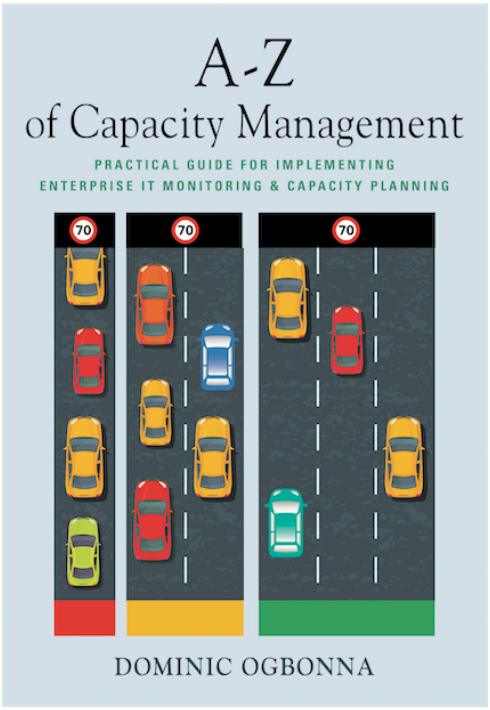


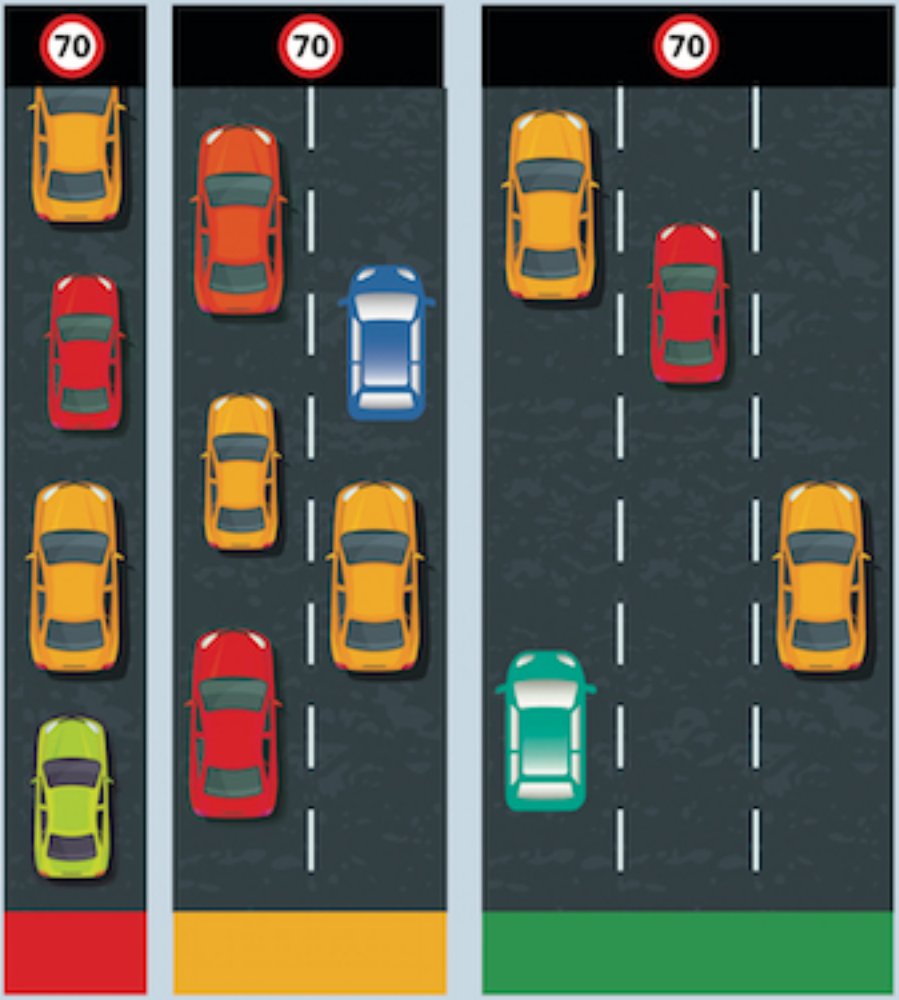
Most often we are told the "what and why" of capacity management, but not how to make it happen. This book provides good practical approach on how to implement the process, with a view to bringing its benefits to the organization. Capacity management is incomplete without business driven capacity planning.

A-Z of Capacity Management: Practical Guide for Implementing Enterprise IT Monitoring & Capacity Planning

by Dominic Ogbonna

A-Z of Capacity Management

PRACTICAL GUIDE FOR IMPLEMENTING
ENTERPRISE IT MONITORING & CAPACITY PLANNING



DOMINIC OGBONNA

ISBN 978-1-63492-757-4

Published by BookLocker.com, Inc., St. Petersburg, Florida.

