

Capacity Management: A CA Service Management Process Map

Denise P. Kalm

DIRECTOR, PRODUCT MARKETING

Marv Waschke

SR. ADVISOR, PRODUCT MANAGEMENT

Table of Contents

Executive Summary	1
--------------------------	----------

SECTION 1: CHALLENGE	2
----------------------	----------

Simplifying ITIL

How to Use the CA Service Management
Process Maps

Finding the Right Path to IT Service Excellence

SECTION 2: OPPORTUNITY	4
------------------------	----------

Capacity Management

Capacity Management Sub-Processes

Monitor Demand

Analyze

Monitor Performance

Forecast

Creating the Capacity Plan

New Services (Application Sizing)

Procure

Load Test

Adjust

Demanding Management—A Form of
Adjustment

SECTION 3: BENEFITS	12
---------------------	-----------

Benefits of Capacity Management Best Practices

SECTION 4: CONCLUSIONS	12
------------------------	-----------

SECTION 5: ABOUT THE AUTHOR	13
-----------------------------	-----------

ABOUT CA

BACK COVER

Executive Summary

Challenge

The Information Technology Infrastructure Library version 3 (ITIL® V3) process framework addresses the IT Service Management (ITSM) service lifecycle. The Service Lifecycle is an organizational model providing insight into the structure of ITSM, with critical guidance for IT organizations seeking to improve service quality and align closely with business to create customer and business value.

However, ITIL V3 best practice guidelines for the five phases of the service lifecycle are complex and challenging to interpret. Moreover, they do not provide definitive guidance for implementing ITSM processes. Many IT organizations consequently undertake an ITIL journey without firm goals or plans.

Too often overlooked, the ITIL Capacity Management process focuses on proactive management of capacity and performance. It provides critical support for many other processes, such as Availability, Service Level, Financial, Problem, and more. As such, it is a critical process. The quality of an organization's Capacity Management can make or break its reputation with its customers.

Opportunity

Capacity Management seeks to optimize IT capacity to meet Service Level Agreements (SLAs) in the most cost-effective manner. Capacity Management proactively ensures that an IT infrastructure will meet both current and future business demands. Gathering and analyzing relevant data, Capacity Management helps plan for upcoming needs and optimize purchasing decisions—continuously balancing cost against capacity, and supply against demand. CA has developed a unique approach to representing the ITIL framework and its interdependent ITSM processes at a high level in the form of an easy-to-use subway map. This map is an ideal starting point for understanding and communicating about ITIL in support of successful planning and implementation.

Benefits

The CA Capacity Management process map enables IT organizations to manage by:

- Spending money on the right things at the right time
- Utilizing the systems and budget you have in the most effective manner
- Giving you the tools to proactively reduce system incidents and problems
- Increasing customer satisfaction
- Anticipating capacity requirements before business is adversely impacted
- Reducing the threat of service degradation through better resource management
- Diminishing the risk associated with changes to the IT infrastructure

SECTION 1: CHALLENGE

Simplifying ITIL

The ITIL V3 process framework focuses on the service lifecycle and the structure and linkage service management components. It embodies critical guidance for IT organizations that are seeking to improve service quality and align more closely with business goals.

But, the ITIL V3 best-practice guidelines across the five phases of the service lifecycle are complex and challenging to interpret. Moreover, they are not designed to provide definitive advice about implementing IT Service Management (ITSM) processes. Many IT organizations consequently undertake an ITIL journey without a firm idea of their goals and the path to achieve those goals.

CA has developed a unique approach to charting the ITIL journey through a visual representation of the ITIL framework and its interdependent ITSM processes modeled after an urban subway system. This three-part map (Figure A) presents an easy-to-navigate, high-level view of the ITIL terrain. IT executives, strategists and implementers can use these Service Management process maps along with the family of CA Service Management process map technology briefs that expand on them. The maps and technology briefs provide a common reference point for understanding and communicating about ITIL and help you with program planning and implementation.

