Capacity Management

Standard Practises Document

GAVS Technologies N.A., Inc

116 Village Blvd, Suite 200, Princeton,

New Jersey 08540, USA.

http://www.gavstech.com/

	Designation	Name
Prepared By	Lead Executive – Operations & Strategy Excellence	Gouri Mahendru
Reviewed By	Lead Executive – Quality	Rama Vani Periasamy
Approved By	Head – Quality Assurance	Sekar Thanigaimani

Change History

Version No.	Release Date	Process Improvement Proposal Reference No.	Summary of Changes	Prepared By	Approved By
2.0	3/16/2020	Process 2.0	Initial draft	Gouri Mahendru	Rama Vani Periasamy
2.1	3/19/2020	Update	Format & structures revised	Gouri Mahendru	Rama Vani Periasamy
2.2	9-Jun-20	Update	Added ISO 20000-1:2018 Reference Number	Santhiya P	Sekar T
2.3	14-Jul-20	Update	Updated Section A. 6 (Process Triggers), Section C. 3.4 (Component Upgrade – Implementation), 3.5 (Threshold Management & Control)	Gouri Mahendru	Sekar T

Standards/Model Reference

Standard/Model	Process Area Reference/ ISO Clause(s) No.
ISO 27001:2013	▪ N/A
ISO 20000-1:2018	▪ 8.4.3 – Capacity Management

Glossary and Acronyms

Sr. No.	Glossary / Acronyms	Description
1.	Business Capacity Management (BCM)	BCM is the responsible activity that follows the future business requirements for use in the capacity plan. Due to changing future business requirements, business capacity management aims to meet the required capacity and optimum capacity for the provided services which contributes to capacity planning.
2.	Capacity Database (CDB)	A Capacity Management Database (CDB) consists of one or more databases that serve as a repository for holding multiple types of data used by its subprocesses.
3.	Capacity Management Information System (CMIS)	A virtual repository of all Capacity Management data, usually stored in multiple physical locations.
4.	Component	A component is the underlying structure behind a service. For example, it is the database behind the application or the server underneath the website. It is a component that must be purchased, built, maintained, and monitored.
5.	Capacity Plan	A Capacity Plan is used to manage the resources required to deliver IT services. The plan contains scenarios for different predictions of business demand, and costed options to deliver the agreed service level targets.
6.	Capacity Report	The Capacity Report provides other Service Management processes and IT Management with information related to service and resource utilization and performance.
7.	Component Failure Impact Analysis (CFIA)	Component Failure Impact Analysis (CFIA) is a proactive method to determine the potential impact on service delivery if a component (or configuration item) should fail.
8.	Critical Success Factors (CSF)	Critical success factor (CSF) is a management term for an element that is necessary for an organization or project to achieve its mission. To achieve

Sr. No.	Glossary / Acronyms	Description
		their goals, they need to be aware of each key success factor (KSF) and the variations between the keys and the different roles key result area (KRA).
9.	Event Filtering and Correlation Rules	Rules and criteria used to determine if an Event is significant and to decide upon an appropriate response. Event Filtering and Correlation Rules are typically used by Event Monitoring systems. Some of those rules are defined during the Service Design stage, for example to ensure that Events are triggered when the required service availability is endangered.
10.	Install, Move, Add, Change (IMAC)	The term is in IT management use and summarizes the services in a life cycle of a job are important. The individual terms summarize various aspects related to the dynamics of a workplace.
11.	Key Performance Indicators (KPI)	A performance indicator or key performance indicator is a type of performance measurement. KPIs evaluate the success of an organization or of an activity in which it engages.
12.	Performance	Performance is how quickly a system responds to requests. For example, how quickly an application processes data and returns a new screen is one indicator of its performance.
13.	Resource Capacity Management (RCM)	RCM methodology helps to modify and integrate the current assets and tools in realizing end-to-end automated business processes using a Service-Oriented Architecture (SOA).
14.	Service Capacity Management (SCM)	To manage, control and predict the performance and capacity of operational services. This includes initiating proactive and reactive action to ensure that the performances and capacities of services meet their agreed targets.
15.	Service Level Agreement (SLA)	A service-level agreement (SLA) defines the level of service expected by a customer from a supplier, laying out the metrics by which that service is measured, and the remedies or penalties, if any, should the agreed-on service levels not be achieved.
16.	Total Cost of Ownership (TCO)	The total cost of ownership (TCO) is the purchase price of an asset plus the costs of operation. Assessing the total cost of ownership represents taking a bigger picture look at what the product is and what its value is over time.

Contents

Table of Figures

Section A: Introduction

1. Overview

The following Capacity Management Process has been designed for the GAVS IT Service Management program. It will be used as a reference for the implementation and use of the Capacity Management process on an ongoing basis. This process document is based on the best practices described in the Information Technology Infrastructure Library (ITIL®) Framework. It includes Capacity Management goals, objectives, scope, policies, key terms, roles, responsibilities, authority, process diagrams and associated activity descriptions. This Process will have relationships with other Processes and those documents should be read and understood along with this, the primary related processes being Incident, Problem, Change, Configuration, Service Level, Availability, Continuity and Financial Management.

2. Definition

ITIL® defines **Capacity** of a business measures how much companies can achieve, produce, or sell within a given time.

Capacity Management refers to the act of ensuring a business always maximizes its potential activities and production output — under all conditions.

There are several sources of information that are relevant to the Capacity Management process. Some of these are as follows:

- External suppliers of new technology
- The organizations business strategy and plans, and financial plans
- The IT strategy and plans and current budgets
- The Incident and Problem Management processes with incidents and problems relating to poor performance
- The SLM process with details of the contents of the SLAs and SLRs, and possibly from the monitoring of SLAs, service reviews and breaches of the SLAs
- The Change Management process with a forward Schedule of Changes and a need to assess all Changes for their impact on the capacity of the infrastructure
- The IT Operations team with schedules of all the work that needs to be burn and information on the dependencies between different services, and the interdependencies within a service.

3. Process Objective

ITIL Capacity Management aims to ensure that the capacity of IT services and the IT infrastructure can deliver the agreed service level targets in a cost effective and timely manner. The Capacity Management process considers all resources required to deliver the IT service, and plans for short, medium- and long-term business requirements.

The goals for the Capacity Management process are to:

- Monitoring the performance and throughput or load on a server, server farm, or property
- Performance analysis of measurement data, including analysis of the impact of new releases on capacity
- Performance tuning of activities to ensure the most efficient use of existing infrastructure
- Understanding the demands on the service and future for workload growth (or shrinkage)
- Influences on demand for computing resources
- Capacity planning of storage, computer hardware, software and connection infrastructure resources required over some future period.

4. Scope

The Capacity management should be the focal point for all IT performance and capacity issues. It should cover the operational and development environment which includes the hardware, networking equipment, peripherals, software and human resources. It should also make sure that the IT resources and planned and scheduled to deliver a level of service which is consistent and matched to the present and future needs of the business:

- Processor utilization
- Memory utilization
- Network
- Storage
- Database and Infrastructure Services
- Tape
- Human Resources

However, the driving force for Capacity Management should be the business requirements of the organization.

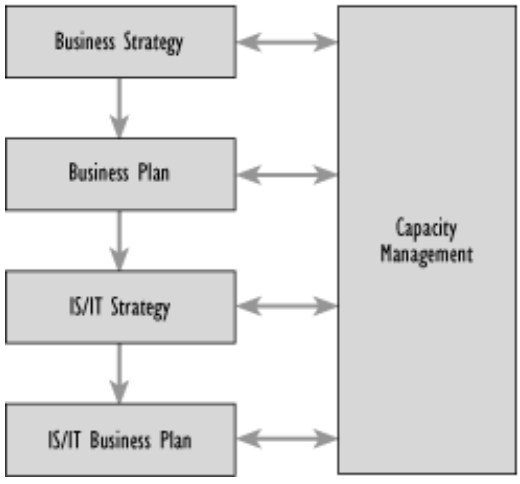

Figure 1: Scope of Capacity Management

4.1. Out of Scope

The scope of Capacity Management can be one service system or multiple service systems. If the service provider is dependent on multiple service systems, Capacity Management can be performed independently on each discrete service system, but the organization will realize reduced value.

5. Policy

The policies defined for Capacity Management process are listed as follows:

- The Capacity Management process shall identify Capacity requirements based on business plans, business requirements, SLAs and risk assessments, and shall be consulted in the development and negotiation of SLA's.
- The Capacity Plans will be reviewed at least annually to ensure requirements reflect agreed-upon changes required by the business.
- The Capacity Management process will be subject to Continuous Process Improvement.
- Capacity Management will endeavor to ensure optimal integration with other ITSM processes.
- The best available demand forecasts should be provided to Capacity Management as soon as
- they are identified.
- Monitoring, data gathering, analysis, reporting, and reviews will be undertaken consistently in a defined manner, with the data being stored in the Capacity Management Database (CDB).
- The contents of the CDB will be shared with other ITSM processes.
- The necessary authority will be delegated to the Capacity Management process to initiate actions which ensure required levels of IT Service Capacity and reliability

6. Process Triggers, Inputs and Outputs

Triggers

There are many triggers that will initiate Capacity Management activities. These include:

- Service breaches, capacity or performance events and alerts, including threshold events
- Exception reports
- Periodic revision of current capacity and performance and the review of forecasts, reports and plans
- New and changed services requiring additional capacity
- Periodic trending and modeling
- Review and revision of business and IT plans and strategies
- Review and revision of designs and strategies
- Review and revision of SLAs, OLAs, contracts or any other agreements.