Printed on acid-free paper.

Booklocker.com, Inc.
2017

First Edition

TABLE OF CONTENTS

FOREWORD

Capacity Management is a bedrock of application stability. Collecting the appropriate metrics, understanding the data and reacting in a timely manner can avoid outages or at least reduce the time taken to fix an issue. This book will teach your organization how to ensure metrics are designed from the beginning of the Software Development Lifecycle and deployed to production to collect the necessary insight into the environment.

One of the common misnomers for capacity management is all metrics are driven from the infrastructure. Dominic provides critical insight into how understanding the business needs and capturing the expected profiles drives teams to make better decisions. For example, a set of servers running at100% CPU utilization might look like a problem - unless those same servers are the batch processing systems or grid computing nodes designed for computational analysis. In this case, capacity decisions would be driven by business expansion or volume/timing and the available headroom left on the servers to run more computations.

Dominic has written a thoughtful and detailed book on planning capacity metrics to provide insightful views to your business cases. He provides both the theory of the capacity management processes and the practical implementation that is so often overlooked - using real-world examples and the specific commands. Having seen Dominic's suggestions in practice, the techniques are invaluable to the organization implementing them.

Megan Restuccia
Former Executive Director at Morgan Stanley

PREFACE

WHY I WROTE THIS BOOK

Most often we are told the "**what and why"** of capacity management, but not **how** to make it happen. This book provides good practical approach on **how** to implement the capacity management process, with a view to bringing its benefits to the organization.

The subject of capacity management is now treated like a theoretical process, and as an oral tradition which has lost its right content. It is now common knowledge that a good number of trainers in this process do not understand the basic concept, how much more offering a practical insight.

This book provides a detailed guideline for practical implementation of the capacity management process, with a view to demystify the management process; a move from theory to practice - using a simple capacity management model that can fit into organizations of any size.

Repeatedly, I have seen individuals and organizations very keen on implementing capacity management correctly, but inadvertently, they end up doing it wrong because, traditionally, the focus is just on monitoring and alerting based on the host server resources usage - CPU utilization, Memory utilization, etc. This book seeks to clarify the process and expose the readers to a simplified way to doing it right, while adding value to the organization through capacity management process.

The full benefit of implementing the capacity management process is usually harnessed when it operates as a value-added process within the organization. This will be possible when the process maturity level is well above average.

Capacity management can hardly be accomplished solely by the capacity analysts/manager working alone; to get its implementation right, this book outlines both the technical and business stakeholders that should be involved. It also contains questions you should ask regarding most IT service/application to ensure you are monitoring the right business and service data.

WHO SHOULD READ THIS BOOK

This book is for anyone who wants to have an in-depth knowledge of how to implement capacity management process in an organization, and those whose functions or services involve mitigating business risk associated with IT service failure.

From a technology point of view, CIOs, CTOs, capacity managers, capacity analyst, capacity planners, business and technical service owners, IT operations managers, service managers , IT and business consultants, IT auditors and risk officers, operation engineers, business managers, senior business managers, application architects and developers, infrastructure support analysts, etc. will find this book very insightful, and useful.

Furthermore, CEO, and senior business leaders who are interested in delivering excellent service to customers, but with focus on reducing IT Infrastructure spending will also find this book very rewarding.

BOOK ORGANIZATION

This book is organised into chapters based on the capacity management process model diagram, with each chapter describes how to practically implement the process. In addition, each chapter has conclusion, success hints and organizational appraisal questions; that are designed to help the reader evaluate the process implementation within their organization.

Chapter 1, What is Capacity Management, introduces capacity management, its goals and benefits, and the need to use business data to drive capacity planning rather than basing it on infrastructure resource usage.

Chapter 2, Capacity Management Strategy, dwells on the guidelines for putting the proper policy and procedure in place to drive the capacity management process.

Chapter 3, Capacity Management Gap Analysis & KPIS, provides an overview of what to look out for when assessing the current state of the process, and how to measure the success of the capacity management process.

Chapter 4, Monitoring and Resource Data Collection, provides ways to go about collecting system resource performance data.

Chapter 5, Business Metrics Data Collection Techniques, provides deep dives for business metric instrumentation, and how to determine the business metrics capacity limits.

Chapter 6, Data Aggregation Methods & Granularity, provides guidelines for transforming the collected metric data to meet the capacity management needs.

Chapter 7, Capacity Database (CDB) & Data Storage Techniques, provides the information that will help towards building scalable and high performance CDB/CMIS.