How to Use the CA Service Management Process Maps

CA's Service Management process maps (Figure A) illustrate every process (or track), each activity (or station) and the key relationships that are relevant to navigating continuous IT service improvement. The ITIL quality cycle takes the form of a "circle" with each Plan-Do-Check-Act (P-D-C-A) step as a process integration point (junction) on the line. Junctions serve both as reference points when assessing process maturity, and as a means to consider the implications of implementing a process in isolation.

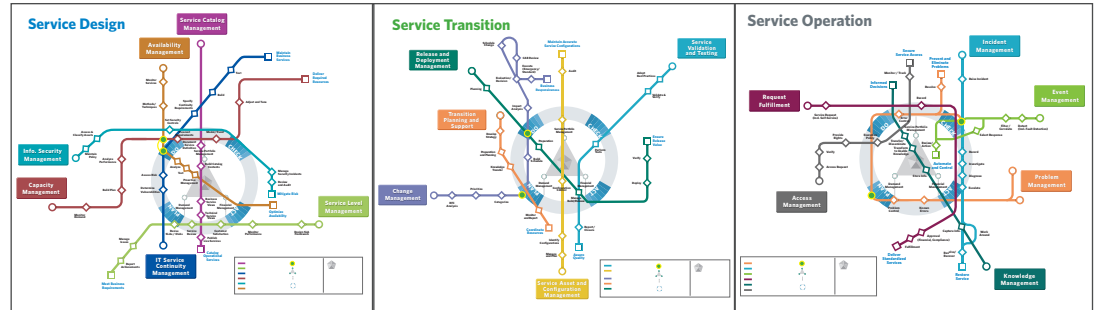
Strategic controls (Service Portfolio Management, Demand Management and Financial Management) are needed to reduce risk and optimize integration across the service lifecycle, as illustrated on the three points of the triangle centered in the P-D-C-A quality circle (seen more easily in Figure B). These strategic controls help in evaluating, prioritizing and assuring the appropriate levels of financial and human resources for existing and new services.

This paper is part of a series of Service Management Process Map technology briefs. Each brief explains how to navigate a particular ITIL process journey, reviewing each process activity that must be addressed in order to achieve process objectives. Along each journey careful attention is paid to how technology plays a critical role in both integrating ITIL processes and automating ITIL process activities.

CA ITSM PROCESS MAPS

CA ITSM Process Maps illustrate at a high level how best to navigate a journey of continual service improvement guided by strategic controls throughout the service lifecycle. Each map describes the relevant ITIL processes and activities you'll need to work with to reach your goals.