Inputs for this process are from: Incidents, Problem, Change, Configuration, Release, Finance, and Service Level Management.

- **Business information**: from the organization's business strategy, plans and financial plans, and information on their current and future requirements.
- **Service and IT information**: from Service Strategy, the IT strategy and plans and current budgets, covering all areas of technology and technology plans, including the infrastructure, environment, data and applications, and the way in which they relate to business strategy and plans.
- **Component performance and capacity information**: of both existing and new technology, from manufacturers and suppliers.
- **Service performance issues**: The Incident and Problem Management processes, with incidents and problems relating to poor performance.
- **Service Information**: from the SLM process, with details of the services from the Service Portfolio and the Service Catalogue and service level targets within SLAs and SLRs, and possibly from the monitoring of SLAs, service reviews and breaches of the SLAs.
- **Financial information**: from Financial Management, the cost of service provision, the cost of resources, components and upgrades, the resultant business benefit and the financial plans and budgets, together with the costs associated with service and component failure. Some of the costs of components and upgrades to components will be obtained from procurement, suppliers and manufacturers.
- **Change information**: from the Change Management process, with a Change Schedule and a need to assess all changes for their impact on the capacity of the technology.
- **Performance information**: from the Capacity Management Information System (CMIS) on the current performance of both all existing services and IT infrastructure components.
- **CMS**: containing information on the relationships between the business, the services, the supporting services and the technology.
- **Workload information**: from the IT Operations team, with schedules of all the work that needs to be run, and information on the dependencies between different services and information, and the interdependencies within a service.

Outputs at the same time are from: Availability, Finance, Contingency, Change, and Service Level Management.

The outputs of Capacity Management are used within all other parts of the process, by many other processes and by other parts of the organization. Often this information is supplied as electronic reports or displays on shared areas, or as pages on intranet servers, to ensure the most up-to-date information is always used.

The information provided is as follows:

- **The Capacity Management Information System (CMIS)**: holds the information needed by all sub-processes within Capacity Management. For example, the data monitored and collected as part of Component and Service Capacity Management is used in Business Capacity Management to determine what infrastructure components or upgrades to components j are needed, and when.
- **The Capacity Plan**: used by all areas of the business and IT management. and is acted on by the IT service provider and senior management of the organization to plan the capacity of the IT infrastructure. It also provides planning input to many other areas of IT and the business. It contains information on the current usage of service and components and plans for the development of IT capacity to meet the needs in the growth of both existing service and any agreed new services. The Capacity Plan should be actively used as a basis for decision-making. Too often, Capacity Plans are created and never referred to or used.
- **Service performance information and reports**: used by many other processes. For example, the Capacity Management process assists Service Level Management with the reporting and reviewing of service performance and the development of new SLRs or changes to existing SLAs. It also assists the Financial Management process by identifying when money needs to be budgeted for IT infrastructure upgrades, or the purchase of new components.
- **Workload analysis and reports**: used by IT Operations to assess and implement changes in conjunction with Capacity Management to schedule or reschedule when services or workloads are run, to ensure that the most effective and efficient use is made of the available resources.
- **Ad hoc capacity and performance reports**: used by all areas of Capacity Management, IT and the business to analyze and resolve service and performance issues.
- **Forecasts and predictive reports**: used by all areas to analyze, predict and forecast business and IT scenarios and their potential solutions.
- Thresholds, alerts and events.

7. Benefits

Capacity management is critical to keeping IT costs down and quality of service up.

Most organizations use it to:

- Get more out of existing IT resources
- Improve IT cost per service unit positions
- Fine-tune applications and infrastructure components
- Improve performance, reduce consumption, and delay upgrades
- Eliminate redundant work

- Ensure consistent reporting
- Provision capacity efficiently
- Make informed business decisions using timely capacity and related cost information
- Provide insight into total cost of ownership (TCO) of IT upgrades and initiatives
- Predict future use based on growth levels
- Uncover bottlenecks with enough time to stop them before service is affected

Capacity management teams also have close ties to ITIL service level management and financial management areas.

In fact, capacity management processes lead to more thorough service level and associated financial information for the business. And this helps business leaders make more informed decisions.

Section B: Roles and Responsibilities

1. User Roles and Functions

The responsibilities of various user roles in Capacity Management are listed as follows:

Roles	Responsibilities
Capacity Manager	▪ Responsible for: ○ Ensuring that the aims of Capacity Management are met. ▪ Ensuring that there is adequate IT capacity to meet required levels of service, and that senior IT management is correctly advised on how to match capacity and demand and to ensure that use of existing capacity is optimized. ▪ Identifying, with the Service Level Manager, capacity requirements through discussions with the business users. ▪ Understanding the current usage of the infrastructure and IT services, and the maximum capacity of each component. ▪ Performing sizing on all proposed new services and systems, possibly using modeling techniques, to ascertain capacity requirements. ▪ Forecasting future capacity requirements based on business plans, usage trends, sizing of new services, etc. ▪ Production, regular review and revision of the Capacity Plan, in line with the organization's business planning cycle, identifying current usage and forecast requirements during the period covered by the plan. ▪ Ensuring that appropriate levels of monitoring of resources and system performance are set. ▪ Analysis of usage and performance data and reporting on performance against targets contained in SLAs. ▪ Raising incidents and problems when breaches of capacity or performance thresholds are detected and assisting with the investigation and diagnosis of capacity-related incidents and problems. ▪ Identifying and initiating any tuning to be carried out to optimize and improve capacity or performance. ▪ Identifying and implementing initiatives to improve resource usage – for example, demand management techniques. ▪ Assessing new technology and its relevance to the organization in terms of performance and cost. ▪ Being familiar with potential future demand for IT services and assessing this on performance service levels. ▪ Ensuring that all changes are assessed for their impact on capacity and performance and attending CAB meetings when

Roles	Responsibilities
	appropriate. ▪ Producing regular management reports that include current usage of resources, trends and forecasts. ▪ Sizing all proposed new services and systems to determine the computer and network resources required, to determine hardware utilization, performance service levels and cost implications. ▪ Assessing new techniques and hardware and software products for use by Capacity Management that might improve the efficiency and effectiveness of the process. ▪ Performance testing of new services and systems. ▪ Reports on service and component performance against targets contained in SLAs. ▪ Maintaining a knowledge of future demand for IT services and predicting the effects of demand on performance service levels. ▪ Determining performance service levels that are maintainable and cost justified. ▪ Recommending tuning of services and systems and making recommendations to IT management on the design and use of systems to help ensure optimum use of all hardware and operating system software resources. ▪ Acting as a focal point for all capacity and performance issues.
Capacity Analyst	▪ The Capacity Analyst performs or directs many of the day-to-day and strategic capacity activities on behalf of the Capacity Manager. ▪ Whereas the Capacity Manager is accountable for most capacity-related activities, the Capacity Analyst is responsible for the gathering and analyzing of data for a specific service support area (e.g. Network), and then forwarding the information to the Capacity Manager, who will provide the holistic view for an entire service. ▪ Possesses a comprehensive knowledge of the service delivery infrastructure and the capacity impacts of those infrastructure components on the service. ▪ When analysis is required, initiates the requests to the appropriate infrastructure teams, receives and analyzes the results, and creates the various reports. ▪ Reviews all Capacity reports with the Capacity Manager and publishes them after approval.
Capacity Management Process Owner	▪ Performs: ○ Define process, policies and standards ○ Specify process, purpose, scope and goals

Roles	Responsibilities
	▪ Responsible for: ○ Capacity management framework ▪ Is responsible for, and owns, the process at an executive or management level ▪ Overall performance and results of the process ▪ Identification and management of critical process success factors ▪ Control and lead process improvement ▪ Approval or rejection of process deviation requests ▪ Enforcement of the process ▪ Reporting of process status and progress to peers and executives ▪ Facilitation, resolution or escalation of cross-functional issues ▪ Accountability for the cost and investment in the process ▪ Representation of the process to all external groups

2. RACI Matrix

The following RACI chart outlines which positions are Responsible, Accountable, Consulted, and Informed for each service desk process.

