials and Peoplesoft are examples of typical packaged applications.

1.3 Capacity management in IT service management

Most IT providers think in terms of providing services (IT service delivery, ITSD) and managing the provision of those services IT service management (ITSM). capacity management is a key aspect of this.

Capacity management is described in most frameworks for ITSM. Prime examples are the Microsoft Operations Framework (MOF), the Application Service Library (ASL) and the Information Technology Infrastructure Library (ITIL).

In ITIL capacity management is presented as a process. In version 2 of ITIL ('ITIL V2'), it is included as one of five service delivery processes. Version 3 of ITIL ('ITIL V3') expands on the business aspects and lifecycle of IT services. It restructures the ITIL V2 processes with capacity management described largely within a book entitled Service Design. However, although some of the terminology has changed and a few extra activities identified, the outline description of capacity management is much the same in the two versions. Although there are references to capacity management as a *function* and as an *activity* in each of the five ITIL V3 books, the Service Design book largely refers to it as a *process*. In this book we will show how capacity

management can be perceived as a function in an organizational context, and how it relates to dimensions of process, organization, and technology. For practical reasons we will use the term **capacity management practice (CMP)** to indicate all activities involved with capacity management, and **capacity management team (CMT)** to indicate the staff with a commitment to fulfilling capacity management activities. Appendix A addresses the distinctions between functions, processes and practices in some detail.

Capacity is perceived as one of the core attributes of service quality (see figure 1.7), ideally to be found in service agreements. Like most other service attributes, capacity is influenced by applications as well as systems. That is, the demands for, and usage of, capacity are affected both by the application itself as well as the system it is running on. However, whereas the influence of functionality on service delivery in general is largely determined by the application coding, the influence of capacity in general is largely determined by the system (being the hardware/network/facilities infrastructure that applications run on). Capacity management can be applied to all areas of IT and the weighting of factors can vary, such as the capacity of VOIP communication depending on business usage rather than application factors.

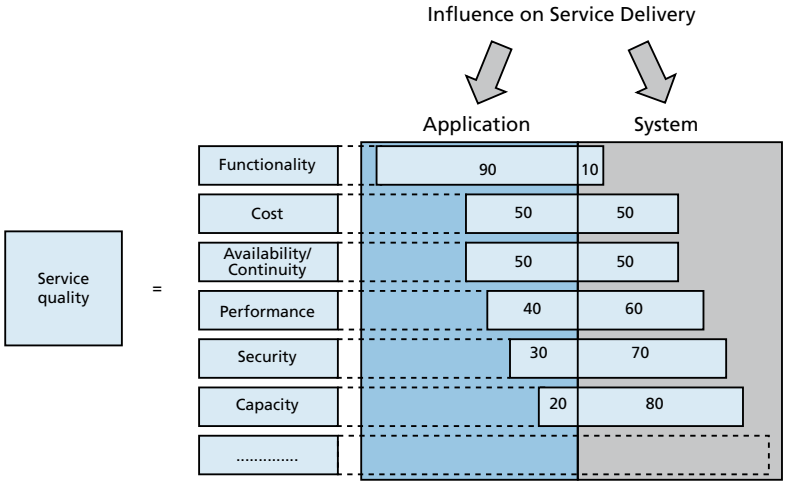


Figure 1.7 Capacity is one of the core attributes of service quality (figures are indicative of their relative contribution, not result of research)

In essence the CMP covers the well known activities of **performance analysis** and **capacity planning** with links to **performance engineering, performance testing** and **service level agreements.** Performance analysis is mostly concerned with production systems, shorter time-scales and detailed data. Capacity planning tends to be longer time-scales with more aggregated data. Performance engineering is mostly concerned with building good performance into the service design and development stages. Performance testing should be part of quality assurance and be done in conjunction with conformance testing, typically in a test lab. Service level agreements ideally contain key requirements for performance and capacity.

Capacity management can be applied at three levels; technology, service and business. These correspond to the interests of the optimum management of resources or components (primarily technology units), the services they support and the business requirements that they are intended to meet. These are described as three sub-processes in ITIL.

Capacity management is applicable to most financial, commercial, industrial, retail, public and non-profit organizations with IT systems designed to provide production level services. It began in the 1970s in the centralized world of mainframes and their datacenters. Initially it was a specialist activity for systems programmers who understood the technical issues to do with the performance of a mainframe and who learned to apply this technical understanding in conjunction with a business understanding of the supported application transactions

In those days there may have been a **capacity management team (CMT)** of performance analysts and capacity planners supporting a single machine (or possibly a small number of machines).

Nowadays, the same functions are required, but applied to a much greater number of machines. As a result, the same team's efforts are typically spread across so many hundreds or even thousands of machines. The solution architecture is often more complex in that there may be many levels or tiers involved with pools of machines at each level. So although there is less time available per machine and each machine may have a different role with specific workload profiles, the discipline remains technically similar. However, there is now the added burden of gaining understanding of and focusing on business criticality issues, as the applications have become more mission critical as well as more numerous. Furthermore, as there is often less technical appreciation of the relevant factors and less time to understand the normal behavior of any one machine, so there is more need to automate the practice as much as possible and to avoid the most extreme dysfunctions.

*The **essential task** of capacity management is to look at the current performance of a service, identify any bottlenecks, understand the workload placed on it and the underlying business drivers that may affect future traffic. The next task is to assess the workload growth history and define potential future scenarios, then to map that workload demand on to the existing resource configuration to assess likely future performance. This needs to take into account the likely future resource configuration in the light of improved device service times. The final task is to predict the impact of any suggested changes to ensure maintenance of the required level of service.*

Thus capacity management is generally viewed as embracing both the activities involved in performance management as well as capacity planning. These tend to work in different timeframes and with different objectives but together help to establish the optimum balance between demand and supply as well as costs and resources.

Related activities typically include the establishment of a reporting regime, usually to pagers, email or intranets, to provide a reference point for all interested in the performance of the machines and the services running on them. This also helps with the assessment of new applications as they emerge, preferably in pre-production saturation testing before they are launched into production.

Although a CMT is often set-up in many enterprises, its role is not always clearly understood by those working in the business areas or the IT service infrastructure. It may be dominated by domain experts who represent the detailed technical resource expertise within the enterprise. So there is always a danger of the activities being focused on fire-fighting, in the sense of resolving current problems and panics, at the expense of planning to avoid them. This is often caused by project managers who tend to ignore non-functional requirements. There are many reasons for firefighting becoming dominant. One is the CMT's pleasure at becoming 'heroes'. But it would be better to try to find the gremlins causing the issues which are making it all go wrong.

Gremlins are well known. They are creatures, commonly depicted as mischievous and technically oriented. Their origins were with World War II airmen, claiming that gremlins were responsible for sabotaging aircraft. Grumlins are similar but smaller and more aggressive (check the author's surname) and are responsible for sabotaging the best endeavors of the CMT.

Grumlin #1 is borrowed from Murphy and states that 'anything that can go wrong, will go wrong'. This means that the CMT should look for program loops, memory leaks, missing indexes, fragmented discs, housekeeping schedules et cetera to avoid building forecasts based on a badly tuned system.

1.4 Capacity management and ITIL

The Information Technology Infrastructure Library (ITIL) offers a systematic approach to the delivery of quality IT services. ITIL was developed in the 1980s and 1990s by CCTA (Central Computer and Telecommunications Agency, now part of the Office of Government and Commerce, OGC), under contract to the UK government[5]. Since then, ITIL has provided not only a widely used framework, but also an approach and philosophy that is shared by the people who work with it in practice. ITIL has been updated twice, the fist time in 2000-2002 (V2), and the second time in 2007 (V3).
The ITIL approach has been widely adopted throughout the world in all sizes and shapes of organizations, and is the de facto framework for IT service management.

ITIL V3 uses the concept of a service lifecycle for the provision of IT services. There are five core books that describe the lifecycle. In order, these are Service Strategy, Service Design, Service Transition, Service Operation, and Continual Service Improvement. ITIL is technology independent, for example avoiding mention of specific operating systems, database platforms, or service management toolsets. The library provides information on concepts, processes, functions, activities, organization, methods, techniques, tools, implementation considerations, and measures. In V3, capacity management is viewed largely as part of the service design phase of the service lifecycle. In ITIL V3, a service has two dimensions, utility (functionality, what it does) and warranty (provisioning, how well it does it). Capacity management is described as a warranty

5 www.best-management-practice.com lists various documents produced by the OGC, including ITIL (for ITSM), PRINCE2 (for project management), Managing Successful Programmes (MSP for programme management), Management of Risk (M_o_R) and Portfolio, Programme and Project Offices (P30). The Structured Systems Analysis and Design Method (SSADM) is in some references under www.ogc.gov.uk

attribute. Capacity management is described at a high level in some fifty pages, with content extracted from the orginal book in version one[6] and similar to a single chapter in version two.

Top level frameworks such as ITIL are usually introduced with numerous lower level details about processes, sub-processes, interfaces and required data flows. There is always a danger that the description of a reference framework can become viewed as a formal definition of recommended procedures and activities. Sometimes the mere formalization of a view of a subject from the top down introduces a structure that becomes an entity in itself. This has unfortunately happened with ITIL, with many people incorrectly believing that you must explicitly follow its recommendations. This is not the case. The ideas from ITIL should be 'adopted and adapted', tailored to meet the specific needs and circumstances of an organization. In fact, very few organizations have adopted the whole of ITIL; most adopt just a few areas of the whole library.

1.4.1 Processes, functions and activities

Because ITIL covers the full service lifecycle, there are a high number of subject areas, which ITIL calls 'processes'. There are twenty-six[7] of these, most of which have sub-processes. There are also subject areas that ITIL calls 'functions': e.g. service desk, application management, technical management, and IT operations management. Apart from these, ITIL also describes a lot of other practices that it calls 'activities'. These terms are discussed in some detail in appendix A1.

Most of what ITIL calls a 'process' does not follow the definition of what a process is (according to the definition presented in ITIL[8]). One solution to this is to consider the twenty-six ITIL 'processes' as twenty-six ITIL 'practices'. So capacity management in ITIL, which uses several real processes to achieve its aims, can be considered to be a 'practice', and not a 'process'.

In the early days, the organizational functions needed within IT were clear: Development and Operations. Since then Development has separated into functions for analysis, design, coding, and testing, before merging into software engineering and then splitting again. Operations has similarly split into functions such as systems management, network management, and desktop management.

For IT service management there is no need to create twenty-six functional departments, just because there are twenty-six ITIL practices described in ITIL. How practices are mapped within an organization is as variable as the organizations themselves, and the people within them. However, most sites will try to establish standards to reduce the variations.

---

6 The original book was written by Brian Johnson, one of this book's referees.

7 Estimates of this number vary, mostly in the twenties. The excellent IT skeptic (www.itskeptic.org) has shown the inconsistency within seven ITIL V3 references containing 22 processes in all sources and 40 mentioned in at least one source (with 11 only in one). So the likely range is 22-29, depending on which sources you most respect.

8 ITIL V3 definitions are:
- A process is "a structured set of activities designed to accomplish a specific objective. A process takes one or more defined inputs and turns them into defined outputs".
- A function is "a team of people and the tools they use to carry out one or more processes or activities".
- An activity is "a set of actions designed to achieve a particular result".

Grumlin #2 is that the great thing about IT standards is that there are so many to choose from.

Grumlin #3 is that on average, new standards last for eighteen months so that in any large enterprise there are typically five active standard definitions in various degrees of implementation in place for each aspect of IT and five to the power 'n' permutations for the 'n' regimes of standards defined.

Capacity management is viewed largely as part of the service design phase of the service lifecycle. ITIL does not mention information specifically required for capacity management such as performance data sources, virtualization levels, consolidation trends, utility computing, blades, parallel processing, multi-core processors, clustered technologies and so on. These architectures are very dynamic and their performance issues are very real and have to be addressed. Vendors of capacity management tools have a continuing challenge to ensure that they keep up with new architectures and new releases. Members of the CMT have a continuing challenge in keeping up with new technology.