FIGURE A

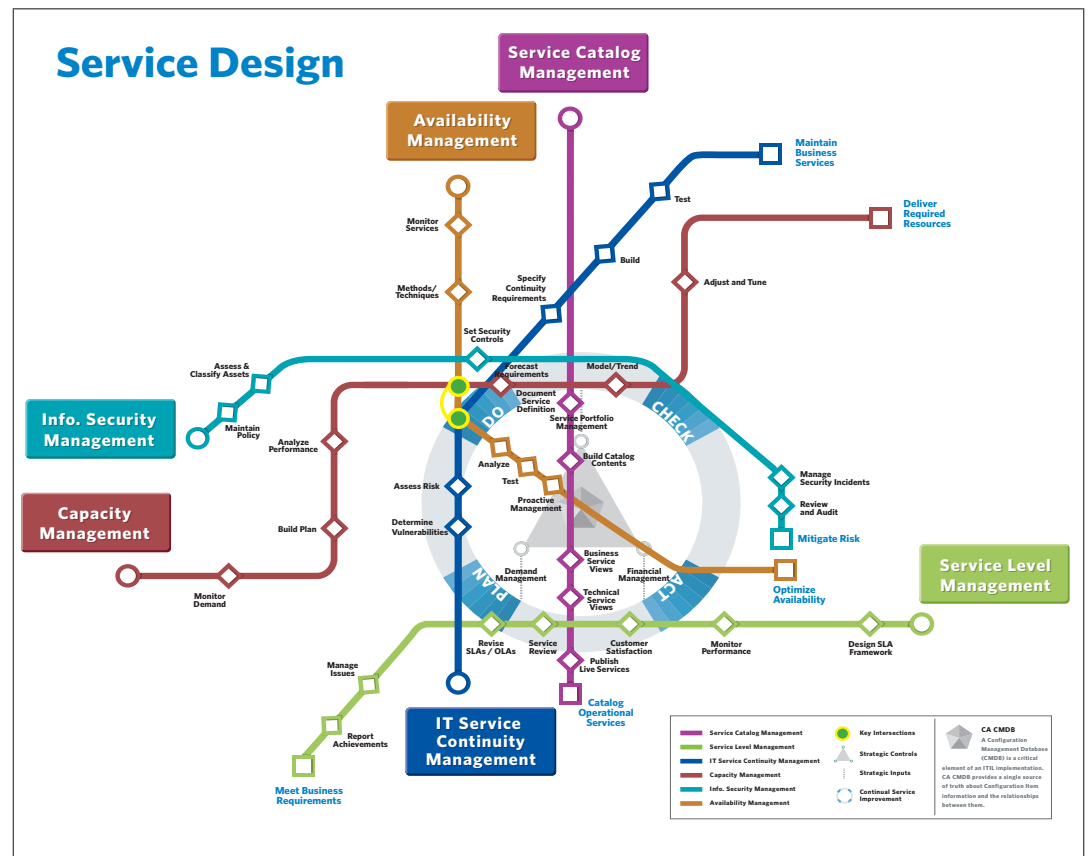


CA has developed three maps that portray the critical ITIL disciplines that most ITSM discussions focus on: Service Design, Service Transition and Service Operation.

SERVICE DESIGN MAP

The Service Design map represents a journey of developing and improving capabilities for engineering and maintaining the appropriate levels of services in production. The Capacity Management journey is drawn in this map.

FIGURE B



Finding the Right Path to IT Service Excellence

Most organizations have multiple disciplines in IT that respond directly to the matter of service availability. That is, many of the functions that decide or constrain availability are active already but are not strategically coordinated to optimize their collective impact.

Key points in this optimization include:

- Embracing the role of architecture
- Unifying service visibility
- Managing service components using business priorities

A primary outcome of the Capacity Management process is the assurance that cost-justifiable capacity is available when required across the enterprise, and that there is a close match between present and planned capacity and the current and future needs of the business.

SECTION 2: OPPORTUNITY


Capacity Management

The objective of Capacity Management is to ensure that your infrastructure provides the cost-justifiable resources needed to meet current and future business service requirements, while ensuring that IT resources are provisioned, managed, utilized and acquired in a cost-effective manner.

Capacity Management activities receive inputs from a number of the other ITIL processes. For example, the Incident Management process informs Capacity Management when incidents occur due to capacity problems. IT Service Continuity Management provides failover and disaster recovery capacity requirements for capacity planning. The outputs of Capacity Management also support many other processes. For example, Capacity Management can inform Change Management of the need for additional capacity, or about the potential impact of a new service on current capacity levels. In addition, Capacity Management is instrumental in ensuring compliance with service levels established through the Service Level Management process. These integration points are represented in Figure B.

The goals of Capacity Management are to manage all capacity and performance-related service and resource matters. The Capacity Manager must:

- Create and maintain a Capacity Plan
- Act as an expert and advisor on all capacity and performance issues and needs
- Manage the environment to meet or exceed established service level agreements (SLAs), while understanding the impact on both services and resources
- Take the role of problem manager for capacity and performance problems
- Evaluate all changes as they might relate to performance and capacity
- Be proactive while being cost-aware



This translates to the following roles, which may or may not be the function of a single person or single group:

- Performance monitor — observe patterns of both IT and business activity, perform reporting activities and solve performance problems noted
- Performance tuner — provide for efficient delivery of SLAs with existing resources
- Capacity planner — understand current and future resource demands and create a forecast and capacity plan with sufficient future focus to allow time to react
- Capacity and performance policy manager — define thresholds and other policies for planning purposes
- Resource demand manager — manage resources efficiently through policies

The key Capacity Management activities take place at the various “stations” illustrated in Figure C, en route to the process terminus, delivering the resources required to ensure that IT operations and IT Infrastructure support the future capacity of IT services that are required by the business in the most cost-effective manner. In other words, Capacity Management assures that currently delivered services and the service infrastructures support the needs of the business now and in the future. It is important to keep in mind, however, that Capacity Management is also about understanding the potential for future Service Delivery, so that new technology can be leveraged where appropriate to cost-effectively deliver the services required by the business.

This process is a balancing act:



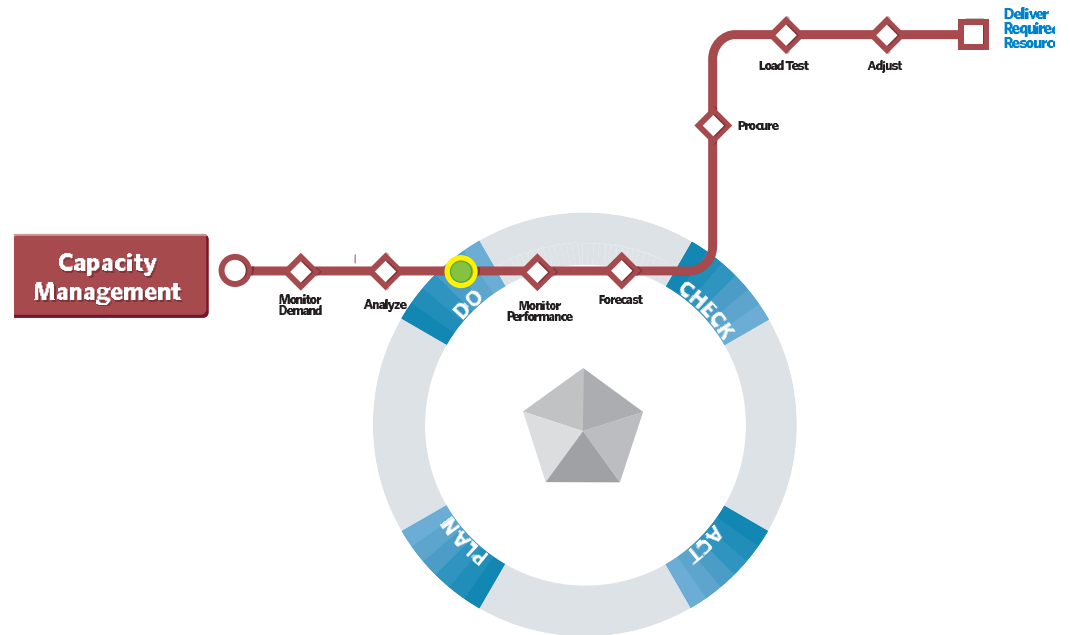
There are a number of stops on the Capacity Management process line:

- Monitor demand — Ensure that the capacity of the IT infrastructure matches the evolving demands of the business in the most cost-effective and timely manner. Capacity Management can also influence the demand for resources, perhaps in conjunction with Financial Management.
- Analyze — Process input from performance management, resource management, workload management, demand management, and application sizing, in order to understand the demands currently being made for IT resources and producing forecasts for future requirements.
- Monitor performance — Identify and respond to variations in performance, ensuring the most efficient resource usage while meeting SLAs.
- Forecast — Produce an annual Capacity Plan enabling the IT service provider to continue to provide services of the quality defined in SLAs. This activity includes recommendation and acquisition of resources, testing and delivery.

CAPACITY MANAGEMENT

The Capacity Management track ensures that cost-justifiable IT capacity in all areas of IT always exist and is matched to the current and future agreed needs of the business in a timely manner.