Chapter 8, Capacity Reports, looks at the report audience and what they need to know.

Chapter 9, Capacity Planning, gives in depth guidelines on how to get started with capacity planning, and the basic inputs and tools required.

Chapter 10, Building Analytical Capacity Planning Model, an extensive guideline for building capacity planning models, and a step by step description of a sample model.

Chapter 11, Capacity Planning Review with Stakeholders, covers how to get the business users' co-operation in the capacity planning.

Chapter 12, The Capacity Plan, guide to writing formal capacity plan document.

Chapter 13, Capacity Threshold Alerting & Response Types, introduces effective threshold breach management.

Chapter 14, Cloud Computing Capacity Management, reviews the place of capacity management in the cloud computing and machine learning era.

Chapter 15, Auditing the Capacity Management Process, how to ensure the capacity management process is kept on track and fit for purpose.

Appendix A, UNIX Server Performance Data Collection Techniques, focuses on vmstat, sar, and iostat; and turning their output to csv format.

Appendix B, Windows Server Performance Data Collection Techniques, overview of Logman.exe for performance metrics collection

CONVENTIONS USED IN THIS BOOK

The following typographical conventions are used in this book:

Italic

Indicates quotes from people, formulae, command-line options

Italic Bold

Indicates text that should be replace by the user with the appropriate values

1 CAPACITY MANAGEMENT - ISSUES, GOALS, AND BENEFITS

"Facts do not cease to exist because they are ignored"

– Aldous Huxley.

INTRODUCTION

Capacity management is the information technology risk management process for ensuring there is adequate infrastructure and computing resources to meet the current and future demand of the business in a cost effective and timely manner. This management process primarily seeks to proactively ensure that applications and infrastructures have the ability to provide the resources required to meet the organization's current and future business demand needs in a cost-effective and timely manner. Capacity management is also a risk management technique for ensuring that an IT service meets SLA target in a cost effective and timely manner.

It is one of the processes defined in the Information Technology Infrastructure Library® (ITIL®) framework, and belongs to the Service Design phase of service lifecycle. Within an organization, the maturity level for implementing capacity management can vary for different IT services used by the business depending on their criticality to the business.

A desired maturity level is where the capacity management process can be proactively applied to support the business' current and future demand without

reacting or fire-fighting to restore IT service outage or performance degradation arising from inadequate IT resources to cope with the business demand. This implies that to attain this maturity level, capacity planning will not only be driven by the current utilization of the IT infrastructure resources, but also by how the future demand of the business will affect the infrastructure resources utilization. Resultantly, at this level of capacity management maturity level, capacity is represented using terms that the business users understand, and not technical jargons.

Capacity management is not only about having adequate infrastructure resources for business, it is also about right-sizing and cost-savings; by ensuring that excess capacity provisions are detected and retracted.

Having a good capacity management process in place is not an antidote for preventing IT service incidents, because IT service outage or performance degradation could arise from other sources - human, coding, or IT change management errors, etc. As a result, the capacity management process key-performance-indicators (KPIs) should be based on eliminating incidents with capacity risk as the root cause.

BENEFITS OF CAPACITY MANAGEMENT

Capacity management brings about the goal of right-sizing the application and infrastructure resources, by aligning the current and future business demand at the right cost. There are several other benefits associated with it when correctly implemented, amongst these are:

- Makes it easy to transfer infrastructure and computing resources from places of excess capacity to where needed, without additional spend
- Capacity management helps in the development of the application performance testing function within an organization
- Increases customer or end-user experience satisfaction, loyalty, and retention
- Provides data needed for incident investigation, and problem root cause analysis.

HOW DOES CAPACITY MANAGEMENT NEED ARISE?

The need for capacity management has arisen because IT infrastructure and computing resources are limited in supply; increasing these resources will usually involve the organization parting with money. In contrast to the limited resources, the demand for them increases as the business grows. As a result, capacity management deals with **balancing** IT infrastructure, computing, and processing resources along:

- Cost of getting resources versus resource capacity available
- Supply by IT providers versus demand by business users.

These are further illustrated below.

Practical illustration:

You have an application that can allow up to 1000 online buyers logins at the same time, and operates with performance service level agreement of login average completion time of within 3 seconds. The cost of upgrading the application to

maintain the stated service level agreement is $5000 for each additional 100 buyer-logins.

Cost versus Capacity scenario

- If the current capacity is 1000 (maximum logins)
- Any attempt to increase the concurrent logins beyond the current capacity of 1000 will require spending more money - a cost to the business
- Increasing IT processing capacity always has cost implication for the organization. The cost could be both fixed and recurring.