So there is a need to consider such matters and that is usually achieved by reference to domain experts. They in turn attend conferences and read papers to keep up-to-date in this specialist area. Formal certification on general principles is a useful starting point, but the real education comes from sharing experiences with other experts in the field addressing today's practical issues. This is achieved for capacity management primarily by the **Computer Measurement Group (CMG)**. This organization was founded in the USA in 1975 when capacity management was sometimes known as computer performance evaluation. Related topics have been included over the years with various descriptions of ITSM, IT infrastructure planning (ITIP), performance engineering and capacity management. The CMG annual conferences and meetings in the USA, UK, Central Europe and other local and international chapters are a primary source for the latest approach to current topics. For more information about CMG see www.cmg.org. There are typically some 200 papers given yearly at CMG (and more in the international chapters such as the United Kingdom Computer Measurement Group, UKCMG) so there is a vast amount of material to review.

1.4.2 Key parameters

ITIL and other books talk of People, Process, Products (or Technology) and Partners as the key interrelated entities concerned with the establishment of an effective ITSM practice regime. This book extends the alliteration by adding Price and Planet.

Price ensures that the costs of providing the different elements of IT service management are taken into account, prioritizing the implementation of processes or functions that can achieve the largest benefit. A categorization of the importance of services to the business is often introduced to identify those that are 'mission critical', 'standard', 'specialist' and 'other' (or something similar, and sometimes called 'service tiering' with typical levels of gold, silver and bronze). ITIL is then applied in full for the mission critical services, with a reduced application for services that are less critical.

Planet implies care of the planet, in that IT services should be sustainable and green so that capacity management should limit the enterprise IT negative impact on the environment. Examples are less powerful processors for lightly utilized servers, consolidation of lightly utilized servers (maybe including virtualization), smart data archiving solutions, demand management to set quotas and remove false traffic peaks and so on.

Playfully, the alliteration of the six P's above (People, Process, Products, Partners, Price and Planet) can be even further extended. The main attributes of the people and partner entities include politics, policies and psyche. This is using the word 'psyche' to imply the impact of 'company culture' which often dominates ITIL implementation considerations (and sometimes described as 'ABC' considerations[9]). A company with poor results and increasing competition is likely to be frugal in all its thinking. A growing company with a critical IT service is liable to focus on maximum practical performance. Most companies lie in the middle, trying to get the balance 'just right'.

People is used in this case to include partners, policies, politics and psyche. The issues that affect process include projects, practice, plans, prognostications and procedures. Products can include platforms and proprietary tools. The main output from price concerns is the allocation of priorities with due consideration to issues affecting the price impact on the planet in terms of CO_2 emissions et cetera. See figure 1.8.

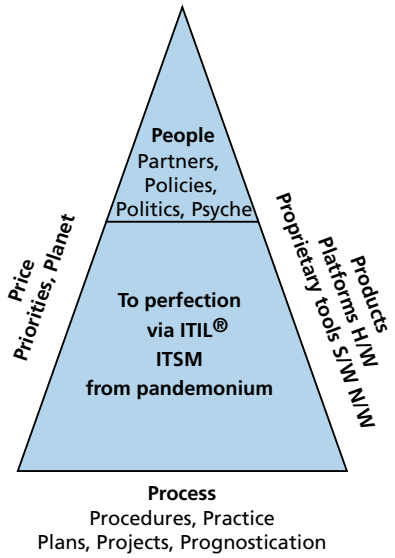


Figure 1.8 ITSM parameters

9 Paul Wilkinson has written about Attitude, Behavior and Culture as the vital, 'soft skill', key aspects for success in implementing ITSM practices.

ITSM parameters include:

- **People** – are the entity least inclined to conform to predefined process definitions; **partners** – involve relationships between people and so the complexity is exponentially increased with the numbers involved; **policies** are the declared objectives for the organization; **politics** are how these are interpreted; and **psyche** is the 'company culture' which sets an overall attitude framework.
- **Process** – is a set of prescribed activities using resources to transform inputs to outputs, i.e. 'what is to be done'; **procedures** – are the detailed instructions to carry out the activities of a specific process, i.e. 'who/how to do it'; **policy** – is the overall intention and direction of a service provider formally expressed by senior management; **practice** is what actually happens; **plans** are what is meant to happen; **projects** are the workloads arising from development and **prognostication** is the forecasting of business demand.
- **Product** – is often the clearest on the pricing issue (unless developed in-house), but perhaps not on the cost-benefit; typically based on **platforms** for hardware and software; with **proprietary-tools** for hardware, software, networks et cetera.
- **Price** – is the largest factor on pragmatic decisions, and control of it requires the allocation of **priorities** to business demand and resultant services; and **planet**, the consideration of sustainability and green issues.

All of these parameters influence capacity management practices (CMP).

1.5 Capacity management maturity

A common analysis of success in business suggests it can be ascribed to the following key elements, which are here interpreted in the adoption of ITIL:

- vision (ITIL offers a good framework to achieve IT and business objectives)
- mission (to implement ITIL practices)
- belief (ITIL will make IT service management more effective)
- strategy (define practices for consideration)
- tactics (select and define activities for planned action)
- operations (do it)

Process 'maturity' levels are often defined based on five levels (or six if inactivity is included), according to the Capability Maturity Model (CMM) or CMM Integration (CMMI) model as defined by the Carnegie Mellon Software Engineering, as illustrated in figure A.10 (Appendix A). A similar approach can be followed for capacity management maturity. The precise terminology associated with maturity levels varies in different presentations, but the levels described below are typical.

In capacity management terms (to be discussed in more detail later) the levels may be thought of as moving from simple monitoring through analysis to the three sub-practices of capacity management as described in ITIL. The related activities and tasks vary from simple monitoring of utilization through trending to capacity planning and corporate performance management with executive dashboards. Although the detailed interpretations of each of these levels and what

they mean in terms of capacity management is open to debate, the following are reasonable indicators:

- **Level 0** – means that there are no activities or procedures of CMP in place.
- **Level 1** – implies an ad hoc response each time to any event with ad hoc or post hoc attempts to gather relevant information to address a capacity-related incident.
- **Level 2** – moves up to a reactive response based on precedent, with a basic measuring regime in place to gather core metrics such as utilizations for capacity and performance and uptime for availability.
- **Level 3** – moves up to a proactive practice established to minimize the potential events, with a capacity management database (CDB/CMIS) and resource or component capacity management (R/CCM) in place, some simple event management around performance alerts, basic utilization trend reporting and core metrics published to the intranet.
- **Level 4** – is where that practice can be measured for its effectiveness with an effective service portfolio or catalogue including meaningful performance objectives and capacity planning in place for service capacity management (SCM).
- **Level 5** – is where the practice is optimized, with business capacity management (BCM) in place and key performance indicators (KPIs) measured and optimized for the CMP itself.

An appreciation of leading enterprise sites indicates informally that only a very small percentage (the 'bleeding edge') is in the area of level 5. A larger but still small percentage (the 'leading edge') would be at level 4. The bulk of sites would place themselves somewhere within levels 2 and 3, maybe accounting for more than half of all sites. The remaining sites do little or only in an ad hoc manner at levels 0 and 1. See figure 1.9.

#	CMMI	ITSM	Capacity management	Task	%
6	Optimized	bITa	Business level	Dashboard CPM	2%
6	Measured	ITSM	Service level	SLAM Capacity plans Service catalogue	8%
6	Proactive	Center	Resource level	CDB Trends Web reporting	30%
6	Reactive	Tickets	Analysis	Utilization Uptime Some event monitoring	50%
6	Ad hoc	Help calls	Monitor	Ad hoc alerts Ad hoc investigations	10%

Figure 1.9 Capacity management maturity – showing the top 5 levels of CMMI and their relationship to capacity management and typical adoption in practice

Key findings from Gartner[10], using their own proprietary maturity model but with a similar overview, indicate that the users, customers, management and boards who express satisfaction with their IT services have a close correlation with the 'bleeding edge' or 'leading edge' sites. The best sites in practice work across the levels according to the priority of the service and have a spectrum of activity something like one of the bracket curves shown in figure 1.9 where the most effort is only applied to the key business critical services.

There is clearly a possibility of over-simplification in this approach. The level of maturity of solution applied to a service changes during its lifecycle. A long–running application that has few changes (maybe a payroll) requires little attention beyond monitoring resource level utilization. Maturity level is also impacted by the volatility of the business and the criticality of the service to the business. A major new application requires application sizing as early as practicable in its lifecycle and should continually be reviewed but typically less rigorously as time goes by.

Further, the definition of each level is subjective and some activities in different levels may be applied to any given service. Particular activities will be selected to meet the particular issues of the service. This is particularly true for a multi-tier solution with pools of machines at different levels being employed to provide the service. The degree of focus overall will depend on mission criticality of the service, but within each tier it may well depend on the cost of the machines.

The practical conclusion in most sites is to adopt a flexible approach as indicated by the dotted bracket lines in figure 1.9. This is a reflection of what actually happens in most cases; it is not the integrated solution that would ideally be proposed, it is more a pragmatic recognition of what is feasible in most sites. Most sites start with a performance monitoring regime across the board (as it is impossible to know which machine is likely to be next at risk). A standard service reporting regime based on R/CCM is usually adopted for the majority of services. This is typically delivered to an intranet, where the number of hits by users outside the capacity management team should become a matter of interest. Ad hoc reporting is typically used by the team itself to deal with new issues. Then for a selection of services, maybe mission critical, key and major production services, a capacity planning regime is exercised. This needs to be based on historic trends in terms of likely traffic growth and likely resource utilization levels. For the machines where there is a workload mix and a significant financial investment, formal capacity planning is performed using modeling techniques. Sometimes this will also be applied to a sample of other services or servers. This is true particularly where there are hundreds or thousands of servers of similar type doing similar jobs for a distributed solution. In that case, a categorization by market segmentation or customer type is usually defined, such as 'typical urban', 'typical suburban', 'typical rural' et cetera (where the location of a branch office indicates the likely population of users and hence related IT traffic and demand for services).

10 D. Curtis, HP Software Forum Denver. June 2005, http://hpsf.conferencearchives.com/2005

1.6 Demand management

A typical management edict for IT is to 'do more with less'. But typically there are more requests for work than resources. Demand management is commonly proposed as a way to understand and throttle demand from customers. It is important as requests for projects often outstrip the resource capabilities of service providers.

Demand management is described as a capacity management activity within service delivery in ITIL V2 with a constrained view of its scope (focusing on degradation of service due to unexpected increases in demand or partial interruptions to service due to hardware or software faults and establishing the redistribution of capacity in order to minimize the impact on business critical services). In ITIL V3 it is allocated to service strategy with a wider view of its scope and links with capacity management identified, but still focused on patterns of business activity and user profiles. In this book it is treated as a capacity management related activity and is interpreted as most practitioners use the term including both of the above as well as establishing longer term practices to deal with handling requests for new services, avoiding un-necessary peaks in workload, provisioning of resources, setting service priorities and quotas, chargeback and related activities.

The objective of demand management is to optimize and rationalize the demand for the allocation and use of IT resources. It covers the entire spectrum, from one extreme of over-provisioning without regard to cost to the other extreme of under-provisioning such that there is no headroom and hence capacity problems.

Effective demand management and capacity management ensures the timely provisioning and efficient allocation of IT resources at three levels:
1. forecasting business demands
2. applying IT strategy and assessing service trends
3. assessing and controlling resource or component utilization levels

At an IT component level, most large organizations have adopted capacity management by monitoring capacity and performance from servers, storage, networks, and so on. They have also used modeling and trending tools to predict future requirements. Fewer have successfully shifted the emphasis from day-to-day needs to a more proactive, business-centric view of future requirements. Effective capacity management must factor in future business developments, including step changes in demand that may arise from longer-term business initiatives or advances in technology. The future plans for the business in terms of growth or change are key to this. It is not always readily available outside the Board Room and so often some simple rules of thumb need to be applied by the CMT, such as 'if the current demand is X and the performance is Y, then the service needs to cater for 2X and the worst performance acceptable is 2Y'.