FIGURE C



Capacity Management Sub-Processes

Capacity Management consists of a number of sub-processes, within which there are various activities. The sub-processes of Capacity Management are:

BUSINESS CAPACITY MANAGEMENT This sub-process is responsible for ensuring that the future business requirements for IT services are considered, planned for and implemented in a timely fashion. This timeliness can be achieved by analyzing data on current resource utilization and gathering business projections and application design plans, in order to forecast and model future requirements. As with many ITIL activities, successful Business Capacity Management involves close collaboration between IT and the business.

SERVICE CAPACITY MANAGEMENT The focus of this sub-process is the management control and prediction of the performance of the operational IT services that enable users to perform day-to-day activities. The focus must be on end-to-end service SLAs rather than the individual infrastructure components. If you focus narrowly on each IT component first, without understanding the end-to-end service, the organization risks investing in resources that may not be needed, and not investing where they are truly needed. In this sub-process, not only is engagement with the business essential, but cross-platform, cross-technology collaboration will also be required.

COMPONENT CAPACITY MANAGEMENT This sub-process focuses on the management of the individual components of the IT infrastructure. This is the “traditional” IT view, and is responsible for ensuring that all IT components that support a particular IT service are closely monitored and measured and that exercises are constantly performed to improve service capacity requirements. Again, a structured and top-down approach to Capacity Management ensures closer business alignment, and enables the prioritization of capacity investments to be based on business needs. It is critical to automate thresholds, where possible, to enable proactive management and minimize degradation and outages.



Monitor Demand

The Capacity Management journey (like many of the other journeys) begins with monitoring your current state; in this case, your current service capacity state. Monitoring demand is about understanding how well you currently support your users' demands for capacity. Monitoring of capacity demand levels can tell you how many users require resource and how much resource each business activity consumes. It also provides valuable information on the current capacity levels of resources so that IT will know how to support new services as they are introduced into the environment, as well as how new services will affect current SLAs. In this case, it is essential to understand effective capacity of a resource, not the theoretical textbook capacity. A given device may only be able to deliver acceptable performance at 50% busy; if you do not understand this threshold and assume demand can double in a 50% busy environment while still meeting SLAs, you may delay upgrades and tuning too long.


Demand monitoring is important because it allows you to collect relevant data for specific components of services that you can present to the business to facilitate smart and effective decisions. And this must be assessed to see patterns of demand based on date and time. With this information, you can make the right decisions every time.

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. The monitoring activity of Capacity Management includes:

- CPU & memory utilization
- I/O rates
- Device utilization
- Queue length
- Storage utilization
- Transaction rate
- Packet rate
- Response time
- Burst rate
- Bandwidth utilization

This monitoring activity can, in turn, trigger typical capacity initiatives such as: archiving, workload balancing, upgrade projections, tolerance levels, I/O tuning, pricing incentives to foster workload shifting and data storage tuning.

Monitored data can include CPU utilization or transaction response times (see below), among other data. However, there is a difference between the information collected around managing capacity (resource utilization) and the data needed to monitor performance (response time or throughput). Business decisions are driven by an understanding of both. Technology can enhance monitoring of thresholds and normal service levels; for example, if thresholds are exceeded, automated alerts can notify the Service Desk that there is a deviation from the expected level of capacity and the appropriate escalation procedure is followed. Ideally, the thresholds should trigger this activity before user impact is felt (soft threshold) to improve service level compliance.



So regardless of whichever sub-process (Business, Service or Component Capacity Management) is involved, monitoring of IT resources is a key activity of the overall Capacity Management process that should be performed in order to ensure that IT capacity continues to meet the requirements of the business.