Sr. No.	Activity Description	Capacity Manager	Capacity Analyst	Service Level Manager	Support Team Mgmt.	Process Owner	Config. Manager	Change Manager
1.0 – Manage Business Capacity Requirements								
1.1	Quantify Business Impacts	A	R	C	I	I	I	I
1.2	Review SLA's	A	I	C	I	I	I	I
1.3	Decision – SLA changes required?	A	I	C	I	I	I	I
2.0 – Manage Service Capacity Requirements								
2.1	Strive to ensure agreed service levels are maintained	R	C	A	C	I	I	I

Sr. No.	Activity Description	Capacity Manager	Capacity Analyst	Service Level Manager	Support Team Mgmt.	Process Owner	Config. Manager	Change Manager
2.2	Monitor, Evaluate & Report	A	R	I	C	I	I	I
2.3	Identify Trends	A	R	C	C	I	C	C
2.4	Establish normal Service Operation Levels	A	R	C	C	I	I	I
2.5	Define Exception Levels	A	R	C	C	I	I	I
2.6	Report on service breaches & near misses	A	R	C	C	I	I	I
3.0 – Manage Resource Capacity Requirements								
3.1	Monitor individual hardware & software components	I	A	I	C	I	I	I
3.2	Collect Data	I	A	I	C	I	I	I
3.3	Perform preemptive and proactive problem determination	C	A	C	C	I	I	I
3.4	Determine the effects to Change	C	A	I	C	I	C	C
3.5	Plan & budget HW & SW upgrades and HR augmentation	A	C	C	C	I	C	I

Sr. No.	Activity Description	Capacity Manager	Capacity Analyst	Service Level Manager	Support Team Mgmt.	Process Owner	Config. Manager	Change Manager
3.6	Balance services to use existing resources efficiently and effectively	C	A	C	C	I	C	I
3.7	Evaluate new HW, SW & personnel capability	A	R	I	I	I	I	I
3.8	Finalize and agree on the capacity plan	A	R	C	C	C	C	I
4.0 – Create and Distribute Capacity Reports								
4.1	Define or validate requirements	A	C	C	I	I	I	I
4.2	Define or validate audience	A	C	C	I	I	I	I
4.3	Identify data sources	A	R	I	I	I	I	I
4.4	Gather and analyze data	A	R	I	I	I	I	I
4.5	Produce report	A	R	I	I	I	I	I
4.6	Distribute report	A	R	I	I	I	I	I

Section C: Process Flow

1. High-level Capacity Management Process

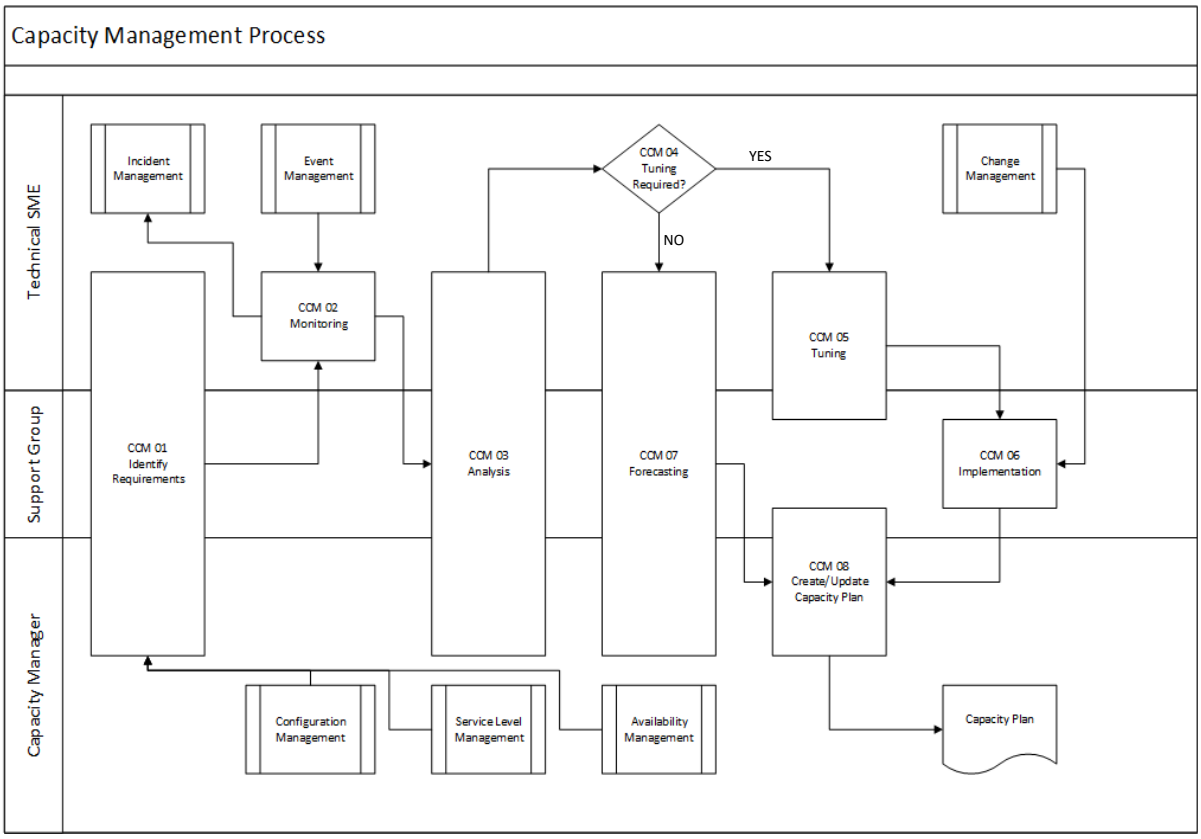

Figure 2: High-level Capacity Management Flowchart

In ITIL®, Capacity go through a structured workflow that encourages efficiency and best results for both providers and customers. ITIL® recommends the Capacity management process follow these steps:

1. Identify Requirements
2. Monitoring
3. Analysis
4. Tuning
5. Implementation
6. Forecasting
7. Planning

2. Capacity Management Sub-process

This section explains several sub-processes of Capacity Management, for which there are various activities.

- **Business (Strategic)**: Strategic capacity management is done when decisions to expand or contract the infrastructure are made due to expected changes in demand by the business (addition of new applications, new infrastructure, etc.).
- **Service (Tactical)**: Tactical capacity management is done when new services are added into the infrastructure (how network devices are configured for services, who can use the capacity, etc.).
- **Resource (Operational)**: Operational capacity management is done by real-time monitoring and adjustments to capacity as needed.

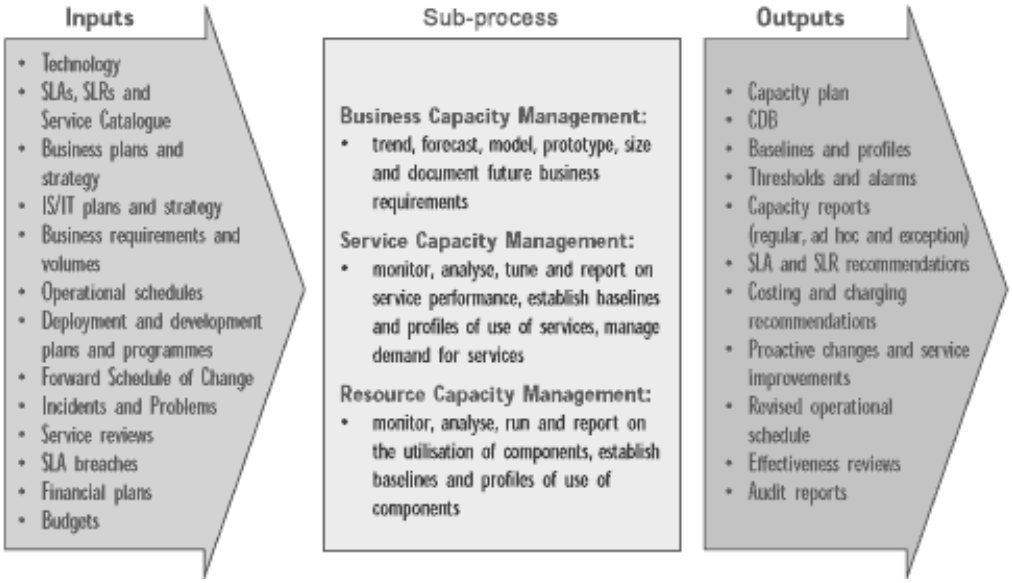

Figure 3: Capacity Management Sub-processes

2.1 Business Capacity Management

Objective: To translate business needs and plans into capacity and performance requirements for services and IT infrastructure, and to ensure that future capacity and performance needs can be fulfilled.

This sub-process is responsible for ensuring that the future business requirements for IT Services are considered, planned and implemented in a timely fashion. This can be achieved by using the existing data on the current resource utilization by the various services to trend, forecast or model the future requirements. These future requirements come from business plan outlining new services, improvements, and growth in existing services, development plans etc.

Figure below shows how new requirements for Capacity cause the Business Capacity Management sub-process to work closely with many of the other Service Delivery and Service Support processes, for example SLM and Change Management, together with other processes such as procurement.

Much of the work that needs to be done by the Business Capacity Management (BCM) sub-process is carried out in conjunction with other processes. In above figure, the three highlighted elements are largely the responsibility of Capacity Management, and of these only "Identify new requirements" is the focus of BCM. "Ensure operational service complies with SLA" and "Resolve Capacity related Incidents and Problems" are the responsibilities of the Service Capacity Management and/or Resource Capacity Management sub-process.

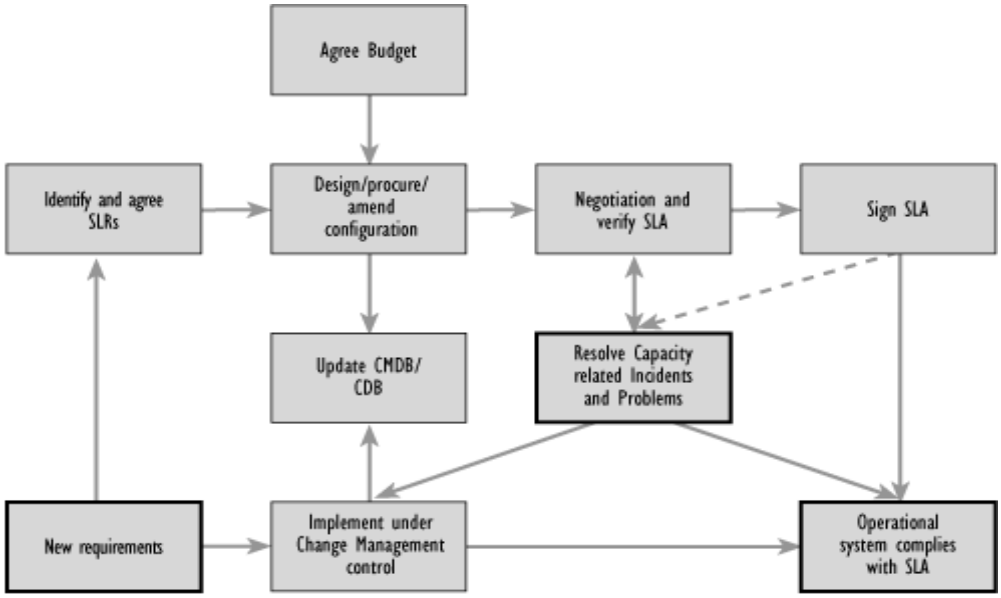

Figure 4: Business Capacity Management

The involvement of the BCM sub-process with other processes is explained below:

- Identify and agree Service Level Requirements
- Design, Procure or Amend Configuration
- Update CMDB and CDB
- Verify SLA
- Sign SLA

2.2 Service Capacity Management

Objective: To manage, control and predict the performance and capacity of operational services. This includes initiating proactive and reactive action to ensure that the performances and capacities of services meet their agreed targets.