Delivering effective capacity management requires forecasting business demands, applying IT strategy and assessing service trends, and evaluating utilization of the current implementation. Key factors include:
- identifying which services are vital to the success of the business

- ensuring that these services are available as the business needs them
- improving efficiency by ensuring that resources are not over-provisioned

The key output is a capacity plan for IT resources that:
- facilitates successful management of IT assets
- monitors and communicates key performance indicators (KPIs)
- is able to evolve as business demands change

1.7 Roles for capacity management

Capacity management touches upon many very practical issues in day-to-day management of an IT service organization. Many of these are addressed in relation to financial management, service level management, operations management et cetera with input from the CMP. These are discussed in more detail in later chapters. In essence, the key role is the provision of decision support in the establishment of the optimum IT services at the minimum cost.

Decision makers from different departments of an IT organization have to make many decisions affecting cost, performance, availability and other service levels for the services provided to end users so that the enterprise can compete successfully in its marketplace. The decisions range across a number of areas including:
- application to be bespoke or packaged or 'Commercial Off The Shelf' (COTS)?
- development to be in-house, outsourced, off-shore?
- service to be outsourced or 'facilities managed'?
- centralized or distributed? Open or proprietary? Thick or thin clients?
- single tier or multi-tier? Pools of small machines or fewer large ones?
- clustered, coupled or multi-core processors?
- applications consolidated or virtualized or both?
- devices duplexed, triplexed or mirrored or striped?
- capacity 'right-sized' or 'safely-sized' or 'over-sized' or 'on demand' to meet local needs?
- hardware standby on fail-over, flip-flop, disaster site?
- storage retention period, classification of data (information life cycle management)

These are all complex issues to be resolved in the light of local factors, typically with significant IT budget implications, as well as major corporate impact. Capacity management can help provide the decision support infrastructure by establishing the instrumentation, metrics, analysis and reporting to aid these activities.

Managing a complex IT environment requires the analysis of many trade-offs. The wrong decisions will negatively affect IT performance and result in lost business opportunities. Businesses, vendors and customers demand accurate and timely information from IT, in order to enable real-time management of business processes and to satisfy governance requirements and regulations. Equally, finance management will expect the infrastructure to be sized correctly to meet the business objectives without overspend. IT is tasked to deliver new applications, support more users, process more data and implement new IT solutions with predictable and low total cost of ownership (TCO).

*The **role of capacity management** is to answer a variety of 'what if' questions, to evaluate alternatives and tradeoffs, to compare options and justify strategic and tactical decisions during all phases of the application and information lifecycle. Essentially to enable continuous, proactive performance engineering, performance management and capacity planning and so provide accurate and timely information that can be used to ensure the delivery of quality, cost-effective services to the business.*

New IT solutions, new applications and more users may also increase the contention for resources and the key balance lies in establishing the right amount of computing resource to ensure acceptable service delivery to meet the real current and projected future business demand. This balanced plan is likely to be also influenced by issues of cost, new technology and sustainability.

1.8 IT service management drivers

ITIL and ISO/IEC 20000 are currently the dominant default descriptions for IT service management (ITSM). Other drivers come in to play when considering the practices to be adopted within enterprises. Quality initiatives are encapsulated by total quality management (TQM) and ISO/IEC 9000. Audit requirements driven by SOX and Basel II and COBIT contribute to the need for defined IT processes. Process improvement lies at the heart of Six Sigma and continuous improvement. Business process re-engineering has paved the way for process definitions and Balanced Scorecards and the Capability Maturity Model (CMM) to refine process control.

The Office of Government and Commerce (OGC), in the UK, owns the intellectual property rights to ITIL. OGC have outsourced the management of ITIL to TSO (for publications) and APMG (for examinations). APMG have licensed several other exam institutes, including EXIN and ISEB.

Also, in the Netherlands, the ASL BiSL Foundation promotes professionalizing application management and business information management and improving the communication between these two IT management domains by spreading knowledge on their process models.
These describe structuring the management and maintenance practices of existing information systems on one hand (ASL – Application Services Library) and defining business information management practices on the other (BiSL – Business information Services Library).

1.9 Basic concepts and terminology

1.9.1 Basic concepts for IT service management

Capacity management, like other practices in the IT service organization, finds its roots in IT service management. Appendix A describes the basic concepts for IT service management. It deals with common terms and definitions, e.g. service, value, system, process, function, line, maturity, and organization. It also provides information about important questions for the structuring of service organizations:

- What is the difference between a process and a function?
- How can process management be organized?

- What are core processes in a service organization?
- How to construct a process model?
- How can functions be constructed in a service organization?
- How can maturity models be used for improvement?

These basic concepts apply to the rest of this book, and to the other books in the Practioner Guide series.

1.9.2 Jargon and acronyms

ITIL terminology is used throughout this book, using both ITIL V2 and ITIL V3 terms. This reflects what is in common usage at the time of writing and may well be for some time.

*ITIL V2 described a **CDB** for capacity management database and a **CMDB** for configuration management database. V3 introduced the **CMIS** for capacity management information system and the **CIS** for configuration management information system. This book adopts the convention of referring to the **CMDB/CIS** and the **CDB/CMIS**. In the same way, V2 described a sub-process of resource capacity management and V3 renamed it component capacity management. This book refers to the activity of resource/component capacity management or **R/CCM**.*

Difficulties with jargon abound within ITSM. There is a risk that each organization develops slightly different approaches to, and interpretations of, related disciplines. **Performance engineering** can cover a variety of activities to try to ensure and improve the end result for a particular service. **Software performance engineering (SPE)** is usually taken to cover assessment of likely performance needs and resulting resource demands of new applications at an early stage in their development. **Performance assurance** is often used to describe the entire domain of development, testing and operations in the area of performance. It includes aspects of performance engineering within development, performance testing within quality assurance and service level management within service management as well as capacity management

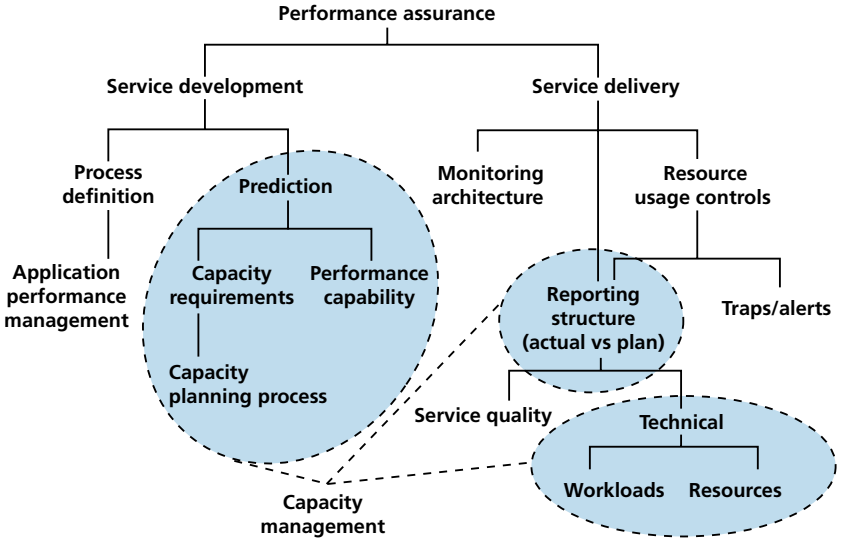


Figure 1.10 Performance assurance and CMP

itself. Capacity management is typically seen as a practice covering most of the technical aspects involved. This is indicated in figure 1.4, showing the overlaps between a performance assurance hierarchy (shown as a tree structure) and capacity management practice (shown in dotted circles).

How this is mapped out in terms of actual teams and functions varies significantly in differing organizations; figure 1.10 is intended to indicate the nature of the overlap rather than make any specific recommendations.

2 What: What is capacity management

This chapter describes the objectives of capacity management and provides some detail of capacity management practices. It then introduces more of the basic concepts involved and discusses the main terms and definitions, such as resource, service, performance, workload and so on.

2.1 Objectives

The **primary objective** for capacity management is to serve the needs of the business by ensuring that the organization understands and tracks demand and can maintain required service levels under both normal and contingency conditions both now and in the future within agreed cost constraints. The essential goal is to achieve the most cost-effective balance between business demand and the size and form of the IT Infrastructure needed to support it.
To put it another way, it means having the right resources, in the right place at the right time for the right costs to do the right work at the right speed to meet the right target for the right users to exploit the right performance to achieve the right results for the business.

The right work has to be assessed in demand management, gauging the priority and fluctuations likely so that the scaling limits of the current situation can be determined. This will be influenced largely by the business context. At one extreme, where downtime and poor performance is not an option, the strategic directive might be to enable the business to continue with its IT services at all times and at all costs. At the other extreme, the strategic directive might be to minimize IT costs to meet a reduced budget no matter what the impact is on IT services. In most cases, the situation is somewhere in between. In all cases, it needs the CMP to put numbers to the current situation and enable the directive to be met.

Arising from the primary objective for capacity management, are a number of **strategic objectives** at tactical and operational level. How these are mapped on to the local CMP and CMT varies across sites.

The **strategic objectives** for capacity management are to:
- ensure the right level of investment, resource and capacity
- match equipment to business need, supply and demand
- identify and resolve bottlenecks
- evaluate tuning strategies
- improve and report/publish performance
- 'right-size' or 'consolidate' or 'virtualize'
- evaluate upgrade plans to ensure accurate and timely procurements
- ensure effective service level management
- plan for workload growth, new applications, new sites
- avoid performance disasters and achieve performance assurance.

These can be further resolved to a tactical level. The **tactical objectives** of capacity management are to:

- identify relevant metrics for KPIs and resource usage
- establish thresholds for values and patterns of behavior for key metrics
- produce workload forecasts, including all workloads for all resources on all servers that are within scope and significance
- track actual usage against plan at resource and workload levels
- provide support to developers in producing capacity requirement forecasts for significant, new or changed applications
- provide financial information for the organization's budgeting activities based on predicted capacity requirements
- Keep informed and aware of new technologies that might impact on capacity

The **operational objectives** of capacity management are to:

- set and use appropriate performance criteria for components or resources (service levels and/or utilization thresholds) to ensure that the tactical objective will be met
- cater for specific constraints such as lead times (internal and external), cost or service compromises, change management and forward schedule of change, shortage of environmental resources (floor space, power, air conditioning, et cetera)
- track all agreed configuration items throughout their lifecycle for all equipment – central and distributed
- liaise effectively with configuration management as the capacity database (CDB) and the configuration management database (CMDB) effectively share a common index, in that both need to manage a lot of information on the same assets

All of these objectives are laid out in the checklist in Appendix C.1 in a form suitable for audit and gap analysis.

2.2 Definition of capacity management practice (CMP)

There are a number of descriptions of this practice. A succinct and good one is:

> *The **capacity management practice (CMP)** encompasses the provision of a consistent, acceptable capacity service level at a known and controlled cost.*

ITIL states that capacity management is responsible for ensuring that the capacity of the IT infrastructure matches the evolving demands of the business in the most cost-effective and timely manner. It expands on the balancing acts involved, weighing costs against capacity and supply against demand. The former ensures that the processing capacity is cost justified and optimized. The latter ensures that the available resources match the demands made to meet the business need.

The balance chosen is necessarily context-driven and thus local to each enterprise: a rich financial institution involved with stocks and shares in London City or New York Wall Street transactions

may decide that it is necessary to triplicate everything in order to minimize downtime which may cost millions, whereas an IT service supplier may try to deliver a service that closely matches imposed service level agreements underpinning contracts, with IT resource provision 'just in time' in order to maintain maximum profit.

Capacity management aims *to understand the business needs, both present and future, and to map these onto services which will be required to achieve the stated goals. Capacity management must also provide a plan showing when, where and what resources will be needed to meet these service requirements. Without this forward-looking activity, there could be any number of unpleasant surprises, such as:*

- *capacity and performance crises*
- *unnecessary hardware expenditure*
- *user dissatisfaction*

Capacity management is responsible for ensuring adequate capacity is available at all times to meet the requirements of the business. It is directly related to the business requirements and is not simply about the performance of the system's components, individually or collectively.