Analyze

The data collected from demand monitoring activities should be analyzed to identify trends, from which you can establish the normal utilization and service level, or baseline. Some key Capacity Management metrics include:

- Processor usage by workload and application
- Relative resource consumption
- Minimum/maximum transactions per second
- Online response times and trends
- Batch throughput
- Resource utilization statistics
- Business usage trends
- Processor and I/O usage trends
- Network usage trends
- Workload trends and forecasts
- Growth forecasts

Trending allows you to understand what is considered “normal,” which serves as a benchmark to compare against any deviation from “normal.” Any deviation from expected utilization levels, thresholds or response times can be immediately detected and acted upon. (Thus, the trends you extrapolate from this analysis can also be highly relevant during the Incident and Problem Management processes.) These trends may be as granular as by hour, by day, because business workloads are not evenly distributed across the day or week. You may find such issues as bottlenecks, unbalanced workload distribution, application design inefficiencies, database design issues, inefficient use of memory and more.

One major value of capacity monitoring and analysis is that the information collected can be used to predict future activity. Technology can help you here, by predicting future resource usage, and by monitoring actual business growth against predicted growth. With the information analyzed, you can make decisions to improve performance or make more efficient use of system resources. A way to improve these capabilities is by “tuning,” which is best supported by technology to perform such techniques such as work load balancing.

Once monitoring and analyzing the data is underway, you need a way of storing the information you are gathering. Addressing this storage requirement is where the Capacity Management Information System (CMIS) comes into play. You might also store the capacity plan and the forecasts in this data store. The Capacity Management process collects data from a variety of hardware platforms and software applications that could be widely distributed, so technology helps considerably to organize that information and make it available in the CMIS. Make sure that the gathering and maintenance of your CMIS data is identified as business critical and controlled and protected by the Change Management process, so that Changes are considered relative to their impact on the CMIS.




Monitor Performance

At the Monitor Demand stop the discussion touched briefly on monitoring performance; specifically monitoring response times. A major intersection here at Monitor Performance is with Service Level Management, since SLAs often reference expected response times, including end user response time, components of response time, expected response times for service restoration, escalation and resolution. The monitoring of response times is a complex process, and technology can greatly ease this activity. Supporting technology can include network monitoring systems, transaction monitoring systems, distributed agent monitoring systems, passive monitoring systems, robotic scripted systems with terminal emulation software or the incorporation of specific code within client and server applications software that allow for real-time monitoring of user performance.

Forecast

Forecasting activities allow the business to intelligently predict future growth and plan accordingly for capacity. This can be done in a variety of ways depending on the technology in question. Armed with this forward-thinking approach, the business can decide to double the amount of Web-based users of a service, for example, and accurately predict the direct costs associated with the new usage consumption rates based on current data and growth patterns. If the capacity forecast indicates a requirement for increased resources, this then becomes an input to the IT budget cycle. Monitoring and analyzing expected service levels will allow you to forecast inadequate levels of service if capacity is not planned for with an understanding of your current state. Some forecasting techniques include:

- **Tuning** — the analysis of the monitored data may identify areas of the configuration that could be tuned to better utilize system resources or improve the performance of the particular service. Some valuable techniques include: balancing workloads and traffic, balancing disk utilization, defining a workable lock strategy to avoid unnecessary lock activity, efficient use of memory.
- **Implementation** — the objective of this activity is to introduce to the live operational service any changes that have been identified by the monitoring, analysis and tuning activities. Following implementation, go back to assessment to measure how closely the results matched expected outcomes.
- **Storage of Capacity Management Data** — Data in the CMIS is stored and used by all the sub-processes of Capacity Management. This repository holds a number of different data types; including business, service, technical, financial and utilization data.
- **Modeling** — A prime objective of Capacity Management is to predict the behavior of IT Services under a given volume and variety of work. Modeling is an activity where you can use activities relating to any of the sub-processes of Capacity Management. Modeling can predict the behavior of services (performance and resource utilization) under different circumstances. Depending on your time and budget, you can employ models using various techniques: base-lining, trend analysis, analytic modeling, simulation modeling, benchmarking.
- **Application Sizing** — the primary objective of application sizing is 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 objective, application sizing has to be an integral part of the applications lifecycle.