Supply versus Demand scenario

- If the current capacity is 1000
- If the monthly peak concurrent buyer-logins over the last 6 months is 200, it will make sense to reduce the capacity, and increase it as demand increases. (this will not only save the business money, it may also reduce other licensing costs associated with this application)
- If on the other hand, the business as a result of new marketing plan informs you that in 6 months time, it expects the buyer-logins to increase by additional 500. At this point, you will need to increase the capacity to accommodate the expected demand increase, however, you need to carry out a planned upgrade close to the expected demand increase
- Increase or decrease in business demand of an IT service should translate to IT infrastructure resources supply upgrade or downgrade respectively.

COMMON ERRORS IN IMPLEMENTING THE CAPACITY MANGEMENMT PROCESS

There are some common mistakes often made while implementing the capacity management process they should be avoided if you desire to get the full benefits of the process.

- No single capacity planning model will be a fit-for-all applications or systems
- For infrastructure resource utilization or service latency, the maximum aggregated value is good for monitoring the system heart beat and incident investigation. However, it is not good for capacity planning. The resource utilization spikes could come from system panic, bad database queries, system command, application bug, or other unexpected sources. As a result, using the maximum aggregation method, a single or transient spike in resource utilization will erroneously be taken as the value of the periodic data set, rather than getting the peak utilization incurred over a sustained time interval. This will lead to infrastructure over provisioning or excess capacity which is a cost to the organization
- Like above, the average value is also wrong because it obscures the real high utilizations over the period interval. This will lead to infrastructure under provisioning or inadequate capacity which is a cost to the organization
- Modelling infrastructure resource utilization, for example, 'Total CPU utilization' using trend line will lead to inaccurate planning, because such resources' response time will no longer operate linearly once the CPU is overloaded
- Capacity planning based on only Infrastructure resources utilization may never be representative of the business volumes and throughput driving your infrastructure capacity, and will lead to inaccurate planning
- Business capacity metrics without performance measurements, (throughput and latency), will not be able to provide the needed end-user perception of the IT service
- Capacity Planning should be carried out based on peak trading period metrics
- In data collection the focus should be on measuring resource used, and not resource available. Capacity management is about reporting, and planning based on resource utilization
- Collecting resource data for which there is no known capacity limit or specified available maximum capacity will not be useful in capacity

CAPACITY MANAGEMENT VERSUS CAPACITY PLANNING

Usually the terms 'capacity management' and 'capacity planning' are used interchangeably, this is not right. In summary, capacity planning is a subset or component of capacity management. Capacity plan is the output from the capacity planning component; and implementing the capacity plan is the end product of proactive capacity management.

Capacity planning is usually done in consultation with the business users/representatives that provide their business demand forecasts as an input to the process; which in turn predicts the infrastructure requirement to meet the future business demand.

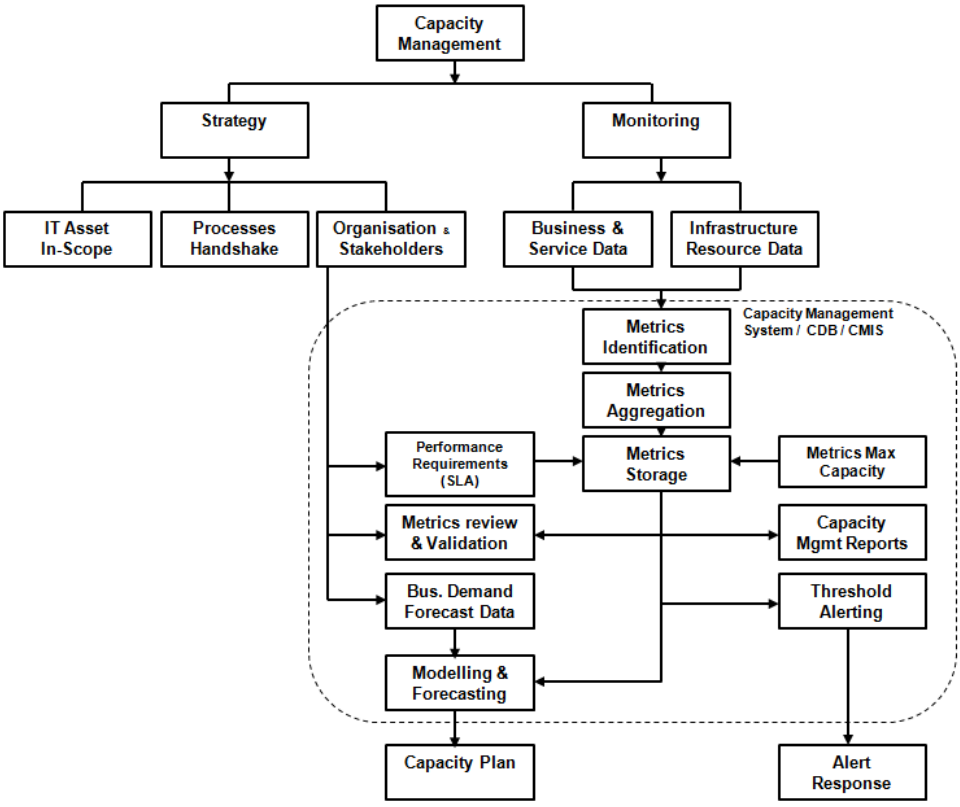
Capacity management adds value to organizations when it can proactively help mitigate service performance degradation or outage relating to inadequate infrastructure resources.