An important distinction to bear in mind in many sites is that there may be teams of people committed to the capacity management function for different domains, the capacity management team (CMT), but there is a separate issue of the practice of capacity management across the enterprise, the capacity management practice (CMP). The precise definitions vary across enterprises and even sites within them, with overlaps between projects, domains, architects, systems specialists, fire-fighting, process implementation and planning. Where the distinction between the team and the practice is particularly critical, the acronyms CMT and CMP are used.

2.3 Main terms and definitions

Appendix B provides a glossary of terms used as well as a list of acronyms and a listof key terms used. This section is intended to provide an initial 'walk-through' some of the main terms and definitions to establish a foundation for the following sections.

2.3.1 Resources

It is common capacity management jargon to say that a physical computer and network configuration supplies a limited set of resources, either directly as 'real' or indirectly as 'virtual', which can be utilized (or consumed) by the systems or applications that run on it. Resources include, but are not limited to:

- CPU time (utilization)
- memory occupancy (free pages)
- page space on disk (paging activity)
- file space on disk (spare file space)
- network bandwidth (network traffic)
- I/O throughput capacity (I/O activity)

2.3.2 Service

A service is a described set of facilities, both ICT and non-ICT, sustained by the ICT service provider to fulfill one or more needs of the customer, and to support the customers' business objectives. The customer perceives a service as a coherent whole.

Typically a (full) service is defined in terms of a group of specific applications on specific machines, real or virtual, designed to service the requirements of a particular business function. These full services can be decomposed into partial services, e.g. some application-specific services that 'use' network services, desktop services, or other generic services which are often shared by multiple customers. A service also calls for support services, such as service desk, service reporting and service requests (how to purchase or decommission the service).

ITIL V3 is expressed in terms of the lifecycle of a service. IT service management is the implementation and management of IT services to the quality that meets the needs of the business. Service level management is the practice responsible for negotiating service level agreements (SLAs) and ensuring that they are met and policed.

It is necessary to define a sensible measure for demand of a service to show its pattern of activity and behavior over time, either in terms of observed values of a useful metric in the past or predicted values for the future. This is usually presented using a single 'representative value' over a period for that metric, reflecting the value during a key period of busy activity, as shown in figure 2.1.

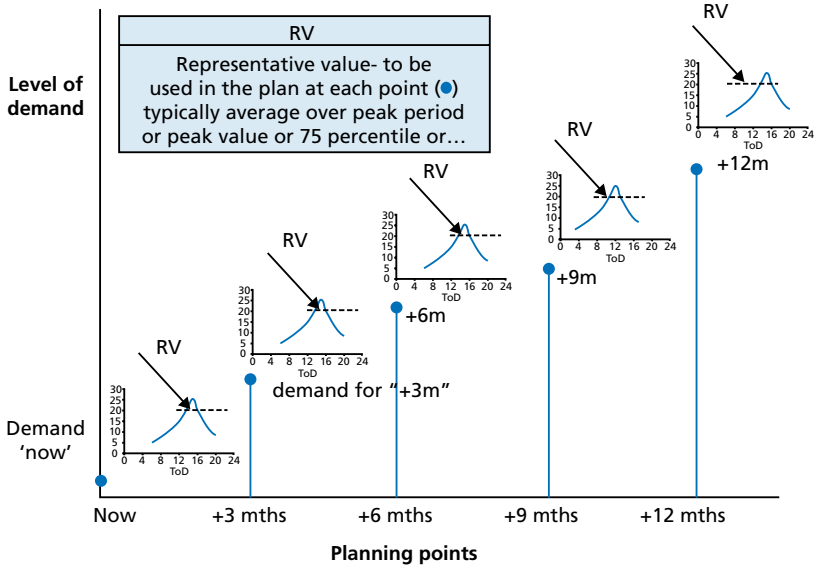


Figure 2.1 Showing representative values over time

Many graphs presented in capacity management are of this nature, in that a single line is often shown which is representing snapshot averages or other metrics over a period when there is a distribution of values involved. It is often the only practical way to show a pattern over a

significant period of time. What should be remembered is that each point typically represents just an average or percentile value for a selected snapshot interval. And that any line drawn through the points is probably an approximation and if extended into the future should include an indication of the probability that the line is valid by showing confidence limits.

2.3.3 Performance and capacity

Performance and capacity are the two key issues for capacity management. Performance is essentially about the speed at which something can be completed and is typically measured as a response time. Throughput is the count of the number of such things completed in a given time. Capacity is the number of such things that could be completed in a given time and is the maximum throughput. Using a motorway analogy[11], the journey time is a measure of performance, the traffic flow per hour is a measure of the throughput and the 'gridlock' level of traffic is the maximum capacity.

The performance of a service is a key issue for most real applications. Particularly in the online industry such as news sites where users will not tolerate waiting and will quickly switch to an alternative site. Once the SLA has basic availability and continuity issues under control, the users are most likely next to express concern that although the service is 'up' it is not really functioning as the responsiveness of the application is so poor that the business can not be supported adequately. Some SLAs define this as 'slowtime' and include slowtime in unavailability. At this point in the discussion, it becomes clear that the use of a stop watch to try to collect information on response times is impractical. The concepts of services, applications and transactions emerge to define the workload and assess the throughput and response times in an objective and measurable manner. However, performance is not only a matter of responsiveness. If the application is poorly written or the logic annoying, then a user may well switch to another service supplier, no matter how quick the system is.

The capacity of a system is not simply the sum of the quoted capacities of all the devices in a given solution. It is the maximum throughput that a service can deliver whilst meeting SLA performance targets. It can be greatly reduced by the presence of a bottleneck device which saturates at lower traffic levels than other devices in the service. Spare capacity is not simply the 'idle' or 'unused' time; it is the measure of the amount of extra useful work that could be undertaken without detriment to the SLA performance targets.

2.3.4 Workload

The workload is the extent of a resource use in a defined period. It tends to be used to imply the throughput of work for particular groups of users or functions within the organization.
In the context of capacity management modeling, workload refers to a set of forecasts, which detail the estimated resource usage over an agreed planning horizon. Workloads generally represent discrete business applications and can be further sub-divided into types of work (interactive, timesharing, batch).

11 An analogy ably introduced and expounded by Michael Ley, a referee of this book.

Workloads may be categorized by users, groups of users, or functions within the IT service. This is used to assist in analyzing and managing the capacity, performance and utilization of configuration items and IT services. The term workload is sometimes used as a synonym for throughput.

The workload on a set of resources can be defined at whatever level matches the analysis required. For a single application server, it may be that the entire workload can be treated as a 'black box' with a 'system level' model showing how the resource behaves overall. If the server supports multiple applications, or if there is a need to separate the production work from system overheads such as back-up, then the workload needs to be 'characterized'. This is the technique of identifying how the workload on a server is split, for example by groups of users in different departments doing different work, or maybe simply by virtual machines. This is discussed in more detail in section 2.5.

2.3.5 Metrics

A metric is something that can be measured objectively to help manage or control what is going on. In order to assess the performance of a particular resource or device, it is vital that it is instrumented and reports on a range of metrics. The dominant examples would be the number of 'arrivals' at a device in a given period, the number held in a stack or queue waiting for service, the time each takes to be serviced, the total busy time of the device and related statistics. There have been many attempts to define standards for such metrics. These have ranged from the universal measurement architecture (UMA) to the application response measurement (ARM) initiatives[12] and also a variety of simple network management protocol[13] (SNMP) and management information base (MIB) definitions. Although each initiative has had its own devotees, none of these have achieved universal implementation. So capacity management has to accommodate a wide variety of metrics of different quality from a wide variety of sources. Most of these measures can only be obtained readily once the service (or a prototype) is running. As such, from the COBIT point of view, they are known as 'outcome measures'.

2.3.6 Data sources

There are three types of data that form the inputs for capacity analysis and modeling:

1. **Machine configuration details** – What kinds and models of host machines are present; what is their physical and logical configuration (processors, memory et cetera): where virtualization is involved what are the physical, logical and management parameters (in terms of the virtual machine view, the hypervisor view and the physical view) for the relevant domains, LPARS, VMs et cetera. Where there are, or have been, significant planned or actual changes to

12 The Open Group has defined the application response measurement (ARM) standard which describes a common method for integrating enterprise applications as manageable entities. It includes definitions for measures on application availability, performance, usage, and end-to-end transaction response time. Ideally, if widely adopted, ARM would allow users to extend enterprise management tools directly to conformant applications creating a comprehensive end-to-end management capability. It was preceded by a similar attempt at proposing a standard called the universal measurement architecture (UMA) which defined an open system for collecting and managing performance data in a distributed open systems environment but failed to achieve wide adoption.

13 Simple network management protocol (SNMP) is used in network management and administration systems to monitor network-attached devices for exception conditions as defined by the internet engineering task force (IETF). SNMP exposes management data in the form of variables which describe the system configuration and can be queried or possibly changed. The information on the managed objects resides in a standard format on a management information base (MIB).

configurations (upgrades, replacements, re-assignment of workload categories) it is very useful to also have this information as a time-line or event list.

2. **Machine performance data** – This is in the CDB/CMIS (which will contain resource-level information about the system utilizations and the different user-workloads which are present), augmented by any available application-level statistics. These are most likely to be transaction counts of some kind, and they may be provided by middleware such as Tuxedo or WebSphere or other business-logic. The usefulness of these is that they can provide a link between the low-level resource behavior (e.g. processor utilizations or I/O counts), and the high-level numbers coming from the final set of data.
3. **Service reporting statistics, and service forecasting quantities** – Since these are the drivers for capacity planning, access is needed to the following types of business numbers:
   - Reports of actual business activity. Examples are number of flights per day for an airline or number of teller transactions for a bank. This is required preferably on a historical basis that extends over at least one business cycle, and in just sufficient level of detail to distinguish the major workloads on the infrastructure: by service, and preferably in the same level of detail that is to be planned for. This data can be correlated with the levels of machine-activity in the historical baseline.
   - Descriptions of any critical scenarios (end-of-month or seasonal peaks, conjunctions of different workloads, current known shortcomings et cetera).
   - Forecasts of future levels – Realistically, these may be in different terms to the actual figures above and also volatile since they are the product of business planning. They will need to be translated into resource requirements where possible to understand the true impact. It is very useful to have some explicit idea of confidence levels attached to business forecasting
   - it allows sensitivity analysis to be performed at the hardware level. This is an area where communication skills are critical. The CMT needs to coach the business users to appreciate the significance of the forecast numbers they provide.

The main data sources for most metrics of interest to R/CCM are those provided by the operating system, RDBMS, any middleware and related statistics from hardware, software logs and catalogs and trails and traces as appropriate. For SCM, the focus moves to the application and how well instrumented it is and how much service oriented data it provides. For BCM, the focus moves up again to a higher level of aggregation and the application instrumentation again is relied on for as much information as possible.

2.3.7 Queuing

Queuing is where a service has a limited resource and customers seeking their requirements form an orderly queue to wait their turn. Forming orderly queues is part of the inherent characteristics of the Anglo-Saxon. Other parts of the world may not have the same word ('waiting in line' being a popular alternative) or even share the concept, but most computer systems rely on servicing queues in the best manner to help improve throughput.

Nearly all servers of any size and complexity will be hosting applications that are accessed by several (possibly very many) users simultaneously. As soon as there is more than one user, there is a finite probability that several users will attempt to gain simultaneous access to the same resource – the CPU or a particular disk, for example. In the simplest case, one will be executing, and the

others will have to wait for it to finish. In other words, there will be contention for resources, and contention causes queues to form.

It is completely normal for queues to form in a multi-user server. This is how operating systems support multiple users – by organizing the various requests into queues and managing these queues as efficiently as possible. This has been a key part of operating system design from the outset in the early sixties, and designers are now very good at it.

The problems arise if the queuing within the multiple-user server becomes 'excessive'. The effect will be degraded performance and – potentially – the inability of the system to support the required workload within agreed performance targets. Typically, excessive queues are associated with long end-user response times and reduced system throughput.