- 
- **Exploitation of New Technology** — Capacity Managers must make time to learn new ways of doing their job and new technologies available to them, so they can make prudent decisions when it is time to implement a new system or perform an upgrade. Each new technology needs to be assessed for its value to the business.
 - **Designing Resilience** — In all plans and decisions made by the Capacity Manager, ability to withstand failures and disasters should be considered and weighed against the cost and risk. In any event, part of capacity planning is to understand the impact of failures on system capacity and performance. Provisioning of appropriate spare capacity is often part of this exercise.

Creating the Capacity Plan

The objective of the Capacity Plan is to document the current levels of resource utilization and service performance; and, after consideration of the business strategy and plans, to forecast the future requirements for resources to support the IT services that, in turn, support the business activities.

Forecasting allows you to implement “what if” scenarios (see the discussion of the Load Test station below for further detail). This does not affect the current state of your IT services, but arms you with information to budget and plan for the future based on accurate data. The Capacity Plan is a forward-looking strategy based on forecasting—but it also looks backwards at the data from monitoring and analysis to accurately predict growth and capacity requirements.

Your Capacity Plan should assess the current IT capacity with respect to future workloads, performance and Service Level Requirements (SLRs); it should also align with the goals of the business. In a nutshell, the Capacity Plan includes management summaries around business, service and resource capacity information, cost information, and service target agreements—all of the information compiled through monitoring and analysis in order to forecast future business lines, which will include new services.

New Services (Application Sizing)

New services under consideration should be included in a Capacity Plan, which includes inputs from the Service Level Management process. This means that the SLRs of the service must be achievable, and that the SLAs of existing services will not be broken by poor Capacity Planning. Note that the Capacity Manager generally must be involved in design, because performance and capacity demands are incorporated into the application as it is designed, and the Capacity Manager can seldom compensate for these issues in production.

Capacity Management can support the plan for new services by helping SLM to understand customers’ capacity requirements; i.e. required response times and service volumes. Capacity Management can help Service Level Management in the negotiation process by providing possible solutions to a number of scenarios. Additionally, Capacity Management can support the design and procurement of new services, arriving at an intersection with Availability Management, to ensure proper services design.



Procure

Design and procurement of new services should be based on the requirements of the business and on your understanding of your capabilities to support those requirements. This is possible when there is an understanding of your current capacity. Capacity Management provides the business with the information required to make financial decisions, such as which components to upgrade and when to purchase new hardware to support future capacity requirements.

Capacity Management should be involved in the design of new services; and it should make recommendations for the procurement of hardware and software, where performance and/or capacity are factors. In the interest of balancing cost and capacity, the Capacity Management process obtains the costs of proposed solutions and the business decides “to procure or not to procure” based on that information.

Load Test

At this stop, the Capacity Manager makes sure that IT infrastructure components do not “break” when all applications are running concurrently at “full load.” This information will provide you with the ability to support the business when business stakeholders are forecasting and drafting time lines for new application service offerings to support a business initiative. Load Test includes scenario testing and technology can assist in creating “what if” scenarios (similar to modeling activities) that can present data to support the best plan of action.

Adjust

In line with ITIL’s emphasis on continuous improvement, at the Adjust station you are making adjustments to your systems to ensure they are running optimally after you have carried out the various load tests. New services are procured and, before you release those services into the environment (per Change Management and Release Management best practices), you conduct activities like load testing and then adjust your systems based on the data you receive. This optimizes the ability of Capacity Management to meet its goal of cost-effectively supplying capacity to match the needs of the business. Adjustments should be considered and evaluated at regularly scheduled intervals to ensure compliance with business needs.

Demanding Management—A Form of Adjustment

Demand Management seeks to influence the demand for computing resources and the use of those resources. In other words, Demand Management is all about manipulation! For example, when your current capacity levels are insufficient, you may want to manipulate usage because this meets the requirements of the business and there is no budget to increase capacity levels.

Either way, Demand Management is a savvy way to maintain services that are critical to the business and maximize the benefit of Capacity Management. But smart Demand Management requires understanding your current state: which services are utilizing certain resources and at what level. Then you can decide if it is even possible to manipulate the use of that resource.