At the lower maturity level of capacity management, there may not be explicit capacity planning process in place, rather the IT support team relies on infrastructure utilization threshold alerting, and users' IT-service-failure complaints. This approach is reactive, and leads to fire fighting for service restoration.

COMPONENTS OF CAPACITY MANAGEMENT

Capacity management process like any other information technology management process has building blocks, or components. The key components of the capacity management process are shown below in Figure 1.1 (Dominic's model of the capacity management process diagram); each of the components will be

Figure 1.1 Capacity Management Process Diagram - Dominic's Model



Each of these components or group of components is discussed in detail in the subsequent chapters.

SUMMARY

The capacity management process is an IT service risk management technique, which should be given adequate attention, so as to ensure that service failure or performance degradation arising from inadequate infrastructure resources is proactively and cost effectively eliminated. This can happen when:

- In this digital age, customers can easily switch service, capacity management should be seen as a business enabler, not only as cost centre
- Organizations should embrace the capacity management process, and give it the needed senior management support.
- Capacity management should not be seen as just reacting and fixing capacity issues arising from infrastructure resource threshold breach alerting. It should be proactive, by focusing on the business capacity drivers which are causing the resources usage to increase.
- The capacity management process is incomplete without the capacity planning component, which makes the process proactive.
- The capacity management process is also aimed at reducing the cost of doing business, by eliminating excess infrastructure provisioning, licenses, and other associated costs.

ORGANIZATIONAL APPRAISAL

1. Is capacity management process implemented in your organization?
2. Is the capacity management implementation yielding the expected benefits?
3. Is capacity planning part of your capacity management process activities?
4. Is your capacity management process determined by business volumetric or just system resource utilizations?

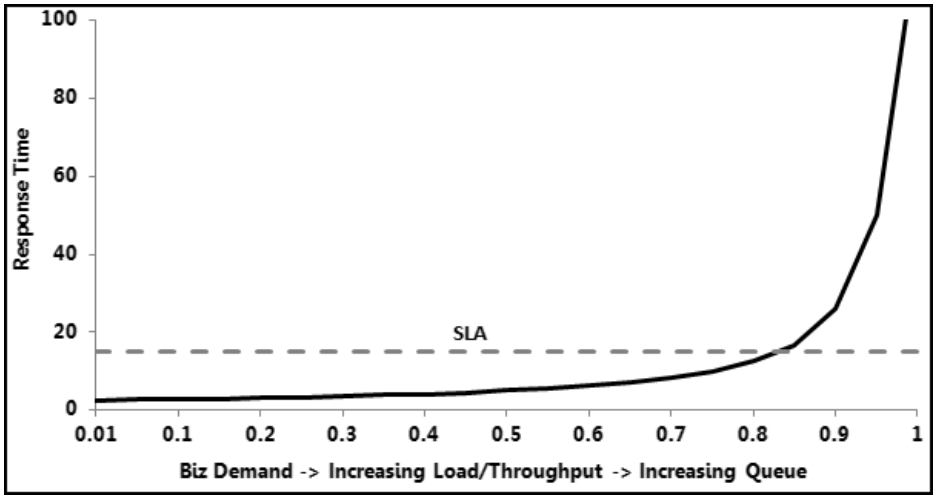
capacity management - proactive capacity planning (to predict future infrastructure resources needed to support the business demand forecast), the three sub processes should be used together in a capacity model.

APPLICATION PERFORMANCE ISSUE - THE ROOT CAUSE

The necessity for capacity planning has arisen to combat the performance problems that users experience when using an IT system/application, which sometimes makes such application unusable. It is important to know what gives rise to this performance problem.

Figure 4.2, shows the typical response of infrastructure resource or device to concurrency request types, one example of such device is the CPU. This resource response pattern is what is eventually translated to the external application response, which the end-users experience.

