

## Document Purpose

This Practices Guide is a brief document that provides an overview describing the best practices, activities, attributes, and related templates, tools, information, and key terminology of industry-leading project management practices and their accompanying project management templates.

## Background

The Department of Health and Human Services (HHS) Enterprise Performance Life Cycle (EPLC) is a framework to enhance Information Technology (IT) governance through rigorous application of sound investment and project management principles, and industry best practices. The EPLC provides the context for the governance process and describes interdependencies between its project management, investment management, and capital planning components. The EPLC framework establishes an environment in which HHS IT investments and projects consistently achieve successful outcomes that align with Department and Operating Division goals and objectives.

Capacity is defined as the maximum amount or number that can be received or contained. For example, the amount of data that a computer hard disk can store is the disk's capacity. The maximum possible data rate received over a communication channel under ideal conditions is its capacity. Capacity can also refer to non-technical things such as the maximum amount of work that an organization is capable of completing in a given period of time or the maximum number of people that can physically occupy a room.

Discrepancies in capacity and demand results in inefficiencies associated with either under-utilized resources or unmet user demand. The goal of capacity planning is to minimize this discrepancy and to provide satisfactory service levels in a cost-efficient manner. The Information Technology Infrastructure Library (ITIL) defines capacity management as supporting the optimum, and cost effective, provisioning of services by helping organizations match their resources to the business demands.

## Practice Overview

Capacity planning doesn't always mean planning for periods of peak demand. Capacity requirements can vary greatly from times of peak demand to times of limited demand. As a result there may be drastic differences in the resources required to maintain normal operations during periods of peak demand. When planning for what levels of capacity will be supported it's important to understand and weigh the costs, benefits, and risks of delivering, or not delivering, a certain level of capacity.

Depending on the type of project, it's possible that true capacity planning may play a small role in the overall project life cycle. For example, assume a project is only responsible for creating a small database application that will be stored on a shared server within a data center. The server's capacity is the responsibility of the data center manager. The project manager may not have any real control over the server at all. However, the project manager is responsible for ensuring that the business requirements are met. For example, if it is required that the database be accessible 100% of the time, the project manager needs to ensure this is the case. Accomplishing this requires working closely with the capacity provider. The data center manager would then be responsible for maintaining adequate server capacity to fulfill the business need.

Capacity planning is simpler if an existing system is in place and needs only adjustments to meet current or expected changes in capacity. However, when planning capacity for a new system, analysis of information such as historical project archives, industry standards, and information obtained from vendors/customers, etc. can assist in estimating expected demands upon that system. Regardless of which situation a project is planning for three basic steps for capacity planning include:

- *Determine capacity requirements* by understanding what work will be performed by the system and then quantify the user experience as it relates to that work

- *Analyze current capacity* to determine how it is meeting the requirements of the system and needs of the users
- *Plan for future capacity* by forecasting future business activities and system requirements. Implement required changes to ensure sufficient capacity to meet demands upon the system

### Determine Capacity Requirements

- *Define required system performance to support associated business and process workloads.* Capacity planning defined in business terms rather than utilization of CPU, memory, hard drive space, etc. simplifies planning for impacts on people, processes, growth, etc. For example, it's easier to predict the growth of a business unit and adjust capacity to accommodate an expanding work force than it is to predict an increase in transaction volume resulting from such growth. Once defined, capacity requirements are then later used as inputs into identifying the planning of the physical infrastructure (technology, building, staff, etc) necessary to support the business requirements.
- *Contact prospective capacity providers as early in the project's life cycle as possible.* It's important to make individuals and departments aware of the prospective impact a project may have on their organizations. The earlier this can be done the more easily it can be mitigated.
- *Agree upon a unit of measure to assess conformance to defined system performance requirements.* Note that the goal of capacity planning is to provide satisfactory services in a cost efficient manner. As a result, this unit of measure should be defined in terms of business units as well as the technical requirements necessary to support them. Associate it with a measurable quantity of work as opposed to the amount of system resources required to accomplish that work. For example, when deciding on a unit of work to measure the satisfactory servicing of patients at a hospital consideration may include the number of patients serviced, the number of patient records processed the amount of drugs dispensed, etc.
- *Test the system against defined business requirements and make adjustments where needed.* This relates more to IT performance, as opposed to business functions, necessary to meet business needs. This may be accomplished by adjusting IT resources such as CPU utilization, memory usage, disk space, etc.
- *Agree upon minimum levels of service (capacity) the system must provide.* These requirements are often defined by clients/users and should outline measurable system performance in terms that make sense to clients/users. This typically includes items such as response time, processing time for each request, minimum number of requests that can be processed in a given period of time, etc. Defined in these terms it's easier to illustrate success and make adjustments to the technical aspects of the system as necessary to meet business objectives.

