DITY™ Weekly Newsletter
itSM Solutions® IT Experience. Practical Solutions
The workable, practical guide to Do IT Yourself Vol. 5.21 • May 26, 2009

Capacity Management In 90 Days

By Hank Marquis
Hank is EVP of Knowledge Management at Universal Solutions Group, and Founder and Director of NABSM.ORG. Contact Hank by email at hank.marquis@usgct.com. View Hank's blog at www.hankmarquis.info.

A key requirement for any IT organization is to ensure capacity to meet the evolving demands of the business. Many IT executives think "upgrade" but then hear "capacity" - relating "a capacity plan" to "a spending plan."

However, best practice can not only identify, justify and meet business needs for capacity, but also reduce costs -- without making new investments in people or products.

The ability to predict future business demand is critical to effective Capacity Planning. Projected Saturation is a means to understand future requirements and plan accordingly for cost effective capacity planning.

This article introduces a practical and easy to understand guide to effective capacity planning for today's dynamic, customer-driven IT environment.

Effective Capacity Planning.

The IT Information Library™ (ITIL®) states the goal of effective capacity planning is to "Ensure that the capacity of the IT infrastructure matches the evolving demands of the business in the most cost-effective and timely manner."

As described in the ITIL Capacity Management process (see V3 Service Design publication), effective capacity planning starts with customer experience and measuring IT Service quality compared to customer requirements as expressed in Service Level Agreements (SLAs).

There are three important "management views" of capacity. The three views build upon one another and deliver a roadmap for effective capacity planning:

Resource Capacity is measuring and monitoring all components comprising IT Services. Practically, this means reporting on the utilization of discrete items like routers, switches, transmission links, servers, etc.

Service Capacity is ensuring that IT Services meet Service Level Requirement (SLR) targets within Service Level Agreements (SLAs). Practically, this means measuring and monitoring IT Service performance. Service Capacity is a function of the Resource Capacity of all the service CIs.

Business Capacity is planning and implementing IT Services to meet future business requirements. Practically, this means trending and forecasting of projected IT Service saturation based on observed Service Capacity.

The three views of capacity build upon one another and deliver a roadmap for effective capacity planning.

The views of Capacity Management provide a top-to-bottom understanding of IT Service consumption. *Resource Capacity Management* provides the individual details of capacity utilization. *Service Capacity Management* delivers the end-to-end view of service performance; effectively the result of the utilization of all its components. *Business Capacity Management* ensures that IT maintains sufficient capacity to meet current and future business demands. Of course, the question remains, how do you actually measure capacity consumption?

Peak Hour Load.

Peak Hour Load (PHL) provides a method for understanding business consumption of IT resources ("load") as well as providing a basis for predicting future load ("demand").

Peak Hour Load (PHL) indicates the busiest hour of the workday for any CI. PHL can provide a target -- if you meet the required PHL then you have enough capacity. PHL is very useful because it describes the CI at its busiest point. Business activities drive the capacity planning process, and PHL shows the maximum required capacity.

While there is no one perfect metric, PHL is useful by itself, and when mapped over time identifies those periods during the week, month, or year that require even higher capacity. For example, charting PHL over a month can show the busiest days of the month as well.

Use PHL as the primary metric for determining capacity utilization and requirements for a CI. Then use PHL to size the CI such that the peak-hour load is smaller than the CI capacity. The difference between peak-hour load and the capacity of the CI reflects the growth capability.

Chart PHL over time to understand all the peaks and troughs unique to capacity requirements for your business and industry. Create a workload catalog showing utilization over the day, week, month, and quarter the PHL for CIs and IT Services. Look for marked contrasts between average and peak loads -- these represent your liabilities and your opportunities.

After understanding the relationships between IT Services and their CIs, the first step in effective capacity planning is to establish the PHL for CIs. The next step is to determine how close the CI is to saturation -- saturation occurs when CI utilization nears its total capacity.

Projected Saturation.

Utilization should be somewhere below total capacity. Nearly every CI has an optimal performance capacity somewhat below its maximum capacity. In addition, the difference between optimal performance and maximum performance limits growth.