Figure 4.2 CPU Response Time vs. Load/Throughput/Queue Behaviour



Based on Figure 4.2, it can be seen that:

- The Response time increases as the users' requests increase (e.g. more users requesting a system resource at the same time)
- More user demand leads to additional load, and additional load leads to increased request throughput
- As the throughput increases such that the device cannot respond to all request, the request queue gets increased
- As the queue continue to increase, the device may become unresponsive; and at this point, users will wait almost indefinitely (this state is popularly described by users as 'the system is hanging')
- SLA is breached when the response time stops responding linearly
- One key objective of the capacity management process is to ensure that there are adequate resources to share the load, so that the application will always operate within the SLA acceptance region.

COMMON CAPACITY MANAGEMENT METRICS

There are common metrics associated with each of the three sub processes of capacity management. In this section, we will focus more on the resource and service metrics that are generic. The business sub process metrics are not generic in nature, but vary for each IT application; therefore, detailed guideline will be provided for it in chapter four.

Table 4.3 Common Capacity Management

Sub Processes	Metrics Type	Common Metrics
Resource / Infrastructure	**Host/Server**	CPU Utilization (Total), CPU Utilization (User), CPU Utilization (System), CPU Utilization (IOWAIT), Queue Length, Memory utilization,

5 BUSINESS METRICS DATA COLLECTION TECHNIQUES

"The secret of getting ahead is getting started"

- Mark Twain.

INTRODUCTION

In the previous chapter, business metrics was briefly introduced and explained, nonetheless, this chapter is dedicated to further discussing this very important and pivotal component of business capacity management.

For an organization's capacity management maturity level to operate in the "value" position, her forecast of IT infrastructure and computing requirements should be driven by the following:

- The ability to measure and express the actual business activities performed by an IT system in business terms

- The business representatives should be able to provide the business demand forecast in business terms

- Capacity reports meant for senior management, and business stakeholders must be expressed in business terms.

BUSINESS METRICS IDENTIFICATION TECHNIQUES

As part of the process of collecting business metrics, the application business owners or their representatives (IT system owners/managers/architects) should be consulted to ensure that the appropriate business metrics are collected. After all, the demand forecast data that is required for capacity planning will have to come from the application business owner.

Sometimes, the application business owners or their representatives may not be very clear about how to determine the most appropriate business metrics to be included for data collection. Guidance can be provided using the "question and answer" technique outlined below:

Table 5.2 Business Metrics Identifier Questions & Answers

Probe Question	Typical Answer	Typical Identification Analysis
1. What does the application or system do?	The application receives customer orders, sends out request messages and accepts uploaded files	The likely volumetric business metrics from the stated key end-user activities based on the provided answer are: Request messages, Uploaded files and Customer orders Note: The identified volumetric business metrics will also have service performance throughput metrics.
2. How do you determine	The time it takes to deliver the request	The likely performance metrics - response time or latency are:

7 CAPACITY DATABASE (CDB) & DATA STORAGE TECHNIQUES

"If you only have a hammer, you tend to see every problem as a nail"

- Abraham Maslow

INTRODUCTION

Capacity management depends essentially on collecting and analyzing data. Therefore, the process, success will depend largely on properly storing the data, and the ease of accessing, analyzing, reporting and extracting the data. The data storage component of the capacity management process system tool is identified by different names, amongst them are: Capacity Database (CDB), or Capacity Information management System (CMIS); we use the name tagged as CDB.

The method adopted in storing the metrics in the database plays a vital role in the usability, reliability, reporting flexibility, and scalability of the CDB system.

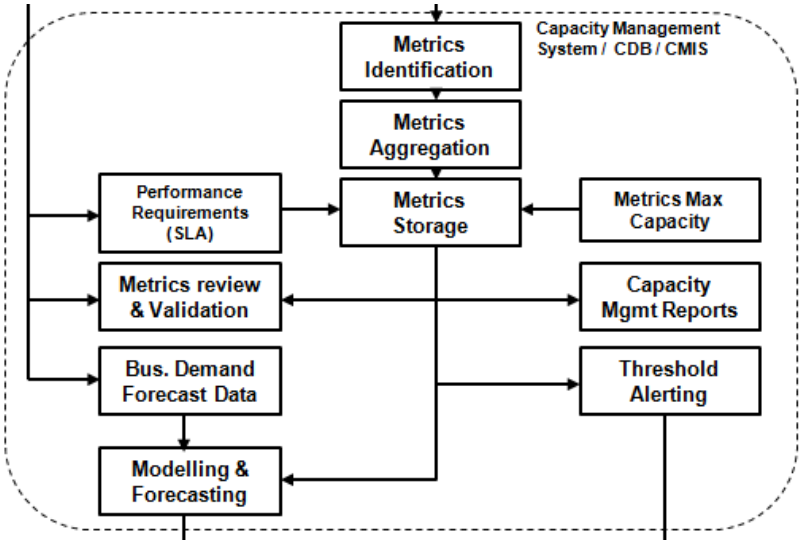
For organizations that choose to implement a commercial / third party solution, this may not be given serious attention; conversely, the reverse is the case for organizations building, and implementing an in-house solution.

