“Effective Capacity Planning in an ITIL World”

An itSM Solutions® LLC
Success-Based White Paper™

This paper provides executives with Information Technology responsibilities a 90-day roadmap to success for effective capacity planning. Based on the best practices of the IT Infrastructure Library™ (ITIL®) but with a practical workable approach, the author shows how to improve customer satisfaction, reduce negative business impact, and lower costs without making large investments.

By
Hank Marquis
hank.marquis@itsmsolutions.com

June 2006
Updated 22 June 2006

Copyright and Trademark Notice

Notice of Rights / Restricted Rights Legend

For additional information, please contact:

itSM Solutions LLC
31 South Talbert Blvd #295
Lexington, NC 27292
Phone (336) 510-2885
Fax (336) 798-6296
Web http://www.itsmsolutions.com

Notice of Liability

This guide is distributed "As Is," without warranty of any kind, either express or implied, respecting the content of this guide, including but not limited to implied warranties for the guide's quality, performance, merchantability, or fitness for any particular purpose. Neither the authors, nor itSM Solutions LLC, its dealers or distributors shall be liable with respect to any liability, loss or damage caused or alleged to have been caused directly or indirectly by the contents of this guide.

Executive Summary

A key requirement for any IT service delivery organization, public or private, is to ensure the infrastructure has adequate capacity to meet the evolving demands of the business.

Many IT executives think "upgrade" with then hear "capacity" -- relating "a capacity plan" to "a spending plan." However, using the methods described in this paper, you can not only identify, justify and meet business needs for capacity, but also reduce costs -- without making new investments in people or products.

Using the methods described in this paper, you can not only identify, justify and meet business needs for capacity, but also reduce costs -- without making new investments in people or products.

To manage capacity effectively requires a model to understand business consumption of IT Services as well as all the components that make up an IT Service. This paper introduces *Peak Hour Load* as one metric for this purpose.

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

The ability to predict future business demand is critical to effective capacity planning. This paper introduces the concept of *Projected Saturation* as a means to understand future requirements and plan accordingly for cost effective capacity planning.

Capacity planning is not always about spending money; and in fact, this paper introduces several methods and techniques that can reduce costs while maintaining customer satisfaction.

In summary, this paper provides a practical and easy to understand guide to effective capacity planning in today's dynamic, customer-driven IT environment.

Effective Capacity Planning

According the best practices of the IT Information Library (ITIL), 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, 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; Business Capacity, Service Capacity and Resource Capacity. The three views of capacity 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 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.

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

Requirements for Effective Capacity Planning

IT Services consist of a collection of discrete components, called Configuration Items (CIs.) IT Services are composites of CIs, and are themselves CIs. Typically, CIs span several broad categories, for example:

Category	Example CIs	Example Measures
Storage	Disk, RAM, Tape	Total; used vs. free
Compute	CPU, Processors	Utilization; 0 to 100%
Transmission	LAN or WAN Link	Throughput; bits-per-second
IT Service	Record lookup, Send email	Latency; delay in seconds

Table 1. Example Configuration Items

The characteristic common to both IT Services and their resource CIs is that they all have a finite ability to perform useful work, and we can measure the work performed. We express the ability to perform work as "capacity." We express work currently performed as "load."

Effective IT Service capacity planning begins with examining the relationships between IT Services and their CIs. Understanding the capacity and load of CIs is the key to effective capacity planning for IT Services. An IT Services' capacity will never exceed the capacity of its most utilized CI; and the capacity and load of all its CIs limit the IT Services capacity.

Understanding Load

The question "What is the peak load?" is meaningless since it only describes the highest observed utilization. For example, a CPU might have a peak load of 100%. However, this 100% utilization might occur for just a moment and may not indicate the true utilization of the CPU.

The stock market is an excellent analogy for understanding the concept of load. Consider the common stock market index. Instead of predicable gains and losses, movements in the stock market appear almost random. At any given moment, the market is higher or lower than it was a moment before.

This movement makes it impossible to determine if the index is up or down without another reference point -- time frame. Worse yet, without a time reference there are several correct answers which directly contradict one another!

This holds true for IT Services as well. The direction and pattern of usage for IT Services and resource CIs are extremely difficult to discern without a reference. In fact, the direction of movement cannot be determined without specifying the period in consideration. Once you reference the period, it becomes easy to determine direction.