Excessive queuing is often caused by a specific bottleneck device, and one of the key objectives of performance management is to identify such devices and to reduce or eliminate their negative impact on overall system performance, by means of appropriate tuning or the provisioning of additional hardware.

2.3.8 Model

A model can be described as:

*A **model** is a simplification of reality built for a specific purpose, preferably incorporating measurements that can be modified under the control of laws.*

There is little point in building a model that takes more effort than the benefit derived from it. The degree of accuracy in the model and the acceptable tolerances in its fabrication are significant factors. But the purpose of a model is always the key factor. It is essential to know what questions it is intended to answer before creating it. Consider two models of an aircraft. One is for the aerodynamicist and is a balsa wood outline of the body to test in a wind tunnel. The other is for an ergonomicist and is a plywood mock-up of the cabin and instrumentation. Both models are totally different and both invaluable in their own right.

In the case of capacity management, different models can be used to assess different issues. The three most popular in practice are trends, simulation models and analytical models. A simple spreadsheet model may well be used to review the history of a given workload, using trend analysis to define a likely growth. A discrete event simulation model may be used to look at the dynamic behavior of a network. An analytical model may be used to assess current performance and predict future performance in the light of different scenarios. These techniques are discussed in more detail in section 4.

2.4 Basic concepts of capacity management

Capacity management is usually described as three sub-practices, at resource or component level (technology infrastructure), service level (delivery) and business level (management view). These levels can be described separately but in practice they overlap. Sometimes the distinctions

correspond to the different timescales involved, in that the hardware view is typically fairly short term (down to hours or minutes) and highly granular. The service view is more likely to be looking at a higher level of aggregation of data and over a longer period of days and weeks. The business view introduces even more aggregation of data, as well as weighting of its significance, to yield interpretations of the data for longer term business assessment. However, all three sub-practices may be involved in addressing similar problems over similar timescales and the distinctions should not obstruct the resolution process.

2.4.1 Resource or component capacity management

Resource or component capacity management (R/CCM) focuses on the technology that underpins the service provisions. The **resource or component** sub-practice refers to the management of individual components of the IT infrastructure (corresponding to the configuration items (CI's) in the CMDB/CIS) and ensures that bottlenecks are addressed. This is traditional performance analysis and capacity planning and is the most popularly implemented aspect of capacity management.

Note that 'resource management' in the sense of human resources, people, personnel and their management or demand control is not within the scope of CMP.

Effective resource or component capacity management needs to capture and store all relevant utilization data for all components of the environment. Data capture agents and import of data from external sources ensure that the CDB contains all relevant performance data. Integral scheduling technology ensures the collection of all data whenever needed irrespective of the source of the data.

2.4.2 Service capacity management

Service capacity management (SCM) addresses the delivery of existing services that support the business. The **service level** relates to the management of the performance of production applications and ensures that Service Level Agreements are defined in a measurable manner and are then policed. This is the performance aspect of service level management which many leading enterprises are now addressing.

More and more, the focus for day-to-day capacity management is moving from individual resource or 'silo' views of the world to a service or application orientated perspective. The need is to assess current and future performance against SLAs and service level requirements (SLRs) across the service. 'application views' enable all the elements that comprise each of the supported services to be reported on as a single entity, with drill down to the individual resource areas as required. This focus is reinforced by service oriented architecture (SOA).

2.4.3 Business capacity management

Business capacity management (BCM) is concerned with current and future business requirements. The **business level** relates to business-IT alignment and ensures that the key business needs are met. This is sometimes known as corporate performance management and is at the leading edge of capacity management.

Open data import technology is required to ensure that IT and non-IT business data is quickly and easily imported and stored in the capacity management Database. Where this data is produced on a regular basis by the business, the collection and storage of the data needs to be automated guaranteeing its availability. Once in the capacity management Database it is available for analysis and reporting alongside all data collected for service and resource/component capacity management.

This enables the correlation of business data over time with observed resource utilization and also enables an effective 'feedback loop' to be established with the business: business users see the impact of their business on services, and of those services on resources or /components, leading to better quality business information being provided over time.

2.5 Scope of coverage

The scope of capacity management practice is indicated in figure 2.2, the capacity management pyramid.

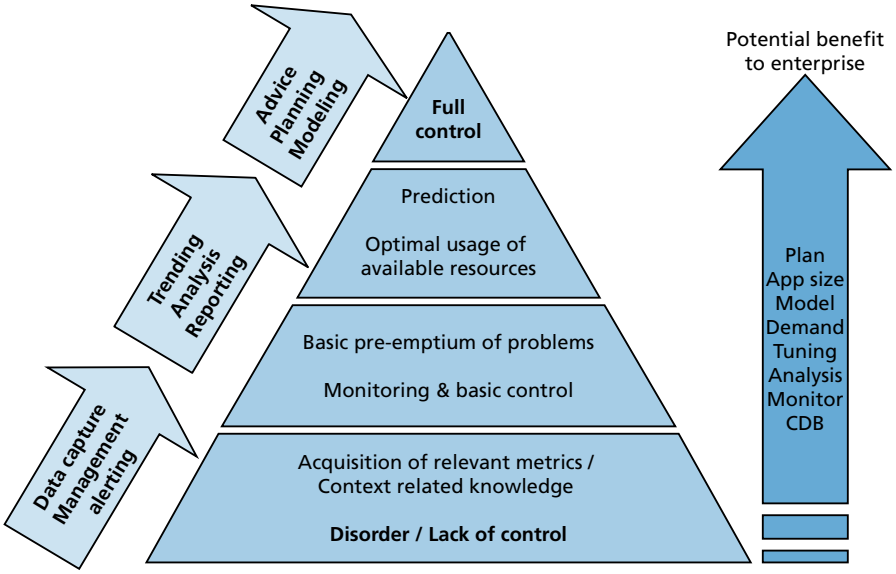


Figure 2.2 Capacity management pyramid – showing the path from chaos to control with the activities involved and showing the increasing potential benefit to the enterprise

The capacity management pyramid covers a variety of tasks which are introduced below to provide a basis for further discussion in chapter 4 of the detailed activities.

The scope of capacity management is described in ITIL with some V2 graphics that have become standard representations and enhanced in V3. Workforce resource management is not within the

scope of capacity management as viewed by ITIL[14]. Although there are some common practices in terms of modeling techniques, the two are usually separated. Nonetheless workforce resource management is central for service providers to ensure that there is not poor delivery of service due to lack of personnel, or any slack in the availability of personnel is otherwise exploited.

Datacenter facility capacity management is a growing aspect of capacity management. With the technological changes of blades, virtualization etc. power and cooling demands per square meter are increasing. Many datacenters were built to older specifications and high density servers place high demands on floor loading and air conditioning. Many companies do not own the facilities and need supplier management to ensure capacity expansion and responsibilities are contractually clear.

An alternative traditional view of capacity management can be shown in figure 2.3. Note that, in the absence of a relevant contractual SLA, service level objectives (SLOs) are often defined internally by the CMT to provide a yardstick for measurements. This is typically a statement such as 'the current performance level must not degrade by more than 50%'. Also, note that the 'stress test results' shown in the figure imply all the results from any performance testing, volume/load, stress, saturation and soak testing. A further aspect is the use of the testing platform to establish the transaction flow for the main transaction(s). That is, the relevant calls it makes as it passes through the system.

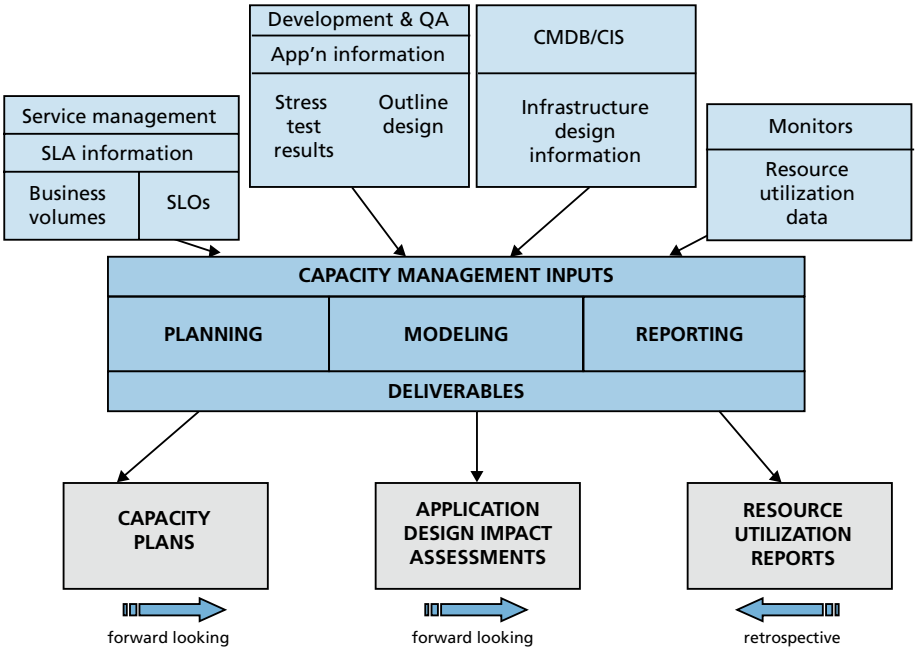


Figure 2.3 Capacity management inputs, practices and deliverables

14 There is a reference in ITIL to lack of human resources causing a delay in end-to-end response time, but that is probably based on operational problems on mainframes in the old days of operators mounting tapes.

In order to gain an understanding of what lies behind the myriad of commands, processes, I/Os within a computer system, it is a useful technique to identify workloads that can be used to simplify analysis. These are typically based on applications or users or services in some way.

2.5.1 Workload characterization

Workload characterization is the task of identifying the workloads within a system, usually breaking it down into workload components. That is, refining the workload into components that represent different aspects of the workload that need to be controlled separately in that they have different business drivers for growth. Also often categorized by key attributes such as its type (interactive, batch), priority (mission-critical) et cetera.

Workload characterization can be as simple or as complex as is required to meet the needs of the enterprise. It forms both the basis of performance reporting, whether by an application view of performance metrics across a number of machines or users or applications. It is the basis of deriving 'workload components' in a model of a multi-tasking multi-user machine. It is worthwhile to consider some real life examples to introduce the ideas.

In the case of Windows, for example, in many sites the Wintel servers are restricted to a single production application which can often be related to a small set of CPU processes that are continually running on that server. The business drivers for such systems are often essentially the number of concurrent users of that service. In order to gain a better picture of the true production workload on the machine, some sites will isolate all the known administrative processes into an 'overhead' workload (typically picking up on security, back-up, anti-virus, anti-spam, defragmentation processes and the like). This will leave a final workload component covering all the rest of the unallocated processes.

Virtualization of lightly used windows machines is a growing trend, which makes the machine essentially multi-tasking and the simplest approach is to treat each Virtual Machine as a separate workload component. In order to gain the whole picture, the information is usually gleaned from the hypervisor or virtualization machine monitor/manager as well as from the physical operating system itself. The required performance metrics are readily available from perfmon.dll reading the object counter information from the registry.

In the case of the mainframe, the concept of workload characterization has been in place for a long time. System programmers have been used to grouping applications into terms such as 'performance groups', 'service classes' and 'service report classes' to achieve much the same benefit. All the performance metrics are readily available as required, already allocated across these classes.

In the case of UNIX the usual approach is typically based on grouping users to reflect a departmental use of a machine, or grouping commands to reflect a particular application, or maybe a combination of the two to identify preferred views. The metrics are available to some degree, either by reading kernel counters (typically via 'devkmem') or by using standard UNIX utilities (typically 'sar', 'iostat' et cetera). However, not all the resource information is available per user and some mapping of I/O has to be done. In other words, the total number of logical block requests is known and the total physical I/O per device is known, but not the number of I/Os per device per command. Thus the performance analyst has to play a game called 'guess who

did the I/O'. To do this well requires detailed knowledge of the mapping of files to devices. This is not always available, so typically simplified pro-rata assumptions are made.