To predict saturation, measure, average and trend the PHL of the CIs that make up an IT Service. Instead of using PHL directly, you have to remove the high and low samples and make the projection based on the midrange of samples. This takes into account the variations in business-cycle activities such as weekends, month-end-closings, etc. You can easily trend the PHL into the future using the built-in trending features of office spreadsheet programs.

Using this method to project future PHL you can determine the projected saturation date. An important element in projecting saturation is the service cycle time, or how long it takes to upgrade the CI. For example, it may take 30 days to upgrade a CI. This means that you need to move the projected saturation date back in time from the actual saturation date. If you determine that CI demand will exceed its capacity in 90 days, then you need to upgrade capacity in 60 days. Failure to factor in this extra time could mean that you exceed capacity, need an upgrade, but cannot perform the upgrade before you need it!

Compare the projected saturation date with service cycle time for the CI. Initiate remedial action for any CI with a projected saturation date that is before the end of the service-cycle-time.

Reducing Waste.

Once the peak-hour load is available, it is easy to identify under-utilized CIs. Consider any CI with a peak-hour load below a given “downgrade threshold” as a candidate for potential downsizing or removal. For example, a T1 (1.536MB/sec) with a peak-hour load of 256KB/sec is a candidate for downsizing, perhaps to 512KB/sec.

The downgrade threshold can be set to indicate the level at which the next lower increment of capacity becomes more economical or the point at which to eliminate or re-deploy the CI.

Another way to conserve capital is to re-deploy assets. For example, initially, business demands required a large, fast, and well-equipped server. However, as the business changed, the requirements upon the server changed, and now an expensive asset is under-utilized. If another application requires a similar server, perhaps because it is approaching saturation, it is often more cost effective to replace the under-utilized server with a smaller unit, and re-deploy the larger unit.

When you know the utilization of assets, and can identify future requirements, you can avoid or reduce purchases by re-deploying existing assets. Most organizations suffer from idle or redundant capacity; locating yours can result in

significant cost reductions.

Balancing Demand.

The final step in effective capacity planning is to look for ways to shift or reduce capacity demands through organizational or procedural changes. CIs with extremely "bursty" usage are good candidates for demand balancing. Comparing the average load of a CI with its peak hour load identifies such candidates. You can even create a metric for this balance of peak-to-average load and track it as a part of capacity planning.

Many demand management activities affect the work habits of individuals and procedures within the organization. Examples include rescheduling batch jobs, deferring data entry, staggering work shifts, placing usage restrictions on individuals during peak usage times.

Balancing utilization often affects the peak hour load of other CIs. Sometimes this balancing provides opportunities for reducing costs through eliminating, downsizing, or deferring upgrades to affected CIs. In this scenario, you can use the "balance factor" of peak-to-average load to identify those CIs that might be candidates for balancing.

This is the most difficult of the capacity planning activities and requires information on the daily usage-pattern of CIs, knowledge of the business activities of the organization, and the ability to institute organizational change. It is, however, an important part of the process. During this step, a highly effective IT manager can produce significant savings for the organization while increasing system responsiveness.

A 90-day Capacity Plan.

Following is a simple 90-day plan to implement effective capacity management.

- Month 1
  - Create a workload catalog charting peak hour load, average load, and maximum capacity of resource-level CIs.
  - Measure service quality based on User ability to work -- that is, latency.
  - Identify CIs that are responsible for the latency "bottlenecks" in those services not delivering as required.
- Month 2
  - Using your workload catalog, check for over and under -utilized CIs.
  - Project the historical demand levels into the future.
  - Supplement current capacity before future bottlenecks occur.
- Month 3
  - Eliminate, downsize, or redeploy any under-utilized CIs to reduce waste.
  - Balance demand to increase utilization of available capacity through organizational change.
  - Repeat the entire process.

Summary.

Effective capacity planning requires much more than a capacity plan. It requires monitoring CIs at a detailed level in order to produce useful metrics like PHL. Trending PHL produces a workload catalog that forms the basis of your capacity plan.

Using PHL and projected saturation you can get a grip on your capacity needs and redeploy, downsize and optimize as appropriate. You will also then have a clear picture of what you really need, and when and where you really need it. This is the basis of a justified and documented cost plan.