### Analyze Current Capacity

- *Understand the organizations current capabilities.* Before assuming that additional capacity is necessary a detailed understanding of current available capacity is required. A capacity study may need to be performed to evaluate the organizations current capacity.
- *Check usage of system resources necessary to support business requirements (CPU, memory, hard drive, etc).* This relates more to IT performance, as opposed to business functions, necessary to meet business needs. If no current system exists it becomes even more important to understand and evaluate all available information regarding expected demands upon that system. This may include information such as historical project archives, industry standards, information obtained from vendors/customers, etc
- *Record and track utilization of system resources* to determine where capacity adjustments need to be made to support business processes as defined by the client

### Plan for Future Capacity

- *Forecast expected workloads for a particular period of time.* Understand how changes in workloads affect business processes and the system that was built to support them. Translate that into technical requirements necessary to maintain the system at a level that satisfies user demands.
- *Plan for future system usage and plan adjustment to accommodate varying demands upon the system.* Create a capacity plan that outlines the current configuration, required future configuration, and the steps necessary to accomplish any necessary system changes. Identify key capacity areas/items needing to be monitored and expected growth rate of these items. Include a defined response strategy for growth. Thresholds should be defined that represent utilization levels requiring action necessary to increase capacity. Such as utilization at 75%, 1,000 users, 10,000 visitors per day, etc. The goal is to mitigate the risk of a capacity bottlenecks before it becomes an

issue. Specific actions should be taken early in response to increase in defined. Contingency plans then define what actions may be taken, in response to identified capacity triggers, in hopes of reducing the impact of capacity issues not mitigated through responses to growth thresholds. Depending on the organization, capacity planning may be applied using one of three strategies:

- Lead strategy adds capacity in anticipation of an increase in demand
- Lag strategy adds capacity after demand has increased beyond existing capacity
- Match strategy adds capacity incrementally in response to changes in demand

It's important to note that capacity planning, from the perspective of the project team, is often different than the capacity planning of those supporting or providing the required capacity. Often tradeoffs need to be made to accommodate capacity limitations, quality issues, budget concerns, etc. Once these tradeoffs have been agreed upon they should be recorded within a supporting document such as a statement of work (SOW), support agreement, formal contract, etc.

For example, from the perspective of the project team, a system must support 1,000,000 users and respond to user commands within 1 millisecond. The hardware and network infrastructure to support these requirements must cost less than \$100K. From the perspective of the organization providing the resources required to meet the objectives, the hardware and network infrastructure needed includes 4ghz CPU, 4GB of RAM, 100GB of hard drive space, etc and will cost \$800K. For the \$100K that the organization can support, only 100k users with a system response time of 3 milliseconds is feasible. The organization may be forced to make decisions based on what capacity is required and what capacity can actually be delivered with in the limitations of the project parameters.

Capacity demands upon an organization change based on demands for that organization's resources. For example, as an organization grows and new users are added to an existing email system, the capacity of the computer hardware and network resources must increase to meet or exceed the demand. Capacity planning determines what capacity is needed to meet current and future demands as well as planning for any extenuating circumstances that may cause a temporarily impact exponentially greater than that normal operations. Capacity planning requires on-going monitoring of existing systems, predicting future needs and impacts, and making intelligent and well-timed decisions to adjust system capabilities to meet user demands. When done correctly the process is seamless to users and business processes are minimally impacted by system adjustments

## Practice Best Practices

- **Understand the work** – Relate capacity requirements to what work needs to be supported
- **Translate** – Translate the capacity requirements necessary to support business functions into technical requirements
- **Evaluate** – Constantly evaluate current capacity against requirements defined by the client
- **Plan** – Plan for future capacity requirements and the incremental step necessary to achieve them
- **Research** – Research other projects, their capacities, and utilizations. If resources are not being fully utilized explore the possibility of leveraging other projects to reduce cost and time to delivery
- **Business Requirements** – Confirm that the individuals measuring and estimating capacity requirements understand the business requirements driving the needs for the capacity
- **Change Control** – A key management component is the practice of managing change
- **Tools** – Some tools for managing capacity may include:
  - **Virtualization Software** – manage computers, share resources, and save money
  - **Source Control Management Tools** – improve software development capacity
  - **Sub-Contractors** – manage temporary fluctuations in staffing requirements
- **Types of Capacity** – The following list is in no way a full range of capacity items. When planning capacity requirements consider items such as physical server consolidation, storage, necessary IT controls, standards and processes, computing resources, required technical skills, additional staffing, requirements for licensing, backups, security, system configuration, monitoring, and availability, etc.

## Practice Activities

1. Determine capacity requirements
  - Define required system performance to support associated business and process workloads
  - Agree upon a unit of measure to assess conformance with defined system performance requirements

- Test the current system against defined business requirements and make adjustments where needed
  - Agree upon minimum levels of service (capacity) the system must provide
- 2. Analyze current capacity
  - Check existing usage of system resources
  - Record and track utilization of existing system resources
  - Understand historical capacity utilization and any available industry standard usage data
- 3. Plan for future capacity
  - Forecast expected workloads
  - Plan for future system usage