A typical pair of workload characterization levels might be:
- level 1:
  - service classification relating service usage and hence resource usage to business lines
  - used to establish accountability, cost management, service and resource usage analysis by business function
- level 2:
  - technical classification to define online or batch, normal or contingency, accurate or inaccurate measures, firm or vague planned commitment
  - used to create workload forecast scenarios and input to resource management.

These categories can then be refined and exploited to establish workload limits such as:
- parameter 1: committed/uncommitted (firm workload)
- parameter 2: critical/noncritical (outage in disaster recovery (DR))
- parameter 3: online/batch (batch also meaning 'can be shifted to outside peak period', rather than simply a batch stream workload and so can be applied to back-up, archive, reorganize, management reports et cetera).

Use parameters 1 and 2 to construct combined workloads and create upper and lower bounds (UB and LB) for normal and contingency situations, thus:
- LBNORMAL is ALL COMMITTED
- LBCONTINGENCY is COMMITTED AND CRITICAL
- UBNORMAL is ALL
- UBCONTINGENCY is ALL CRITICAL

And use parameter 3 to separate online or batch workloads (again, where batch also means 'can be shifted to outside peak period', rather than simply a batch stream workload).

2.5.2 Business drivers

The concept of business drivers is essentially that of relating the traffic on a computer to the business requirement it is intended to satisfy. In a mature capacity management site, this information is available from a number of sources. The application itself records a log file in a time series format of significant transactions completed with key information. The user department maintains a file of workload information and the business management maintains a log of business metrics such as KPIs or business metrics of interest (BMIs). Ideally, some of this information is also incorporated in the configuration management database and the service catalogue which may help to prioritize services. All of these can then be reported on and analyzed with correlations and trends to help relate the business view to the service and resource or component level views.

Other terms for this are 'business forecast unit' (BFU) and 'natural forecast unit' (NFU). These are all terms the business uses to describe their workload that can be related to IT resource usage and may be different for different resource types. For example, business drivers may be orders placed, customer enquiries or web site hits. Measures may be maintained such that business volumes are recorded within application statistics, the SLA creation and review process or a BFU

metric used for dynamic resources within the user area. The links to IT resources have to be identified and are variable, for example, disk space may reflect the number of orders placed, disk I/O the number of users and processor utilization the number of web site hits.

In many sites, however, this information is not so well defined and simpler assumptions have to be made. A frequent, practical parameter used is the number of concurrent users of a system. In the absence of that number, an even simpler assumption is based on the enterprise published overall figures for growth (whether based on turn-over, profit or any other financial measure). Thus the physical resource demands can be viewed in relation to the past business activity numbers and a trend based on the projected activity growth.

When meaningful BFUs can be defined, they:
- relate business volumes to service usage and hence resource usage
- require correlation between service and workload
- have inventories that are typically related to the following IT metrics:
  - user IDs
  - account codes
  - transaction IDs
  - job names
  - command names
  - process names
  - application instrumentation
  - self instrumentation
  - storage quota (mailbox size, personal data share)

The mapping of business drivers to applications and hence to resource demands is a skilled task involving an understanding of a number of different views of IT. It can be compared to a translator's task, although in this case it is mostly a question of jargon rather than language. Figure 2.4 makes the point. Consider 'transactions' for example. The end user will think in terms of a business action such as raising invoices. The system designer will think in terms of some application parameter such as 'creating invoice'. The programmer will think of a software object with interactions and calls to other objects, with activity arising from so many calls to each per transaction visit to the first object.

2.5.3 Service and server classification

There are major issues of scale in implementing capacity management. Firstly there is the number of services. Although there might be many hundreds defined, it is usually possible to categorize them into a few mission critical services and maybe fifty key services and the rest given a lower priority. Such numbers have to be manageable with a reasonable size of team committed to capacity management, such as five to ten full time employees (FTE).

Secondly there is the number of servers. There may well be many more instances than in the past, within a variety of platforms and solutions that are virtualized, multi-tier and have complex configurations with pools or clustered machines involved. To address this, in practice, many sites will also classify servers in order to treat each class appropriately.

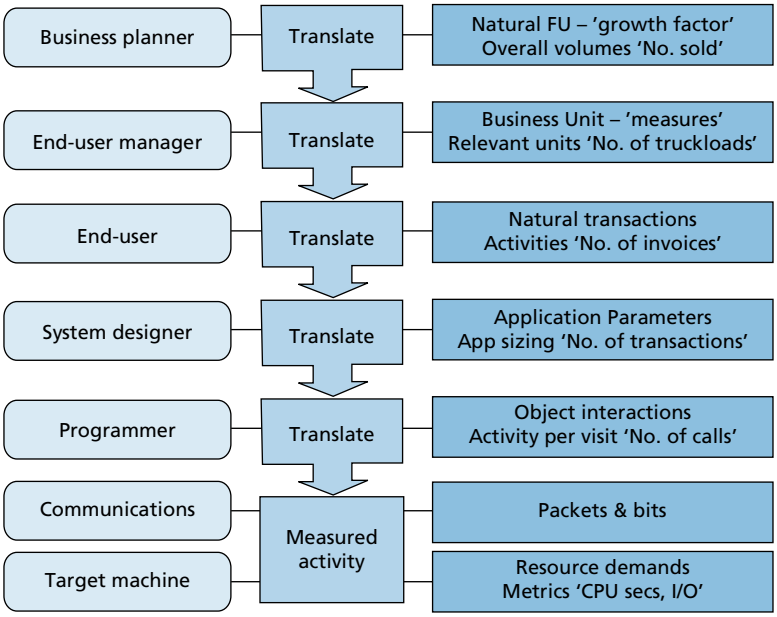


Figure 2.4 CMP workload – resource translation

This is highlighted by the major distinction that most sites still draw between the mainframe and distributed systems. The mainframe is centralized, with a limited number of distinct configurations. It is possible to monitor and report on everything as data gathering is well instrumented and interpretation well proven. Note that although the mainframe is often taken to mean machines from IBM running z/OS, z/VM, or z/Linux, there are other environments rightly thought of as mainframes. The original 'bunch'[15] of alternatives are mostly gone, but some solutions such as HP NonStop (originally based on Tandem), OpenVMS clusters including AlphaServers (based on VAX and Digital Tru64), Sun E25000 (derived from a Cray), and Teradata warehouses are all in the same mould in that there is a well established regime for large installations and well organized ITSM tools.

Distributed systems, on the other hand, have tens, hundreds or even thousands of servers involved with a huge number of combinations (despite standard builds). In this case, most sites try to establish norms, monitor exceptions from the norm and produce report exceptions (alarms). In this way, it is possible to minimize data gathering and processing by using an approach of summarization and selecting representative samples for trending and modeling.

The classification schemes vary but are typically by application (e.g. SMNP server, text processing, email or database software) or by business function (e.g. sales, stores). For each class a standard configuration is often defined. Then for each standard configuration, or band of server, rules of thumb may emerge in the light of experience and measurement, such as a 'medium branch server' will support 'n Notes users' or 'n Sales users'.

15 Burroughs, Univac, NCR, CDC and Honeywell were the bunch. Others in the list could have included DEC (now part of HP) and ICL (now part of Fujitsu).

Servers can also be usefully grouped according to their power rating. This can be based on public benchmarks such as Spec[16] or TPC[17] benchmarks. These numbers are useful guidelines to relative power of machines from the same supplier of the range and same architecture and using the same operating system et cetera. They are only used with great caution in any comparison across platforms.

The approach is one of selecting appropriate 'exemplar configurations'. These have to be one per category which is representative of its class (often the busiest). The exemplars are then used to track and trend growth (just like mainframes) and to identify norms. These can then be translated into application views of the data on all similar machines that can be readily automated into classified reporting regimes.

Relating capacity requirements to business drivers is necessary to plan for non-linear-trend factors and to justify investment and to assess operational quality requirements. As capacity management also has a deep view on the ways systems are managed, there is often some reluctance from co-workers to reveal complete data as the CMP may be viewed as 'unnecessary monitoring activity' and the CMT as a 'menace'.

Capacity planning for the 'top twenty' services (typically on multi-service servers) is required and is practical. It is simplified if there is a simple server/service mapping or simple virtual machine analysis. It is also simplified if there is a smaller number of larger servers, or if there is a rationalized hardware infrastructure and classification based on 'standard builds'.

2.5.4 Consolidation and virtualization

The move towards greater utilization of servers and more cost effective provision of IT services has led to increasing consolidation and virtualization.

Consolidation

'Consolidation' has previously been known as down-sizing or right-sizing. It can mean different things, depending on the entities being consolidated. Most people think initially of server consolidation and maybe moving workloads to a larger machine, possibly taking their disk farm with them. But the server itself may just be moved to a central location, or the disks may be added to a SAN. Also, the workloads may be moved to separate logical partitions on a larger machine, or possibly to virtual machines on a larger machine.

The performance metrics are there in each instance to help address possible scenarios and predict likely outcomes.

Virtualization

'Virtualization' is a buzzword used everywhere but with slightly different meanings depending on the context.
Separating the user or developer from the 'nitty-gritty' has been an integral part of computing

16 See the benchmarks on www.spec.org. SpecIntRate is a useful one for comparing the raw power of CPUs but others can be used to represent particular workloads.

17 See www.tpc.org and select a benchmark that is a reasonable representation of a typical workload of the type you are planning to support. Bear in mind that the actual production workload will be specific to your own applications.

since day one. Layers of abstraction have been added to coding. Disks have been directly attached and then via NAS then SAN. Memory has been virtualized for decades. But recent moves have tried to add processors, servers, services, networks and even data centers to the virtualization mix.

Let's consider some of these in turn.

Coding abstraction
Coding layers of abstraction have gone from machine code to assembler mnemonics to high level languages like COBOL to RDBMS languages or frameworks like ORACLE to object oriented languages and environments like Java and .NET and to automated application generators and coding generators like OmniBuilder.

In the past people have ascribed 'generations' to these languages but this gives a false impression. There are 'horses for courses' and depending on the situation any mix of any of these approaches may prove right. Although most new applications will aspire to the latest trend, most data centers still have a mix of languages of different generations in production use.

Disk arrays
RAID is now dominant. Although there are numerous logical definitions of levels (from 0, 1, 2, 3, 4, 5, 6, 7, 8, 0+1,10 et cetera) just a few are particularly prevalent. Directly attached storage devices are still popular and can yield optimum performance for dedicated applications. NAS is widely in use, largely because of the ease of implementation. SAN is also prevalent, especially where high performance is required alongside effective back-up and recovery.

Processors
There are a number of approaches to try to increase the effective power of a server. Processors can be closely coupled, loosely coupled in clusters or interconnected by special devices or buses in a mesh or a grid.

Clusters are part of this scene. They enable a number of servers to be linked and work in conjunction. They are typically established for the use of numbers of machines in a distributed processing environment with parallel processing, or for failover or for high availability load balancing. Thus pools of servers can be dedicated to a single service such as web servers, mail servers, database or file servers or applications with appropriate software to coordinate the servers and provide effectively a single large computing resource.

Multi-processors
Multi-processors are another route towards more power by adding more processors to a given server to exploit parallel processing or multiple threads across processors. The architectures for this have developed over the years. Initially, some were asymmetric, where one processor was the master and a number of others were slaves to it. This was replaced by symmetric processing which is now dominant but has limitations such that they are mostly around 4-16 and maybe 32 processors sharing a memory bus. Ways to increase further were defined as massively parallel processors (MPP) and non-uniform memory access (NUMA), which either use memory links or a crossbar switch to connect memories in a large matrix. Both remain as niche products but

enable super-servers with 128 or more processors, although other techniques now achieve the same end. Now there is also a trend to stack more processors on a single chip, with dual core and soon multi-core. This may well be a reflection of the fact that the frequencies involved in current chips are now at very high levels and the challenges in production are being limited by various practical factors such as heat as well as electro-magnetic issues and cable lengths. So stacking more chips on the same chipset is adding an alternative axis for improvement.

Partitioning
Partitioning has a long and well established history. Early machines had physical partitions to isolate applications that were essentially static, although some operating systems allowed them to be changed if the relevant applications were in quiescent state. This led to the introduction of logical partitions or domains that allowed for multiple instances of operating systems on the same server. This proves useful in separating development and production, or allowing for different instances of operating system on the same machine, or with appropriate 'front-end' software, dynamic workload balancing across partitions. This is all now part of the super-server environment as well as the mainframe.