The focus of this sub-process is the management of the performance of the live, operational IT Services used by the Customers. It is responsible for ensuring that the performance of all services, as detailed in the targets in the SLAs and SLRs, is monitored and measured and that the collected data is recorded, analyzed and reported. As necessary, the action is taken to ensure that the performance of the services meets the business requirements. This is performed by staff with knowledge of all the areas of technology used in the delivery of end-to-end service and often involves seeking advice from the specialist involved in Resource Capacity Management (RCM).

There will be occasions when Incidents and Problems are referred to Capacity Management from other Service Management processes, or it is identified that a service could fail to meet its SLA targets. On some of these occasions that cause of the potential failure may not be RCM. However, the design or programming of the application is efficient, and so the service performance needs to be managed, as well as individual hardware or software resources.

The key to successful Service Capacity Management (SCM) is to pre-empt difficulties, wherever possible. So, this is another sub-process that must be proactive and anticipatory rather than reactive. However, there are times when it must react to specific performance problems. From a knowledge and understanding of the performance requirement of each of the services being run, the effects of Changes in the use of services can be estimated, and actions taken to ensure that the required service performance can be achieved.

The activities that need to be carried out as part of this sub-process are described in Section 3.

2.3 Component/Resource Capacity Management

Objective: To manage, control and predict the performance, utilization and capacity of IT resources and individual IT components.

The focus in this sub-process is the management of the individual components of the IT Infrastructure. It is responsible for ensuring that all components within the IT Infrastructure that have finite resource are monitored and measured and that the collected data is recorded, analysed and reported. As necessary, action must be taken to manage the available resource to ensure that the IT Services that it supports meet the business requirements. In carrying out this work, the Capacity Management process is assisted by individuals with specialist knowledge in the areas of technology.

As in SCM, the key to successful RCM is to pre-empt difficulties wherever possible, therefore this sub-process must be proactive and anticipatory rather than reactive. However, there are times when it must react to specific Problems that are caused by a lack of resource, or the inefficient use of resource.

3. Activities in Capacity Management

The activities described in the section are undertaken when carrying out any of the sub-processes of Capacity Management and these activities can be done reactively or proactively.

The major difference between the sub-processes is in the data that is being monitored and collected, and the perspective from which it is analyzed.

There are four main activities in Capacity Management which support the creation and update of the Capacity Plan:

- **Iterative Activities**: the monitoring, analysis and tuning of devices and services,
- **Demand Management**: influencing the demand for and usage of computing resources
- **Modeling**: predicting the behavior of IT Services under a given volume and variety of work.
- **Application Sizing**: estimating the resource requirements to support a proposed application Change or new application.

Several activities need to be carried out iteratively and form a natural cycle as illustrated in the figure below:

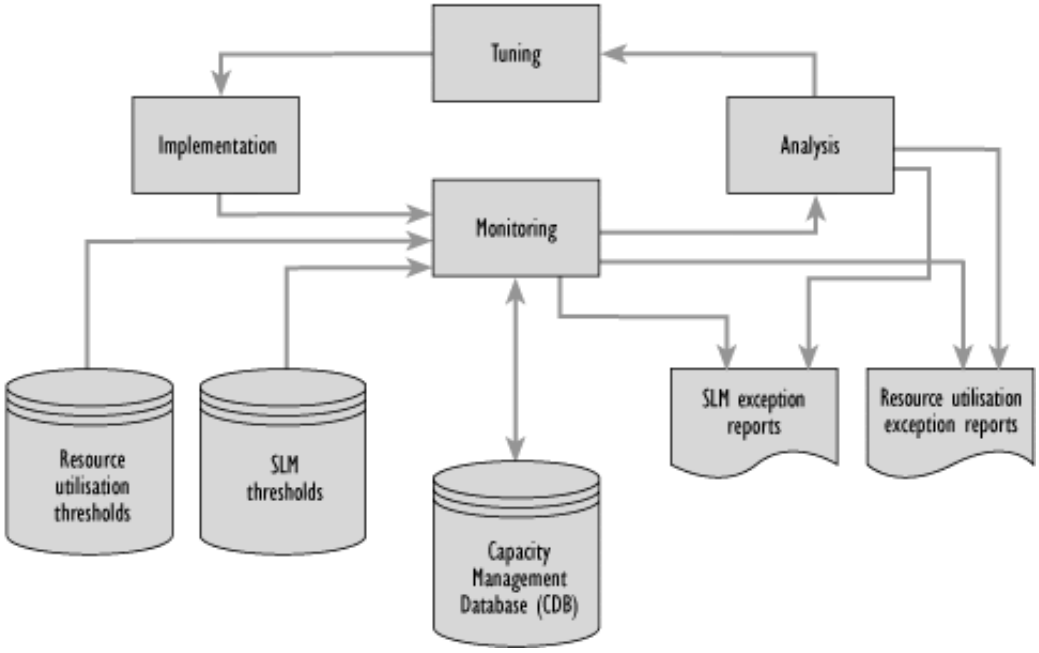

Figure 5: Iterative Ongoing Activities of Capacity Management

Monitors should be established on all the components and for each of the services. The data should be analyzed, using wherever possible, expert systems to compare usage levels against thresholds. The results of the analysis should be included in reports, and recommendations made as appropriate. Some form of control mechanism then be put in place to act on the recommendations. This may take the form of balancing services, changing concurrency levels, and adding or removing resource. The cycle then begins again, monitoring and Changes made to ensure they have had a beneficial effect and collecting the data for the next day, week or month.

- **On an on-going basis**: Iterative Activities, Demand Management and the storage of the data on the CDB
- **Ad-hoc**: Modeling and Application Sizing
- **Regularly**: The production of the Capacity Plan

Any one of the sub-processes of the Capacity Management may carry out any of the activities, with the data this is generated being sore in the CDB.

3.1 Capacity Monitoring

Objective: It is important that the utilization of each resource and service is monitored on an on-going basis to ensure the optimum use of the hardware and software resources, that all agreed service levels can be achieved, and that business volumes are as expected.

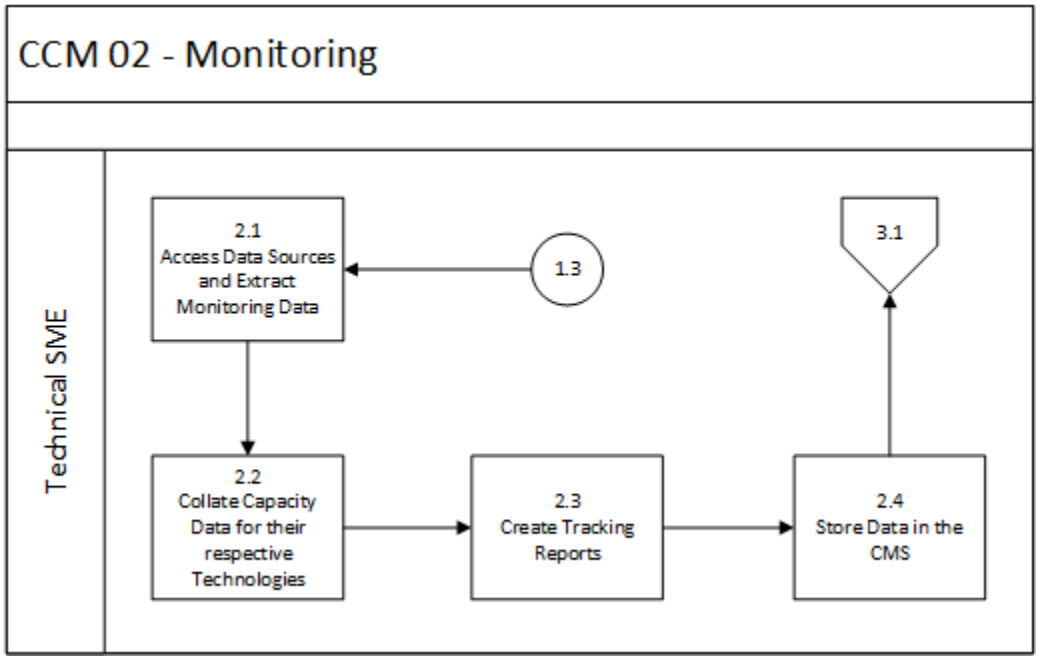

Figure 6: Capacity Monitoring Sub-process

Typical monitoring data includes:

- CPU utilization
- Memory Utilization
- % CPU per transaction type
- IO rates (physical and buffer) and device utilization
- Queue Length (maximum and average)
- File store utilization
- Transactions
- Transactions per second (maximum and average)
- Transaction Response Time
- Batch duration profiles
- Number of Hits
- Number of logons and concurrent users
- Number of Network nodes in use (e.g. Network devices, PCs, Servers, etc.)

