SYSTEMS CAPACITY PLANNING

It is very common for an IT organization to manage system performance in a reactionary fashion, analyzing and correcting performance problems as users report them. When problems occur, hopefully system administrators have tools necessary to quickly analyze and remedy the situation. In a perfect world, administrators prepare in advance in order to avoid performance bottlenecks altogether, using capacity planning tools to predict in advance how servers should be configured to adequately handle future workloads.

The goal of capacity planning is to provide satisfactory service levels to users in a cost-effective manner, ensuring that information systems are properly designed and configured to meet current and future organizational capacity and performance requirements.

Three Steps for Capacity Planning
There are three basic steps for IT systems capacity planning:

• **Determine Service Level Requirements**
  The first step in the capacity planning process is to categorize the work done by the systems and to quantify user’s expectations for how that work gets done.

• **Analyze Current System Capacity**
  Next, the current capacity of the system must be analyzed to determine how it is meeting the needs of the users.

• **Plan for the Future**
  Finally, using forecasts of future business activity, future system requirement are determined. Implementing the required changes in system configuration will ensure that sufficient capacity will be available to maintain service levels, even as circumstances change in the future.

STEP 1: Determining the Service Level Requirements

• **Define workloads**
  From a capacity planning perspective, a computer system processes workloads (which supply the demand) and delivers service to users.
During the first step in the capacity planning process, these workloads must be identified and a definition of satisfactory service must be created.

A workload is a logical classification of work performed on a computer system. A workload can be classified by either company’s departments or by types or number of transactions being processed per hour.

For example the following figure gives a view of workload on an organization’s server by the departmental usage:

![Pie chart showing workload distribution by department]

It is useful to analyze the work done on systems in terms that make sense from a business perspective, using business-relevant workload definitions. For example, if you analyze performance based on workloads corresponding to business departments, then you can establish service level requirements for each of those departments.

Business-relevant workloads are also useful when it comes time to plan for the future. It is much easier to project future work when it is expressed in terms that make business sense. For example, it is much easier to separately predict the future demands of the human resources department and the accounts payable department on a consolidated server than it is to predict the overall increase in transactions for that server.

- **Determine the unit of work (measurement units)**
  For capacity planning purposes it is useful to associate a unit of work with a workload. This is a measurable quantity of work done, as opposed to the amount of system resources required to accomplish that work.

  We accomplish work using resources such as disk, I/O channels, CPUs and network connections. Measuring the utilization of these resources is important for capacity planning, but not relevant for determining the amount of work done or the unit of work. Instead, for an online workload, the unit of work may be a transaction. For an interactive or batch workload, the unit of work may be a process.

  In large organizations that rely heavily on its information systems, the measurement unit ‘transaction’ becomes too simple and thus inappropriate to compare such heavy volumes. In these cases we can derive more complex measurement units from this, for example:

  - For databases, we may use the number of transactions being executed per second (i.e., Transaction/sec.)
  - For CPU usage, we may use number of million of instruction executed per second.
  - For network usage, we may use number of packets being transmitted per second. (i.e., packets/sec.)
• **Identify service levels for each workload**
The next step now is to establish a service level agreement. A service level agreement is an agreement between the service provider and service consumer that defines acceptable service. The service level agreement is often defined from the user’s perspective, typically in terms of response time or throughput. Using workloads often aids in the process of developing service level agreements, because workloads can be used to measure system performance in ways that makes sense to clients/users.

Ideally, service level requirements are ultimately determined by business requirements. Frequently, however, they are based on past experience. It’s better to set service level requirements to ensure that you will accomplish your business objectives.

**STEP 2: Analyze the Current Capacity**

There are several steps that should be performed during the analysis of capacity measurement data.

• **Measure service levels and compare to objectives**
  First, compare the measurements of any items referenced in service level agreements with their objectives. This provides the basic indication of whether the system has adequate capacity.

• **Measure overall resource usage**
  Next, check the usage of the various resources of the system (CPU, memory, and I/O devices). This analysis identifies highly used resources that may prove problematic now or in the future.

  It is important to take a look at each resource within your systems to see if any of them are saturated. If you find a resource that is running at 100% utilization, then any workloads using that resource are likely to have poor response time.

• **Measure resource usage by workload**
  Look at the resource utilization for each workload. Ascertain which workloads are the major users of each resource. This helps narrow your attention to only the workloads that are making the greatest demands on system resources.

• **Identify components of response time**
  Determine where each workload is spending its time by analyzing the components of response time, allowing you to determine which system resources are responsible for the greatest portion of the response time for each workload.

  It should be determined what system resources are responsible for the amount of time that is required to process a unit of work. The resources that are responsible for the greatest share of the response time are indicators for where we should concentrate our efforts to optimize performance.

**STEP 3: Planning for the Future**

How do you make sure that a year from now your systems won’t be overwhelmed and your IT budget over extended? Your best weapon is a capacity plan based on forecasted processing requirements. You need to know the expected amount of incoming work, by workload. Then you can calculate the optimal system configuration for satisfying service levels.
• **Determine future processing requirements**
  Systems may be satisfying service levels now, but will they be able to do that while at the same time meeting future organizational needs?

  Besides service level requirements, the other key input into the capacity planning process is a forecast or plan for the organization’s future. Capacity planning is really just a process for determining the optimal way to satisfy business requirements such as forecasted increases in the amount of work to be done, while at the same time meeting service level requirements.

  Future processing requirements can come from a variety of sources. Input from management may include:

  • *Expected growth in the business*
  • *Requirements for implementing new applications*
  • *Planned acquisitions or divestitures*
  • *IT budget limitations*
  • *Requests for consolidation of IT resources*

  Additionally, future processing requirements may be identified from trends in historical measurements of incoming work such as orders or transactions.

• **Plan future system configuration**
  After system capacity requirements for the future are identified, a capacity plan should be developed to prepare for it. The first step in doing this is to create a model of the current configuration. From this starting point, the model can be modified to reflect the future capacity requirements. If the results of the model indicate that the current configuration does not provide sufficient capacity for the future requirements, then the model can be used to evaluate configuration alternatives to find the optimal way to provide sufficient capacity.

**Capacity Planning Process**

In summary, we have shown these basic steps toward developing a capacity plan:

• **Determine service level requirements**
  o Define workloads
  o Determine the unit of work
  o Identify service levels for each workload

• **Analyze current system capacity**
  o Measure service levels and compare to objectives
  o Measure overall resource usage
  o Measure resource usage by workload
  o Identify components of response time

• **Plan for the future**
  o Determine future processing requirements
  o Plan future system configuration

By following these steps, you can help to ensure that your organization will be prepared for the future, ensuring that service level requirements will be met using an optimal configuration. You will have the information necessary to purchase only what you need, avoiding over-provisioning while at the same time assuring adequate service.