Processor virtualization
To virtualize is to 'decouple workloads and data from functional details of physical platform they are hosted on'. The approaches to virtualization tend to depend on where the microcode and software interrupts lie. Clearly there is the hardware level, the operating system and the applications. Most people are aware of hardware virtualization, where the operating system talks to a virtual layer which in turns talks to the hardware drivers, thus presenting an easier interface to the hardware for the operating system and known as para-virtualization, with hypervisors or virtual machine manager software to control the virtual machines. Software virtualization is where the application talks to a new virtualization layer representing the operating system rather than the operating system itself. Processor virtualization is achieved by hyper-threading where the O/S is led to believe it has more processor resource than it actually has.

Implementation of virtualization

Virtualization is a buzzword used widely with slightly different meanings depending on the context, but the use of virtual machines (VMs) to contain the workload of (previously) separate physical servers on a single server is key.

Separating the user or developer from the hardware limitations has been an integral part of computing since day one. Layers of abstraction have been added to coding. Disks have been directly attached and then via network attached storage (NAS) and then storage attached networks (SAN). Communications networks have evolved with local area networks (LANs) and wide area networks (WANS). Memory has been virtualized within operating systems for decades. But recent moves have tried to add processors, servers, services and even data centers to the virtualization mix.

Virtual operating systems and the use of hardware, software and virtual partitions has been in place for some time, with a rapidly growing range of options and terminology from the leading suppliers. The jargon and acronyms for each supplier are ever changing as new releases and new concepts emerge, but a general current picture is listed below (without expansion on the

meanings or interpretations of each set of acronyms which are best viewed on the supplies' web sites):

- IBM with mainframes zSeries and z9 and use of LPARs, PR/SM, z/VM, z/OS, z/Linux; AIX with dynamic LPARs and micropartitions; and Linux pSeries and iSeries (OS/400, Linux)
- HP with Superdomes and a wide range of servers using nPars (logical) and vPars (virtual) and secure resource partitions
- Sun with UltraSparc or Intel/AMD servers and dynamic system domains, logical system domains and global zones with containers
- For Windows and Linux there are options from VMware (Virtual infrastructure), Microsoft (Hyper-V) and Open Source Xen

These all provide performance information on the VMs from their management systems.

The key elements of virtualization are often described as partitioning, isolation, encapsulation and hardware independence. Once the solution is adopted, it becomes an enabler for other features. Dynamic relocation of virtual machines across servers opens a route for dynamic resource scheduling (workload balancing), high availability (standby-nonstop) and consolidated backup (off-line).

In order to decide on the right mix of physical and virtual approaches, it is worth assessing where the market share has taken off. This is primarily in the consolidation of lots of extremely lightly used x86 servers that are typically dedicated to a single application service and have been widely distributed as part of the local autonomy boom over the last decade or so. As the costs of under-utilized equipment and the costs of support, management and maintenance begin to dominate, so there is a move towards consolidation, which can be made much easier by using virtual machines to reflect each local setup.

The virtualization of x86 architecture has been accomplished in mixtures of three basic ways: full virtualization, paravirtualization and hardware assisted virtualization. The borders between them are ever-changing.

Paravirtualization offers important performance benefits, but also requires modification of the operating system source, which may impact application certifications. Hardware virtualization offers benefits in that the hardware itself supports some of the virtualization. Full virtualization relies on sophisticated software techniques to trap and virtualize the execution of certain sensitive, 'non-virtualizable' instructions in software via binary patching. With this approach, critical instructions are discovered at run-time and replaced with a trap into the VM to be emulated in software. These techniques incur significant performance overhead which becomes a problem in the area of system calls, interrupt virtualization, and frequent access to privileged resources.

For machines running database applications or with significant traffic or with different sets of users running different tasks, some issues involved with virtualization begin to cause concern. Primarily, the issue of overhead, which depends primarily on how much physical IO is done and how many kernel interrupts etc cause changes of state in the CPU. A pure CPU loading will be low on overhead, but a practical SQL Server or Oracle application will have a significant overhead such that users will notice significant degrading in performance on virtualization.

Virtualization is widely adopted for consolidating lightly used servers, separating development, test, quality assurance, pilot and production systems. In some cases, it provides the opportunity for hardware upgrades to outdated operating systems. It provides a vehicle for easy roll-out of cloned player systems with a complete environment suitable for training or marketing. It has managed to establish a great reputation for reliability, fiduciary and integrity. Experiences have grown rapidly, with many sites having initially virtualized and consolidated by a typical 'density' factor of 10:1 (ten VMs to one physical machine). These have often still yielded remarkably light usage so that higher consolidation ratios can be actioned, with more analysis of the nature of the workloads per VM. As the power of the physical servers grows, so the techniques of traditional capacity management apply, with VMs being treated as workload components, but still requiring characterization. Assessment of virtualization in practice has also led to review of many VMs and retirement of outdated or unused solutions.

2.6 Capacity management versus general ITSM process flow

Capacity management is described as a process in ITIL but in this book it is considered simply as a practice. Capacity management requires knowledge of, or at least awareness of, particular technical skills to implement it rather than just pure project definition and project management techniques as for some other ITIL processes. Although tools can reduce the need for some detailed knowledge, the use of the tools is best exercised with understanding. Each site has to find its own balance of members of the CMT and sharing information with its domain and application architects.

Also, CMP involves various processes described elsewhere in ITIL (e.g. incident management is used to handle capacity incidents, change management is used to change capacity infrastructure), and it requires understanding of the underpinning technology. This book uses CMP to discuss the activities, and CMT to describe the team (whether nominated as capacity managers or not; full time or part time involvement). This is expanded in appendix A for those who want to consider the relationships between processes, functions, and activities.

2.6.1 Data archiving and information life cycle management

An important aspect of capacity management is that of formal data archiving under the aegis of information life cycle management control. There are potentially huge amounts of data captured, collected and stored in the CDB. There is long term value in much of it, but it can be summarized. A typical regime will keep all the most detailed data which has been captured at five minute intervals or less for a limited time of days or weeks when the data is essentially of a performance tracing nature. Then it will be aggregated into fifteen minute or hourly averages and retained for weeks or months. Then it will be aggregated into daily averages and retained for months or years and possibly then weekly averages only retained.

Theory of Constraints and capacity management

Theory of Constraints (TOC)[18] is an overall management philosophy. It is geared to help organizations continually achieve their goals. A system's constraint is "that part of the system that constrains the objective of the system." Goldratt states that: "In reality any system has very few constraints and at the same time any system in reality must have at least one constraint." He extended this approach to describe how to optimize steps, applications and operations largely within a factory environment, although it can also be applied to IT. End-to-end capacity management requires minimizing the performance bottleneck in the service chain in the light of the relative leverage factor. For example, amending the structure of a database query may be more useful than adding additional processors.

2.6.2 Capacity management and service lifecycle

The CMP ranges across the lifecycle of a service. Other parts of the enterprise are involved to differing degrees at different stages. Clearly for new services involving in-house development, there is a need for liaison with development and testing functions to ensure performance requirements are considered. As a service is implemented, interfaces with the service desk need to be established to identify performance issues. Throughout there needs to be an understanding of the business needs and expectations, which can be enhanced by establishing the relationship with business demand and service activity by trend analysis of new users, dormant users and leavers.

This can be summarized as in figure 2.5, showing the partners involved, the range of liaisons and the stages of the service as it moves from 'plan' to 'develop' to 'deliver' to 'improve'.

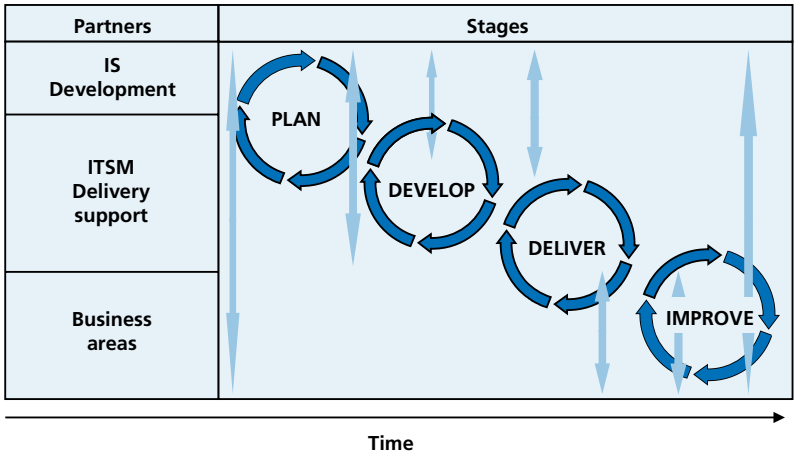


Figure 2.5 Iterative development of services

18 Introduced by Dr. Eliyahu M. Goldratt in his 1984 book entitled *The Goal.*

The essential points to notice in this figure is that each of the stages is iterative (and potentially with feedback loops from one to the previous). The partners involved at each stage tend to have differing levels of interest, but the key point is that all should be kept aware of each others' issues throughout the life cycle.

The lifecycle of a service can also be summarized as in figure 2.6. This shows the sweep of activities from development to 'productization' to service delivery.

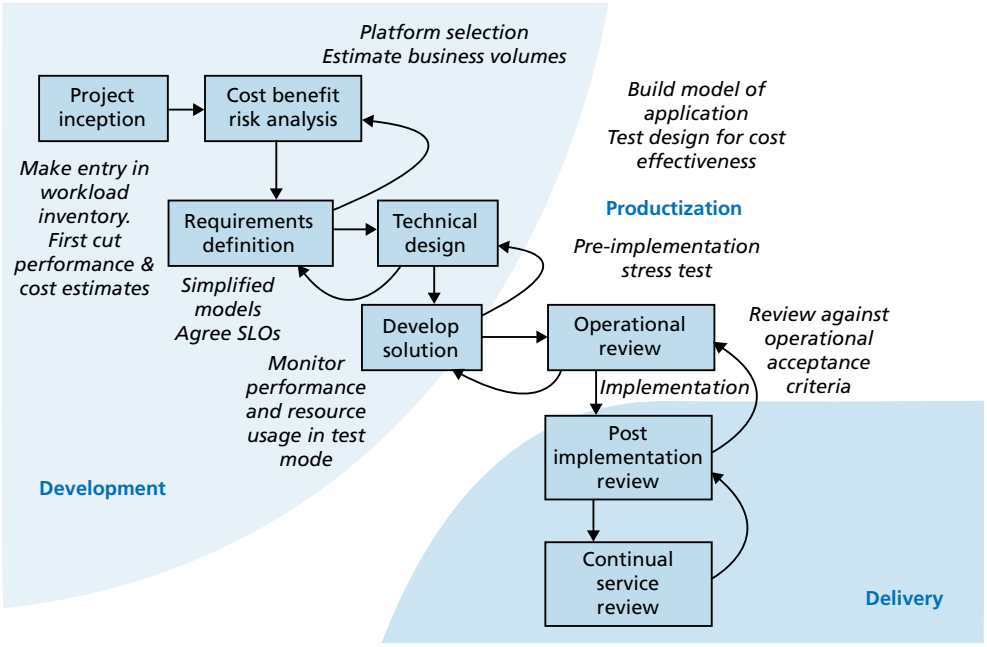


Figure 2.6 Service development lifecycle model

The iterative cycles within each stage are indicated as is the general flow from development through testing to delivery, with some of the main capacity management actions indicated.

The relationship between development and the service life cycle model on one hand and the IT infrastructure and ITSM on the other is a key issue in the establishment of the CMP. What is needed is a sensible balance between projects and practices.

Equally, there is a need for some use of forces to do work, requiring energy and generating momentum to effect some actions. These terms have all been described within mathematics and engineering for a long time, with the initial derivation of their relationships defined by Isaac Newton[19].