Another part of monitoring activity is of thresholds and baselines or profiles of the normal operating levels. If these are exceeded, alarms should be raised, and exception reports produced. All thresholds should be set below the level at which the resource is over-utilized or below the targets in the SLAs.

Most monitoring tasks are near term in nature and rely on underlying tools and principles for operations. The collected information must be recorded or sampled over a determined period. The amount of sampling and resources required to do so must be examined also. The capacity management database (CDB) should contain information points to identify historical trends and patterns.

Data needs to be gathered at total resource utilization level, but also at a more detail profile for the workload that each service places on each resource. This needs to be carried out across the whole infrastructure, host or server, the network, local server, application and client-side or workstation. Similarly, data needs to be collected for each service, for example, availability and a user screen response time.

Part of the monitoring activity is of a baseline or profiles of the normal operating levels. If thresholds beyond the norm are exceeded, alarms are raised, and exception reports produced. These thresholds and baselines are determined from the analysis of previously recorded data, and can be set on:

- Individual components, for example, monitor that the utilization of a CPU does not exceed 80% for a sustained period of one hour
- Specific services, for example, monitor that the presentation time of a web page does not exceed 3 seconds, or the transaction rate does not exceed 1000 transactions per minute.
- It is also important to remember monitoring takes up system capacity, thus can influence the performance of the system. Focus performance measurement and monitors on client service level agreements (SLAs). Operating level requirements and other necessary elements for monitoring often fall out of their overall contribution to meeting the SLA. Monitor at successive levels of control (for example, key IT layers: network, OS, hardware, application, and so on) to ascertain OLOs are met.

The operating system, applications management, associated hardware agent, and systems management tools may dictate which monitors are most readily available. Business rules can correlate element data to service levels in many cases. Many monitors are included as part of the operating system, or free as part of a hardware and software vendor solution, while others form part of a larger systems management tool set and need to be evaluated and purchased separately. It is important that the monitors can collect all the data required by the capacity management process, for a specific component or service.

3.2 Capacity Maintenance - Analysis

Objective: The data collected from the monitoring should be analyzed to identify trends from which the normal utilization and service level, or baseline, can be established. This data can be used to predict future resource usage, or to monitor actual business growth against predicted growth.

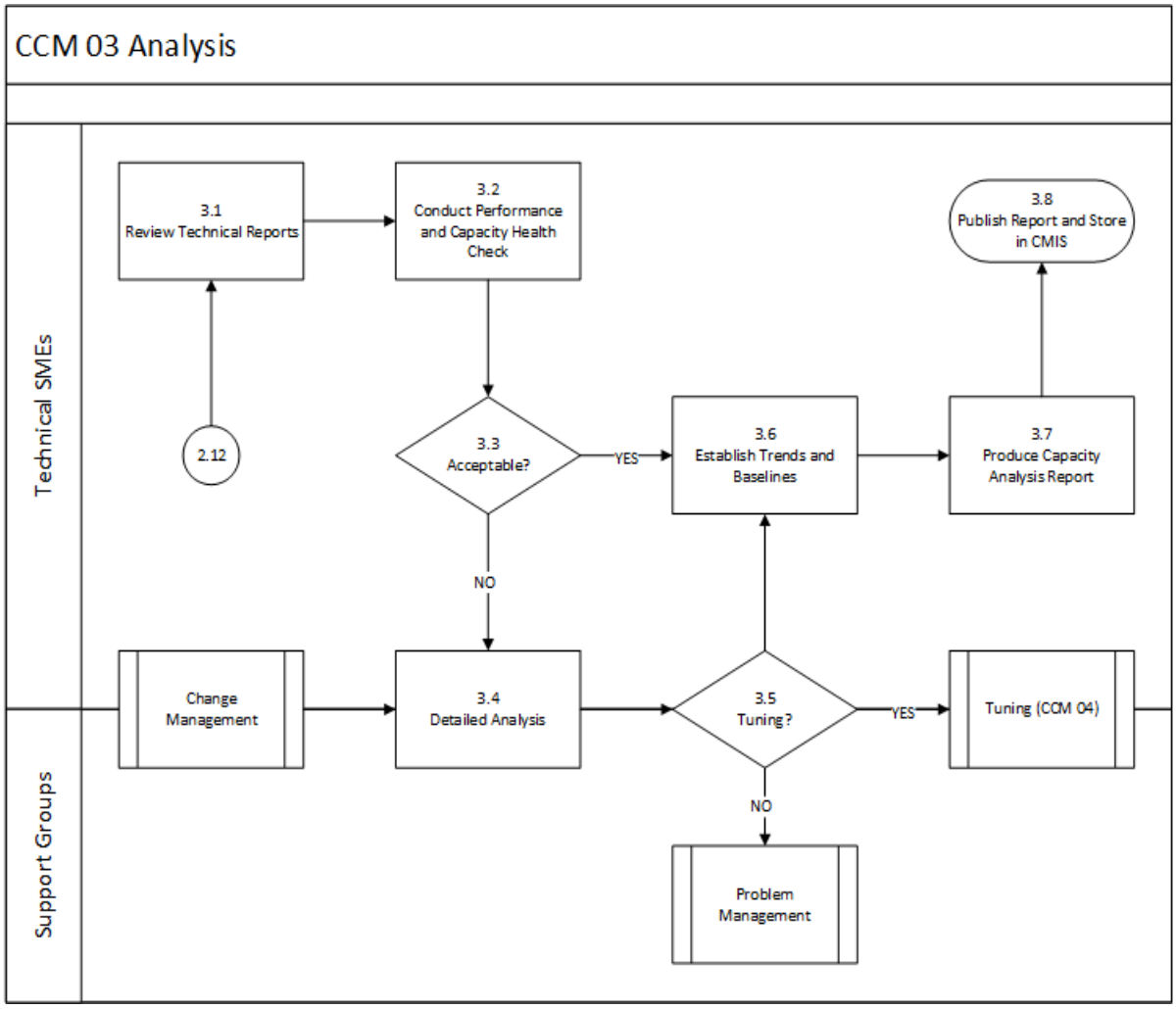

Figure 7: Capacity Analysis Sub-process

Analysis of the data may identify issues such as:

- Contention (Data, File, Memory, Processor)
- Inappropriate distribution of workload across available resource
- Inappropriate locking strategy
- Inefficiencies in the application design
- Unexpected increase in transaction rates
- Inefficient use of memory

When the utilization of a resource is considered, it is important to understand both the total level of utilization and the utilization by individual services of the resource.

Data monitored and collected is analysed for identification and adjustment of thresholds and alarms. In reactive organizations these will trigger exception reports and/or which then need to be analysed and reported upon, and corrective action taken. Ideally, all thresholds should be set below the level at which the resource is over-utilized or below the targets in the OLA or layered OLO. This enables capacity management to take corrective action before the targets in the OLAs have been breached, or the resource has become over-utilized and there has been a period of poor performance.

In proactive organizations, the data collected from the monitoring should be analysed to identify trends from which the normal utilization and service level, or baseline, can be established. By regular monitoring and comparison with this baseline, exception conditions in the utilization of individual components or service thresholds can be defined, and breaches or near misses in the OLAs can be reported. In addition, the data can be used to predict future resource usage.

Analysis of the data may identify issues of:

- Contention (data, file, memory, processor)
- Inappropriate distribution of workload across available resource
- Inappropriate locking strategy
- Inefficiencies in the application design
- Unexpected increase in transaction rate
- Inefficient use of memory

The use of each resource and service needs to be considered over the short, medium, and long-term, and the minimum, maximum and average utilization for these periods recorded. Over time, the trend in the use of the resource by the various IT services becomes apparent.

One key to determining whether a solution is operating at an acceptable level is latency, or the length of time a user must wait for a response once a request for information is complete. Heavy workload on a server might create unacceptable wait times even though the server may be capable of handling every request. As a rule, try to isolate components that have repeatable, high percentage contribution to performance levels and report them at varying workloads.

It is important to understand the utilization in each of these periods, so that changes in the use of any service can be related to predicted changes in the level of utilization of individual resources. The ability to identify the specific hardware or software resource on which an IT service depends, is improved greatly by an accurate, up-to-date and comprehensive CMDB. Any relevant detail performance information should be related or reside and maintained in the capacity database (CDB).

When the utilization of a resource is considered, it is important to understand both the total level of utilization and the utilization by individual services of the resource.

The analysis and tuning activities may also benefit from general observations and guidelines in the Guidelines for Effective Capacity Management and Designing Information Technology Solutions for scalability sections in this document.

3.3 Capacity Maintenance - Tuning

Objective: The analysis of the monitored data may identify areas of the configuration that could be tuned to better utilize the system resource or improve the performance of the service.