IMPLEMENTING CAPACITY DATABASE (CDB/CMIS)

The Capacity Database (CDB) is an ITIL version 2 (in ITIL version 3, it is called capacity Management information System (CMIS)) term used to describe the data repository that holds capacity management process data. It includes but is not limited to business, service, and resources actual metrics from IT services in scope. Also stored in CDB are: metrics capacity limit, SLA, metrics' alert usage threshold, business forecast data, and modelling parameters.

Even though CDB by nomenclature is referred to as "database", it is just a repository which can be implemented using a relational database system (RDBMS), spreadsheet, no-SQL database, etc. Irrespective of the implementation, minimally, it should be able to support the generation of desired capacity reports by users. Figure 7.1, shows some components of the capacity management process model that can be embedded or implemented within a typical CDB.

Figure 7.1 Components of CDB / CMIS



For a base CDB system, the "Metrics Storage" functionality is pivotal, and the key input is the uniquely identified and mapped business, service, and resource metrics; while the minimum output is "Capacity Management reports". However, if proactive capacity planning is to happen, all CDB functionalities shown in Figure 7.1 and more are expected to be provided by a CDB system. Each component of CDB shown in Figure 7.1 is further discussed in other chapters.

CDB DATA AGGREGATION IMPLEMENTATTION STRATEGY: HARD VS SOFT

Usually, capacity management decisions are made using the aggregated values from data collected over a period of time. Such aggregated data are fed into the reporting, modelling, and alerting mechanism of the process.

The aggregated data can be generated and maintained in a CDB using any of the following techniques:

1. **Hard Aggregation**: Compute and store (all aggregated values are stored in the CDB)
2. **Soft Aggregation**: Full dynamic compute (No aggregated value is stored in the CDB)
3. Using the combination of both hard and soft techniques (some aggregated values are stored in the CDB)

For example: An application generates its trading log (containing timestamp, and number of transactions traded every second) , and the log is fed into CDB. The typical daily data points are analyzed thus:

a) Based on the per second log data for each day, a total of 86400 transactional metric data points / records will be stored in CDB repository

Table 7.3 (continued)

Feature	Hard Aggregation	Soft Aggregation
	desired feature.	

[1] Avoid storing average values as they pose a lot of problems maintaining their values, rather store and maintain "Sum" and "Count" aggregated values and use them to compute average value when required.

> ***Success Hint:***
> *For high volume data systems, performance is gained when intraday average values are computed from periodically stored "Sum" and "Count" aggregated values.*

CDB DATA REPOSITORY IMPLEMENTATION STRATEGY: SEPARATE VS. COMBINED

In designing a CDB system, consideration should be given to the appropriate repository technique that will be used to store the metrics data from the different IT services. This consideration is driven by:

- The number of IT services within the organization to be hosted in the CDB
- The frequency of metrics data upload to the CDB
- The volume of data coming from each IT service
- The scalability of the CDB system

There are mainly two types of metrics storage techniques that can be adopted for storing each IT service in a CDB; these are:

1. Separate metrics repository per IT service

9 CAPACITY PLANNING

"The person who says it cannot be done should not interrupt the person doing it"
– Chinese proverb

INTRODUCTION

Capacity planning is one of the key components of the capacity management process. The terms capacity planning, and capacity management are erroneously taken to mean the same thing, this is not right.

An organization's capacity management process is incomplete if the capacity planning aspect of the process is not undertaken. This is what is obtained in organizations where capacity management is taken to denote having different dashboards displaying infrastructure resource usage, and business volumetric.

Capacity Planning is about forecasting the computing and infrastructure resources that will be required to meet the business demand in the future. This is where the business, service, and resource metrics collected are translated into future business demand planning.

Capacity planning is driven by the business demand forecast information (data) provided by the business users or their representative about their anticipated increase/decrease in business activities or users.

Capacity planning involves the following main steps:

1. **Inputs:** Gathering information/data from the various capacity planning Inputs (discussed below, and shown in Table 9.1)
2. **Model:** Using the information/data from the inputs, build the capacity planning model; to forecast the resource requirement corresponding to the business demand forecast's input information/data.
3. **Output:** Using the capacity planning model, produce the **capacity plan** (the main output of the capacity management process).