19 Newton's *Philosophiæ Naturalis Principia Mathematica* written in 1687 derived most of the laws of motion.

Various mathematical laws involved with capacity management will be introduced later. In order to set the scene, a few other laws are reviewed. Newton's laws of mechanics can be related to capacity management.

The first law states that things stay restful unless there is an external force applied to it (or in ITSM an organization can become stagnant without some catalyst such as a consultant for change).

The second law states that the rate of change of momentum of a body is proportional to the impulse applied to it (so that the larger the enterprise, the greater the force required to make a change to ITSM and the longer it takes, thus suggesting that long assignments of large capacity management evangelists are recommended).

The third law deals with the balance between equal and opposite forces, typically expressed in terms of actions and reactions (which encourages ITSM enthusiasts to find ways of working with those less keen, rather than against them; so that the deliverables are seen to be to the benefit of the organization).

3 Why: Benefits of capacity management

There are many benefits in adopting, implementing and enhancing effective capacity management. This chapter identifies some of them, as well as the costs and a few cost-benefit analyses and risk considerations.

3.1 Primary benefits

The primary benefit of capacity management is to provide the optimum balance of resource to meet the demands of business at a cost that ensures the appropriate service levels. This can be extended to a more complete checklist of primary benefits:

- improved quality service provision and cost justifiable service quality
- services that meet business, customer and user demands via integrated processes
- increased efficiency – cost savings
- deferred upgrade expenditure (cash flow)
- effective consolidation and virtualization (maintenance/licenses)
- reduced impact on the environment, by optimizing the IT infrastructure
- reduced costs of accommodation, energy consumption, raw materials
- planned acquisition (allowing time to negotiate discounts), introduction of new technology and meeting sustainability requirements
- reduced risk of poor performance
- fewer capacity/performance disasters (appreciated most by customers)
- fewer capacity/performance problems (appreciated most by users)
- more confident forecasts of performance
- application lifecycle enhanced (fewer performance disasters/abandoned apps)
- learning from previous experience with demonstrable KPIs

Each site will weigh these benefits differently, in the light of local circumstances, objectives and (bitter) experience.

The related capacity management activities are intended to provide these benefits by helping to:

- predict when the IT services will all fall apart leading to performance disasters and abandoned applications
- take action to avoid potential performance disasters in time
- provide early feedback as applications evolve or degrade
- have the right equipment in the right place at the right time
- avoid wasting money on redundant or unsupported equipment
- consolidate and virtualize servers effectively
- avoid wasting time on un-necessary interim upgrades or pointless tuning
- do it right first time
- optimize the IT infrastructure by studying new technology developments

In essence, all of the activities are designed to enable proactive performance and capacity management and all the benefits that such assurance brings.

In summary, capacity management is a 'Good Thing'. This is a reference to a classic slim volume entitled "1066 and all that" in which Sellar and Yeatman summarized 2000 years of confused schoolboy memories of English history into 124 pages. They also said that 'History is not what you thought. It is what you can remember".
This might well also be applied to ITIL capacity management, where ITIL summarizes many person-years of capacity management experience into 50 pages. Sellar and Yeatman sub-titled their book as "A Memorable History of England, comprising all the parts you can remember, including 103 Good Things and 5 Bad Kings". Perhaps Knot ITIL (Appendix E) describes "all the parts of ITSM you can remember, with 6 Good Things (core-processes) and 5 Required Teams (functions)".

3.2 Operational benefits

The primary benefits listed above may all have an effect on operations but specifically the main operational benefits lie in having a team dedicated to performing effective capacity management and so enable specific benefits to operations such as:

- more efficiency in operations as less capacity and performance issues
- less diagnostic skills required as the CMT address capacity issues proactively
- standardization benefits of demands, priorities and alerts
- less intervention required due to proactive capacity management
- less out-of-hours costs to cope with capacity or performance crises
- common tool means better exploitation and/or fewer bodies needed to drive it
- 'radar' to warn against possible potential problems
- smaller or greener footprint and less space or heat requirements with fewer servers
- smoother transition of new technology and better tuning of options
- improved awareness of 'out of support' issues and resultant planned actions

Operations and capacity management will share a lot of information and should establish a relationship that will lead to more effective ITSM.

3.3 Management benefits

Similarly, from a management viewpoint, the main benefits are probably seen as having effective capacity management from a governance point of view, but also explicitly:

- trustworthy IT services
  - customer trust directly reduces costs:
  - procedures can be simplified
  - budget can be allocated functionally rather than against unknown contingencies
- more control
- less degradation
- better information for decisions
- improved working relationships between teams (e.g. applications support, database support, network support, server support) as end-to-end capacity management requires good communication on performance and capacity between different functions
- benefits of more maturity, such as agility and business – IT alignment

In order to achieve all these benefits, management will have to liaise effectively with the CMT rather than assume it provides a merely technical function.

3.4 Business benefits

The business benefits will vary in interpretation in each site even more than the operational and management benefits but can be thought of under the following headings:

- cost reduction or control overall (despite cost of reporting tools)
- performance improvement
- flexibility and agility
- improvement to management of suppliers and delivery chains

The benefits of capacity management are most dramatically demonstrated when the discipline is not in place and is subsequently introduced. Sites have reported huge initial savings by objective assessment of proposed upgrades leading to the realization that, for example, 80% of the benefit of a proposed upgrade may be achieved by an alternative option at 20% of the cost. This is the traditional justification for the CMP and applies best to the case of a significant upgrade to a large machine costing a lot of money. However, it applies across all machines to some degree. The total cost of IT for most enterprises does not drop as the unit cost of machines drop. So the total expenditure remains the same or increases and the leverage in getting the optimum performance from those machines remains valid.

3.5 Costs

A typical set of cost headings is difficult to generate as budget headings and financial controls are so site specific, but a core checklist is shown below:

- procurement of required tools
- monitoring hardware, OS, applications
- CDB/CMIS for holding record of all capacity management data
- modeling tools for 'what-if' and statistics
- graphical reporting tools (web-enabled)
- project management as required
- staff costs including recruitment, training, accommodation et cetera
- development liaison and data flows
- QA load, soak and saturation testing data
- business liaison and attitude questionnaires

Each site will also find its own level of maturity for all of the practices discussed so that not all of the above costs are necessarily incurred in establishing capacity management.

3.6 Cost-benefit analysis

The current direct cost of a formal CMP within an organization is usually difficult to establish. Work undertaken in this area is often carried out as a by-product of other functions as well as the CMT. Any general time recording systems in use could only give figures for current activity in these areas if they have been identified as separate tasks. In sites and data centers where there are specific resources dedicated to CMP there are often limited levels of commitment and adoption.

Ultimately, the purpose of the CMP is to provide sufficient capacity to satisfy the needs of the business. The need is to match the computing infrastructure to the needs of the applications to achieve the business requirements. This discipline is increasingly needed, despite hardware price-performance improvements as the aspirations of end-users and resource demands of application software are increasing even more rapidly.

CMP has become highly visible in most corporations as e-commerce response times affect business transactions and the service level is manifested as web-based transactions. Equally, although modern severs may introduce smaller incremental upgrades and shorter time-scales for procurement, it still merits accurate sizing of demands for new procurements. New technology and how it is to be implemented to achieve sustainability is part of the challenge. The open competition between suppliers should allow even more benefit to be obtained from predictions of the likely impact of alternative procurement solutions. Also, the potential disruption and resource costs of every upgrade need to be reviewed before action.

Cost-benefit analysis is notoriously difficult in any discipline which is essentially directed towards providing an insurance against disasters, and offering a guideline for good practice and efficient operation.

The savings as a result of introducing performance assurance are not readily quantified on the basis of a 'per application' or 'per configuration' basis, as the previous efficiency of matching business needs to resources will vary widely from one site to another. Certainly, once established, few sites abandon the practice of formalized capacity management which is now widely adopted. The following notes indicate some of the factors in making financial estimates:

- **Performance prediction** – there are some extra costs per project as the procedures will be a new requirement imposed on the developers. The savings gained by performance engineering in terms of reduced development work in rewrites and maintenance are significant. There is also a major saving in avoiding corrective software development projects to resolve performance problems.
- **Performance testing** – depending on the level of previous procedures, some extra costs may be incurred due to the increased workload placed on QA conformance testing. Performance testing will extend the formality of benchmarking and may well incur extra costs in analysis and prediction. The savings are achieved through the earlier identification of problems and often substantial savings are possible because improved diagnosis and tools find errors and excessive consumption earlier in development when changes are cheapest to fix.
- **Performance assurance** – this area potentially adds extra costs in that a new discipline may have to be introduced at each service delivery site or data center. The costs of the staffing

and tools are highly variable depending on the extent of adoption. A 'typical' large computer centre might have a team of some two to five planners, with special areas of responsibility and expertise. The savings lie in the controlled acquisition of necessary upgrades.

- **CMP audit** – this is often a new activity requiring specialist expertise to liaise with all the other activities. The basic cost of the central function would be that of establishing the team, training it and providing appropriate tools. The savings lie in the integration of all the activities and the overall benefits of ITSM. The demand comes from a need to demonstrate compliance with various requirements

The costs identified above may turn out to be significant. But no site can survive without CMP, whether formalized or not, as no modern town can do without a (reactive) fire engine and (proactive) water hydrants to supply it. Formalizing the practice allows management to hold it to account and forces it to perform in the best interests of the company. Further, the savings and benefits in both financial and service terms will typically far outweigh those infrastructure costs.

Experience of introducing CMP into large organizations indicates that over the first few years of introduction there is likely to be a minimum average net saving in the order of 5%-10% on capital expenditure and operating costs, plus an unquantifiable (but possibly more significant) saving where systems have not been delayed (or cancelled) due to inadequate resourcing[20].

> "There is nothing more difficult to execute, nor more dubious of success, nor more dangerous to administer, than to introduce a new system of things: for he who introduces it has all those who profit from the old system as his enemies and he has only lukewarm allies in all those who might profit from the new system."
> Nicolo Machiavelli, The Prince, VI, 94.

In practice, some quantifiable data is available from a number of case studies. For example, the author worked at one site (that prefers to be anonymous for obvious reasons) which saved a million pounds on a proposed upgrade (on the wrong devices) which would have had insignificant impact on performance. And another site saved the costs of establishing a new performance assurance function in its first study. But these are difficult cases to refer to, as they imply criticism of those proposing upgrades before they are properly assessed. And, like keeping fit, the benefits of good practices are sometimes only demonstrable by stopping doing them.

Further, a major benefit lies in the guarantee of a known and controlled IT service, with the impact of proposed changes to workload or configuration being formally sized. In this way, the discipline provides insurance that performance disasters due to bad design or under-resourcing can be avoided at the same time as not overspending on equipment.

The unsurprising overall conclusion of this book is that it is feasible, desirable and cost-effective to establish a fully integrated CMP environment. The implementation, managing and optimization of the practice is considered in chapters 5-8, but key ingredients are establishing the right practice owner, budget champion and discipline evangelist. The last will be core members

20 See the papers referred to in 'Sources', section 'Useful Papers'.

of the CMT and will probably set the tone of the team in many ways, such as being energetic, good communicators, out-ward looking, business aware and altogether key to the business and IT relationship.

Again, paving the way for various laws of capacity management, consider those for thermodynamics. The three laws of thermodynamics can also be related to capacity management. Capacity management can use its energy to instigate better control of the infrastructure and improve insight in the relations between business demand, IT services and resources. To achieve this, the team needs to be good communicators and well energized.

The principle of the conservation of energy (the first law of thermodynamics) states that "The energy of a closed system remains constant during any process". So an enterprise considering CMP and related changes, needs to be open to change and establish an energetic CMT.

The second law is the conservation of entropy. Entropy is a measure of disorder, the more liquid the less ordered and the more entropy (for example, consider a glass of some alcohol with some melting ice). A pub can be viewed as an entropy pool. Perhaps what is needed to energize a lethargic CMT is the right amount of the right liquid to optimize entropy in the system. This may need a bit of practice.

The third law states that as the temperature of a body approaches absolute zero, so its entropy approaches zero. A restaurant serving a good chilled wine with seafood is also an entropy pool. Again, enough of the right liquid served at the right temperature can ensure a well oiled team.