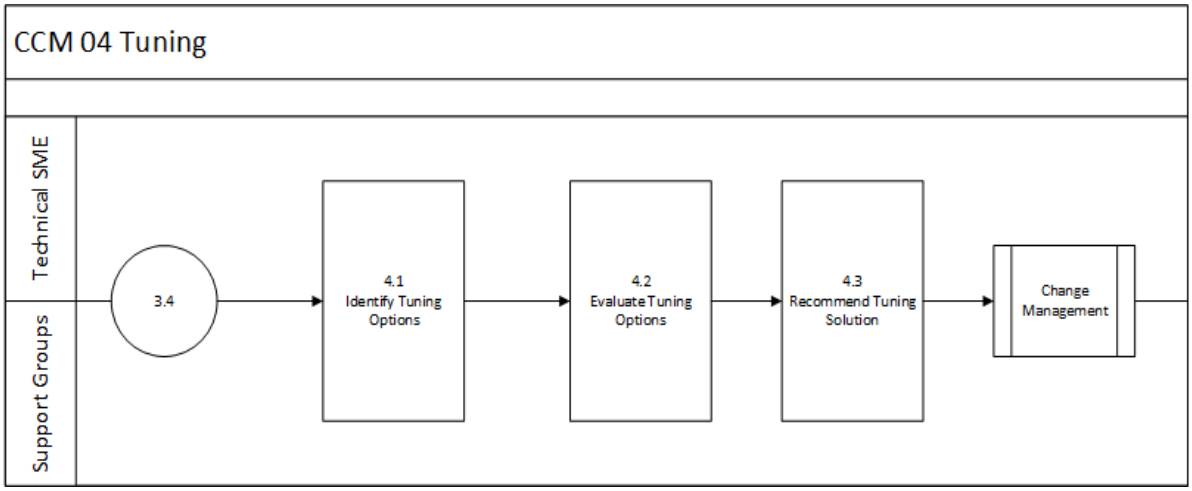

Figure 8: Capacity Tuning Sub-process

The analysis of the monitored data may identify areas of the configuration that could be tuned to better utilize the system resource or improve the performance of the service.

Tuning techniques that are of assistance include:

- **balancing workloads** - transactions may arrive at the host or server at a gateway, depending where the transaction was initiated; balancing the ratio of initiation points to gateways can provide tuning benefits
- **balancing disk traffic** - storing data on disk efficiently and strategically, e.g. stripping data across many spindles may reduce data contention
- **definition of an accepted locking strategy** - specifies when locks are necessary and the appropriate level, e.g. database, page, file, record, and row - delaying the lock until an update is necessary may provide benefits
- **efficient use of memory** - may include looking to utilize memory depending upon the circumstances.

Regarding the efficient use of memory, note that a process may utilize resources more efficiently if data is read into memory and manipulated there rather than a sequential read through files. Alternatively, many processes may be contending for memory resource. The excessive demands may lead to increased CPU utilization and delays while pages are swapped in and out of memory.

Before implementing any of the recommendations arising from the tuning techniques, it may be appropriate to consider using one of the on-going, or ad hoc activities to test the validity of the recommendation. For example, 'Can Demand Management be used to avoid the need to carry out any tuning?' or 'Can the proposed Change be modelled to show its effectiveness before it is implemented'.

3.4 Component Upgrade - Implementation

Objective – To introduce to the live operation service, and Changes that have been identified by the monitoring, analysis and tuning activities.

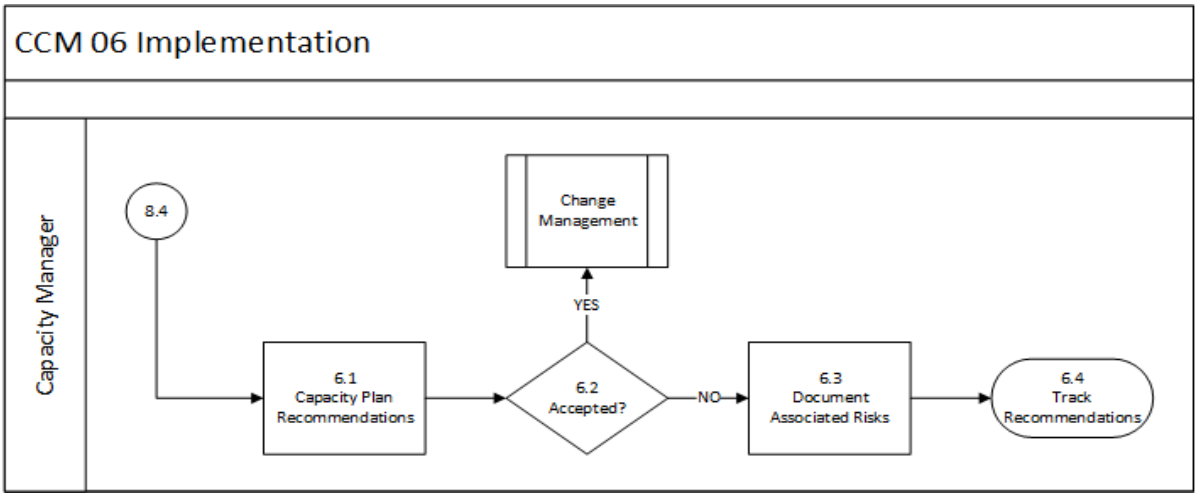

Figure 9: Capacity Implementation Sub-process

Implementing the tuning Changes under formal Change Management procedures results in:

- Less adverse impact on the users of the service
- Increased user productivity
- Increased productivity of IT personal
- A reduction in the number of Changes that need to be blacked-out but the ability to do so more easily
- Greater management and control of business-critical application services

It is important that further monitoring takes place, so that the effects of the Change can be assessed. It may be necessary to make further Changes or to regress some of the original Changes.

3.4.1. Exploitation of New Technology

This involves understanding new techniques and new technology and how they can be used to support the business and innovate improvements. It may be appropriate to introduce new technology to improve the provision and support of the IT services on which the organization is dependent. This information can be gathered by studying professional literature (magazine and press articles) and by attending:

- Promotional seminars by hardware and software suppliers
- User group meetings of suppliers of potential hardware and software
- User group meetings for other IT professionals involved in Capacity Management

Each of these provides sources of information relating to potential techniques, technology, hardware and software, which might be advantageous for IT to implement to realize business benefits. However, always Capacity Management should recognize that the introduction and use of this new technology must be cost-justified and deliver real benefit to the business. It is not just the new technology itself

that is important, but Capacity Management should also keep aware of the advantages to be obtained from the use of new technologies, using techniques such as 'grid computing', 'virtualization' and 'on-demand computing'.

3.4.2. Designing Resilience

Capacity Management assists with the identification and improvement of the resilience within the IT infrastructure or any subset of it, wherever it is cost-justified. In conjunction with Availability Management, Capacity Management should use techniques such as Component Failure Impact Analysis (CFIA, as described in section 4.4 on Availability Management) to identify how susceptible the current configuration is to the failure or overload of individual components and make recommendations on any cost-effective solutions.

Capacity Management should be able to identify the impact on the available resources of failures, and the potential for running the most important services on the remaining resources. So, the provision of spare capacity can act as resilience or fail-over in failure situations.

The requirements for resilience in the IT infrastructure should always be considered at the time of the service or system design. However, for many services, the resilience of the service is only considered after it is in live operational use. Incorporating resilience into Service Design is much more effective and efficient than trying to add it later, once a service has become operational.

3.5 Threshold Management & Control

The technical limits and constraints on the individual services and components can be used by the monitoring activities to set the thresholds at which warnings and alarms are raised and exception reports are produced. However, care must be exercised when setting thresholds, because many thresholds are dependent on the work being run on the component.

The management and control of service and component thresholds is fundamental to the effective delivery of services to meet their agreed service levels. It ensures that all service and component thresholds are maintained at the appropriate levels and are continuously, automatically monitored, and alerts and warnings generated when breaches occur. Whenever monitored thresholds are breached or threatened, then alarms are raised and breaches, warnings and exception reports are produced. Analysis of the situation should then be completed, and remedial action taken whenever justified, ensuring that the situation does not recur. The same data items can be used to identify when SLAs are breached or likely to be breached or when component performance degrades or is likely to be degraded. By setting thresholds below or above the actual targets, action can be taken, and a breach of the SLA targets avoided. Threshold monitoring should not only alarm on exceeding a threshold but should also monitor the rate of change and predict when the threshold will be reached. For example, a disk-space monitor should monitor the rate of growth and raise an alarm when the current rate will cause the disk to be full within the next N days. If a 1GB disk has reached 90% capacity, and is growing at 100KB per day, it will be 1,000 days before it is full. If it is growing at 10MB per day, it will only be 10 days before it is full. The monitoring and management of these events and alarms is covered in detail in the Service Operations publication.

There may be occasions when optimization of infrastructure components and resources is needed to maintain or improve performance or throughput. This can often be done through Workload Management, which is a generic term to cover such actions as:

- Rescheduling a service or workload to run at a different time of day or day of the week, etc. (usually away from peak times to off-peak windows) - which will often mean having to adjust

job-scheduling software

- Moving a service or workload from one location or set of CIs to another - often to balance utilization or traffic
- Technical 'virtualization': setting up and using virtualization techniques and systems to allow the movement of processing around the infrastructure to give better performance/resilience in a dynamic fashion
- Limiting or moving demand for components or resources through Demand Management techniques, in conjunction with Financial Management.

It will only be possible to manage workloads effectively if a good understanding exists of which workloads will run at what time and how much resource utilization each workload places on the IT infrastructure. Diligent monitoring and analysis of workloads, together with a comprehensive CMIS, are therefore needed on an ongoing operational basis.

3.6 Storage

The Capacity Database (CDB) is the cornerstone of a successful Capacity Management process. Data in CDB is stored and used by all the sub-processes of Capacity Management because it is a repository that holds several different types of data. However, the CDB is unlikely to be a single database and probably exists in several physical locations.

The information in the CDB is used to form the basis of performance and Capacity Management reports that are to be delivered to management and technical personal. Also, the data is utilized to generate future Capacity forecasts and allow the Capacity Management to plan for future Capacity requirements.