Manipulating or influencing the use of services can involve either a physical restriction or a financial restriction. Physical restrictions can be supported by technology to limit the number of concurrent users of a resource. A financial restriction might increase the cost of one resource, making utilization of another (less costly) resource more attractive. But again, the key to successful Demand Management is to understand the business requirements and the demands on IT services, and to communicate effectively with your customers.

SECTION 3: BENEFITS

Benefits of Capacity Management Best Practices

The benefits of implementing Capacity Management in line with ITIL best practices include:


- Spending money on the right things at the right time
- The ability to utilize the systems and budget you have in the most effective manner possible
- The ability to proactively manage systems to reduce incidents and problems
- Reducing the risk of service degradation through better management of IT resources
- Reducing the risk resulting from changes to the IT infrastructure, made possible through application sizing and other monitoring and analysis activities
- Increasing customer satisfaction
- The foresight to anticipate and meet future capacity requirements before business processes are impacted

SECTION 4: CONCLUSIONS

Capacity Management is always a balancing act, such as cost against capacity; i.e. the need to ensure that processing capacity that is purchased is not only cost-justifiable in terms of business need, but also that the organization makes the most efficient use of those resources. Similarly, Capacity Management constantly balances supply against demand; making sure that the available supply of processing power matches the demands made on it by the business, both now and in the future. It may also be necessary to manage or influence the demand for a particular resource, perhaps in conjunction with Financial Management for IT Services.

Here are some suggestions for avoiding common problems on your Capacity Management journey:

- Informal or immature service management processes may exist that are key inputs to Capacity Management. For example, SLAs may exist that ignore input from Capacity Management, which therefore leads to under- or over-utilization of some of the capacity resources. Or SLAs may be informal, which leads to misunderstandings and inability to provide the service the business requires. Make sure processes are formalized and documented.
- Some IT organizations view Capacity Management as a responsibility of all senior managers and therefore don't see the need for a Capacity Manager role. With no individual or group held accountable, you will not achieve optimum results from your Capacity Management plan. When an organization appoints someone to "own" the Capacity Management process, the owner must have the level of authority and empowerment required to influence all the relevant areas of IT to achieve the process goal.
- Make sure you have the right resources with the required skills and competencies and the right technology and tools in place to support the Capacity Management process.
- Some roles or individuals may be resistant to the value of Capacity Management. Clear communication of its value from the top down is the best way to communicate successes in business terms to all stakeholders.

- 
- If current levels of capacity are considered adequate it is sometimes perceived that there is no business justification to deploy the Capacity Management process. Since Capacity Management is a much larger issue than simply providing enough capacity, make sure you communicate clearly the role a Capacity Manager plays.
 - Be sure to define an appropriate scope for Capacity Management so you can accomplish the goals and reap the benefits of the process in a timely and cost-effective manner. Start with a smaller scope and increase your sphere of influence as you achieve success.

SECTION 5: ABOUT THE AUTHOR

Denise Kalm is a product marketing director at CA. She has 30 years of experience in IT, including work in application programming, enterprise systems management, and performance management and capacity planning. She is an ITIL Capacity Management practitioner and frequently speaks and writes on capacity planning, performance management, and other issues for organizations such as CMG, where she serves as a national director, regional secretary and Chief Editor of Measure IT.

Marv Waschke is a senior advisor, product management, at CA. He represents CA in the Distributed Management Task Force (DMTF) Configuration Management Database (CMDB) Federation Work Group, and the DMTF Cloud Incubator group. He is also a representative to the W3C Service Modeling Language Work Group and a member of the CA Council for Technical Excellence. He has been designing and building service management solutions for IT for more than 20 years.

To learn more about the CA ITIL solutions, visit ca.com/itil.

CA, one of the world's largest information technology (IT) management software companies, unifies and simplifies the management of enterprise-wide IT for greater business results. Our vision, tools and expertise help customers manage risk, improve service, manage costs and align their IT investments with their business needs.