CAPACITY PLANNING INPUTS

Table 9.1, lists and describes the various input required for capacity planning.

Table 9.1 Capacity Planning Inputs

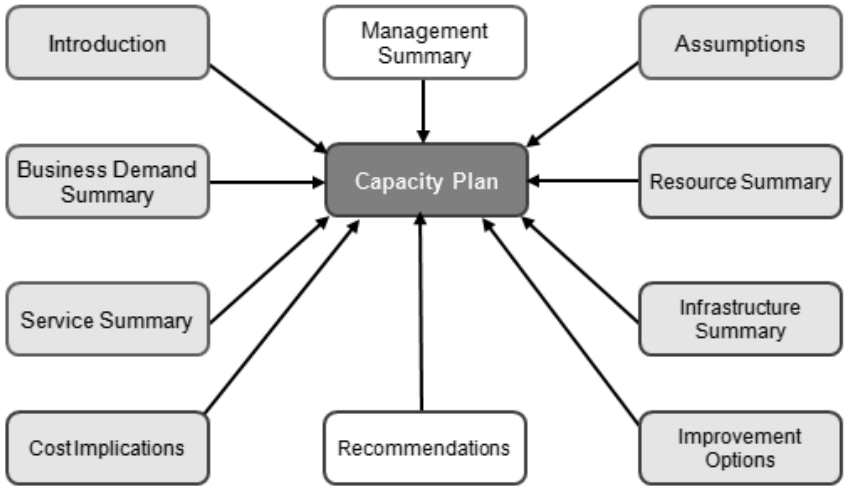
Input Name	Input Description	Input Source
Business demand forecast data / information	• The business plan of the organization should be translated to measurable future business volumes for the IT service capacity planning it is being carried out for. Some examples of business demand forecast information/data are: o The monthly daily trade volume is expected to increase by 30% in the next one year. o For the next 6 months, fifty thousand new users will be	Provided by the business user, using the identified principal business metric that drives capacity.

Irrespective of the content format, based on good practice, certain key information is expected in the document.

> ***Success Hint:***
> *To help drive the quality of the information in a capacity plan, it is good that the organization adopts a capacity-plan-template. The template will help ensure; that the capacity plan for all IT services is consistent, and uniform.*

Figure 12.1, shows the various key informational categories expected in a capacity plan; each is discussed further in detail.

Figure 12.1 Components of an IT Service capacity plan



APPENDIX A – UNIX: PERFORMANCE DATA COLLECTION TECHNIQUES

INTRODUCTION

The Unix operating system and its variants e.g. Linux has an inbuilt utility, for performance monitoring of the provided resources. Amongst the Unix resources monitored are CPU, memory, disk space, input/output time, network time, etc.

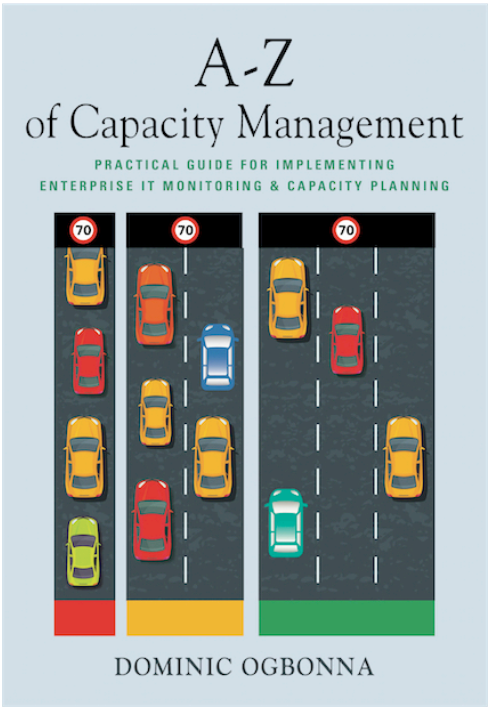
Unix and its variants provide many command-line utilities for monitoring and collecting the performance of the various resource types. In this book only vmstat, sar, and iostat will be discussed.

VMSTAT UTILITY

Description

- vmstat reports information processes, memory, paging, block I/O, traps and CPU activities
- The reports are intended to help identify system bottlenecks. Also, note that Linux vmstat does not count itself as a running process.

Manual



Most often we are told the "what and why" of capacity management, but not how to make it happen. This book provides good practical approach on how to implement the process, with a view to bringing its benefits to the organization. Capacity management is incomplete without business driven capacity planning.

A-Z of Capacity Management: Practical Guide for Implementing Enterprise IT Monitoring & Capacity Planning

by Dominic Ogbonna