The inputs to the CDB:

- Business Data
- Service Data
- Technical Data
- Financial Data
- Utilization Data

The outputs from the CDB:

- Service and component-based reports
- Exception reporting
- Capacity forecasts

3.7 Demand Management

Objective: To influence the demand for computing resource and the use of that resource.

Demand Management needs to understand which services are utilizing the resource and to what level and needs to know the schedule of when they must be run. Then a decision can be made on whether it will be possible to influence the use of resource, and if so, which option is appropriate.

- Short-term Demand Management may occur when there has been a partial failure of a critical resource in the IT Infrastructure and a limited subset of services could be run. Capacity Management should be aware of the business priority of each of the services, know the resource requirements of each service and then be able to identify which services can be run while there is a limited amount of memory available.
- Long-term Demand Management may be required when it is difficult to cost-justify an expensive upgrade.

The influence on the services that are running could be exercised by:

- Physical Constraints – Stop some services from being available at certain times, or to limit the number of Customers who can use a service.
- Financial Constraints – if Charging for IT service is occurring, reduced rates could be offered for running work at certain times of day, that is the times when there is current less demand for the resource.

3.8 Modelling/Forecasting

Objective: To predict the behavior of IT services under a given volume and variety of work. It is an activity that can be used to beneficial effect in any of the sub-processes of Capacity Management.

The different types of modelling range from making estimates based on experience and current resource utilization information, to pilot studies, prototypes and full-scale benchmarks.

- Trend Analysis
- Analytical Modelling
- Simulation Modelling
- Baselines Models

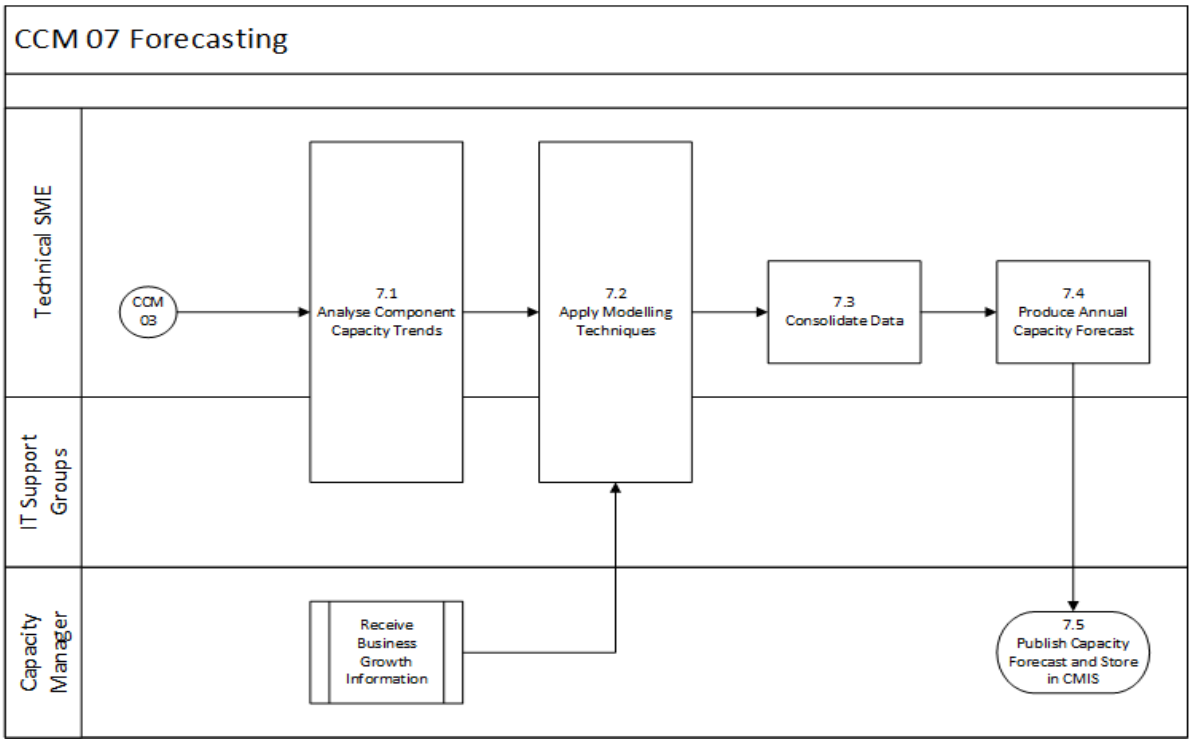

Figure 10: Capacity Forecasting Sub-process

3.9 Application Sizing

Objective: To estimate the resource requirements to support a proposed application change or new application, to ensure that it meets its required service levels. To achieve this application sizing must be an integral part of the application lifecycle.

During the initial systems analysis and design the required service levels must be specifies. This enables the application development to employ the pertinent technologies and products, in order to achieve a design that meets the desired levels of service.

The sizing of the application should be refining as the development process progresses. The use of modelling can be used within the application sizing process.

The SLRs of the planned application developments should not be considered in isolation. The resources to be utilized by the application are likely to be shared with other services and potential threats to existing SLA targets must be recognized and managed.

3.10 Planning

Objective: To produce a plan that documents the current levels of resource utilization and service performance, and after consideration of the business strategy and plans, forecasts the future requirements for resource to support the IT services that underpin the business activities. The plan should indicate clearly any assumptions made, any recommendations quantified in terms of resource required, costs, benefits, impacts, etc.

The Capacity plan should be published annually in line with the budgetary cycle. Ideally it should be updated quarterly.

Typical contents of Capacity plan are:

- Scope of the plan – explicitly name those elements of the IT infrastructure that are included.
- Methods used – details of how and when this information was obtained.
- Assumptions made – any assumptions made, particularly those concerning the business drivers of IT capacity are highlighted early on the plan.
- Management summary – highlights the main issues, options, recommendations and costs.
- Business scenarios – explicitly mention all known business forecasts so that readers can determine what is in and what is outside the scope of plan.
- Service summary – this includes the following:
    - Current and recent service provision
    - Service forecasts
- Resource summary – this includes the following:
    - Current and recent resource usage
    - Resource forecasts
- Options for service improvement – outlines the possible options for improving the effectiveness and efficiency of service delivery.
- Cost model – the costs associated with these options should be documented here, along with the current and forecasts cost of providing IT services.

- Recommendations – summary of the recommendations made in the previous plan and their status along with any new recommendations to be made. The recommendations should be quantified in terms of:
    - The business benefits to be expected
    - The potential impact of carrying out the recommendations
    - The risks involved
    - The resources required
    - The costs, both set-up and on-going

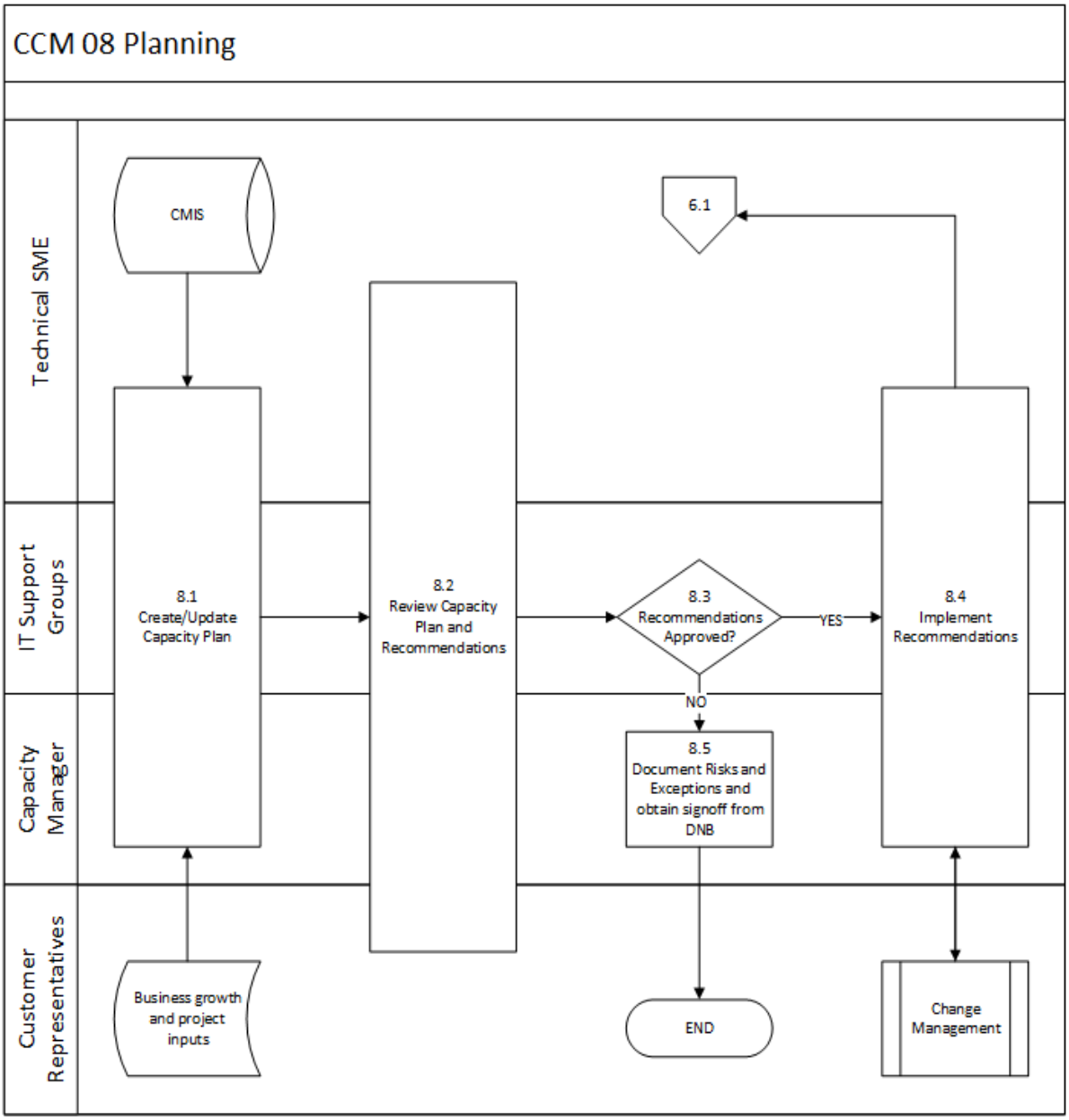

Figure 11: Capacity Planning Sub-Process

4. Capacity Management Reporting

Objective: To provide other Service Management processes and IT Management with information related to service and resource capacity, utilization and performance.

It gathers and then provides other stages with the data related to service capacity, service usage, and service performance. The output of this sub-process is the service capacity report.

Define relevant performance indicators, the systematic and timely reporting of performance and prompt acting upon deviations to ensure the achievement of the performance objectives set for the IT processes. It includes:

- **Collecting Monitoring Data**: relevant capacity indicators (e.g., benchmarks) from both internal and external sources have been defined and data collected for the creation of management information reports and exception reports regarding these indicators,
- **Assessing Performance**: measure capacity and compare it with target levels. Perform assessments of capacity on a continuous basis,
- **Assessing Customer Satisfaction**: Measure customer satisfaction with response, workload, etc at regular intervals to identify shortfalls in service levels and establish improvement objectives,
- **Management Reporting**: Provided management reports for senior management's review of the organization's progress toward capacity targets. Upon review, appropriate management action is initiated.

Section D: Governance and Process Controls

1. Reports and Metrics

For each measurement period in question, the metrics to check on the effectiveness and efficiency of the Capacity Management process should include the following:

- Cycle Time
- Defects
- Change Agents

Typical reports include:

- Utilization Reports on the components in CDB
- SLA breeches reports by weeks, months
- Services usage reports by days, weeks, months
- Reduction in cost report by service
- Up-time reports
- Resource forecasts reports

2. Critical Success Factor (CSF)

- Considering IT Service capacity requirements during service design.
- Ability to plan and implement appropriate capacity to match current and future business needs.
- Provision of organizational strategic business plans.
- Provision of accurate business forecasts.
- Creation of an integrated source of capacity data to allow analysis of the usage of all Configuration Items in scope.
- Ability to successfully analyze capacity data.
- Creation of appropriate thresholds provide warnings and alerts.
- Provision of technology to automatically manage thresholds.
- Senior management commitment in terms of resources and budget for the process.

3. Key Performance Indicators (KPI)

The following table lists the key performance indicators (KPIs) that have been selected for tracking the success of the Capacity Management process. The KPIs will be measured and calculated as a percentage and reflected in the monthly SLA reports.

Sl. No.	KPI	Definition
1.	Incidents due to Capacity Shortages	• Number of incidents occurring because of insufficient service or component capacity
2.	Exactness of Capacity Forecast	• Deviation of the predicted capacity development from actual course
3.	Capacity Adjustments	• Number of adjustments to service and component capacities due to changing demand
4.	Unplanned Capacity Adjustments	• Number of unplanned increases to service or component capacity as result of capacity bottlenecks
5.	Resolution Time of Capacity Shortage	• Resolution time for identified capacity bottlenecks
6.	Capacity Reserves	• Percentage of capacity reserves at times of normal and maximum demand
7.	Percentage of Capacity Monitoring	• Percentage of services and infrastructure components under capacity monitoring

4. Escalation Matrix

Escalation matrix allows you to notify the right stakeholders in the Capacity of critical issues. These contact details are presented to the service delivery team while creating or updating a service ticket. You can notify the right people at the right time about critical tickets based on the escalation matrix. The escalation matrix is time zone specific and can be available 24X7.

Please find the attached list to understand the escalation matrix.

GAVS_Escalation Matrix Template

5. Understanding Priority: Impact * Urgency

ITIL definition for Impact: A measure of the effect of a Capacity, Problem or Change on Business Processes. Impact is often based on how Service Levels will be affected. Impact and Urgency are used to assign priority.

Impact Codes

Impact Code	Description
High	The System is non-operational. It must be corrected immediately. Data is corrupted. Abnormal termination of jobs or user sessions. Production is down. Problem impacts system availability and integrity. No practical workarounds. Public safety compromised. Major Capacity implications.
Medium	Major functions of the system are unavailable, unstable or overall system performance is below normal load requirements. Needed capability is missing. Workarounds awkward or inefficient. Recovery is difficult. Misleading output or documentation. Impacts system credibility. High probability of impact to service. Capacity implications over long period only.
Low	Minor functions of the system are unavailable, unusable or overall system performance is below maximum load requirements. It must be corrected but may wait until the next release. Work can be accomplished with minor inconvenience or loss of efficiency. Temporary workaround available and acceptable. Problem occurrence is infrequent and readily visible.
None	A minor deficiency with minimal impact to users. It can be corrected in subsequent releases. Cosmetic problem. No loss of functionality. Long-term workaround acceptable to End user. No Capacity implications. A corporate opportunity exists to improve the functionality of the system.

ITIL definition for Urgency / Severity: It is the measure of business criticality of a Capacity, Problem or Change where there is an effect upon business deadlines. The urgency reflects the time available for repair or avoidance before the impact is felt by the business. Together with impact, and perhaps technical severity, it is the major means of assigning priority for dealing with Capacity's, Problems or Changes.

Impact Code	Description
High	The Capacity is to be resolved with high urgency to perform a critical business function. Delaying in resolution may lead to significantly high business / productivity loss.
Medium	The Capacity is to be resolved immediately to avoid loss of functionality which may lead to business loss. Non-resolution may lead to loss of group productivity.
Low	The resolution of Capacity can be planned for the near future. Non-resolution may lead to loss of individual productivity.

Impact Code	Description
None	There is no urgency and the action can be scheduled for future implementation. There is no loss of productivity if it not resolved.

Priority Matrix

Impact / Urgency	High	Medium	Low	None
High	P1	P1	P2	P3
Medium	P1	P2	P3	P4
Low	P1	P2	P3	P4
None	P2	P3	P4	P4

Deriving at Priority Codes

Priority	Name	Description
P1	Critical	• Live system is at a halt and unable to process data • Major component not available for use • No viable or productive work around available • Major loss of functionality • Problem cannot be bypassed
P2	Important	• Reduced functionality causing severe disruption to the completion of business-critical tasks • Limited use of product or component • Capacity reported may have a possible bypass. Workarounds are possible
P3	Normal	• Minimal loss of functionality • Reduced functionality causing some disruption to the completion of business-critical tasks • Capacity reported may be bypassed/ redundancy in place. Workarounds are possible
P4	Low	• Few functions impaired • Non-urgent query or request • Service request for Installations, Moves, Additions and Change (IMAC)

Section E: ITIL Inter-relationships and Best Practices

1. Relationships with other ITIL Processes

ITIL describes an integrated set of processes which, collectively, describe an overall approach or framework to service management These interdependencies for Capacity Management process are described below.

1.1. Incident Management

IM provides Capacity Management with information on capacity related incidents which may need to be reflected in Capacity upgrades.

Refer to the link for the standard practice for Incident Management

https://mygavs.gavstech.com/ims/

1.2. Problem Management

PM provides Capacity Management with information on capacity related problems which may need to be reflected in Capacity upgrades. Refer to the link for the standard practice for Problem Management

https://mygavs.gavstech.com/ims/

1.3. Change Management

CM provides Capacity Management with information on changes which may affect Capacity planning. Refer to the link for the standard practice for Change Management

https://mygavs.gavstech.com/ims/

1.4. Configuration Management

The CMDB serves as a source of reference for capacity management. The information contained in the CMDB provides a complete picture of components in the IT environment. This facilitates the activities of capacity management, such as capacity planning and performance management.

1.5. Service Level Management

Service level agreement (SLA) data allows a more proactive measurement of performance based on SLA compliance. This data is an important input to capacity management.

1.6. Availability Management

Availability reporting and measurement highlight availability trends indicating capacity or performance issues.
Availability data allows a more proactive measurement of performance based on SLA compliance. This data is an important input to Capacity Management.

The Capacity Plan needs to be coordinated with the availability plan. The same technology solution can often meet the needs of both plans. The availability and capacity plans should be created in

collaboration. Some solutions that cannot be cost-justified for one plan may be justified in combination with the other.

1.7. Continuity Management

SCM may develop Continuity Plans which have implications for Capacity Management.

1.8. Financial Management

FM acts a filter; ensuring that solutions proposed by Capacity Management can be justified in terms of their cost to implement versus their benefit to the customer.

1.9. Release Management

Capacity Management helps determine the distribution strategy, particularly where the network is used for distribution. It provides the necessary planning data and technical expertise to support the strategy on an ongoing basis and for individual distribution requirements.

Section F: Appendix

1. Capacity Plan Template

ITIL Capacity Planning Template.c

2. Capacity Management Report Checklist

ITIL Capacity Report (Checklist).docx