Consider the case of a transmission facility such as a *Wide Area Network* (WAN) link. If you sample the load at different times, you would have several different measurements, each of which is correct. If the sampling interval is short, the load can appear high if, for example, data is present. Longer sampling intervals will observe lower loads. To be useful, we need a defined sample interval and a defined sample size. We also need a series of samples showing the pattern of usage.

Peak-Load

As it relates to IT Services, you can obtain utilization data by measuring the load during the busiest period for the CI. We refer to this as the *peak-load*. Think of peak-load as a measure of the worst-case or highest utilization of a CI.

In business, people work according to their own schedules, at all hours of the day and night. This makes peak load critical. The question "What is the peak load?" is meaningless since it only describes the highest observed utilization. For example, a CPU might have a peak load of 100%. However, this 100% utilization might occur for just a moment and may not indicate the true utilization of the CPU.

To understand true utilization of a CI we have to take into account its usage over time. Consider the previous question rephrased as "What is the peak hourly load?" If we add a time element to peak load, we can obtain very useful information.

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

There is no one perfect metric, but Peak Hour Load is useful by itself, and when mapped over a larger period identifies those periods during the week, month, or year that require even higher capacity.

While there is no one perfect metric, PHL is useful by itself, and when mapped over a larger period 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.

Avoiding Future Problems by Projecting Saturation

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.

Utilization should be somewhere below total capacity. For example, a car may be able to exceed 150 miles per hours for short periods, but performs best at a speed of 65 miles per hour. Nearly every CI also 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.

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.

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.

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.

Consider the following example: Users complain about slowed system performance on Friday afternoon. Through charting of PHL and average load in your workload catalog, you discover that every Friday at 12:00pm network utilization skyrockets. Further investigation shows that the culprit is the weekly database backup. At this point, you can implement more capacity to meet the needs of both the backup and the Users, or you can try to shift the backup to a point in time where the impact is less.

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. During this step, a highly effective IT manager can produce significant savings for the organization while increasing system responsiveness.

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.

Planning for Success

IT Services use many CIs that have limited capacities to process, store, or transfer information. As CI utilization approaches maximum, these CIs become bottlenecks and impeded other services and applications. The result is negative impact on the business and low customer satisfaction.

IT Service capacity planning is an important job, but it can be difficult. Part of the difficulty is converting raw numbers into information. Good decisions only result from careful analysis of loads, capacities, usage, and time.

The first key to success is to realize that the measure of IT Service success is the Users ability to work. The definition of User requirements are the *Service Level Requirements* (SLRs) contained in *Service Level Agreements* (SLAs.) Capacity targets derive from SLRs.

The second key to success is to realize that an IT Service contains many CIs each of which has finite capacities. Together, these CIs make up an IT Service, and the latency of the service experienced by a User is a direct result of the underlying CIs.

The final key to success is to create a workload catalog CIs, charting *Peak Hour Load* (PHL), average load and maximum capacity. Charting the peaks and valleys of utilization over time illuminates your liabilities as well as your opportunities.

Summary

Settle on Peak Hour Load. PHL is the best single metric to settle upon. PHL accurately shows business consumption, and translates easily to any type of CI (e.g., CPU, RAM, Transmission, etc.) PHL also shows daily, weekly, monthly, quarterly, and annual requirements. Determine PHL for all CIs of your most critical or problematic IT Services first.

Create a Workload Catalog. Charting the PHL, average load and maximum capacity for CIs over time shows how the business consumes service. This data lets you understand and either support or modify these requirements.

Target the Projected Saturation Date. Once you collect enough PHL measurements, you can determine the projected saturation date for an IT Service or any CI. Use projected saturation to drive acquisition. Use your spreadsheets trending capabilities to show when in the future you will need to have more capacity.

Do not forget Service Cycle Time. Compare the projected saturation date to the service-cycle time (that is, how long it takes to perform the upgrade) for the CI. Take remedial action for any CI whose projected-saturation is sooner than the service-cycle-time.

Know your “Balance Factor.” Compare the average load to the peak hour load. If a CI has a high peak-hour load relative to average load then it may be a good candidate for demand balancing. Remember that demand management and balancing normally occurs through changing work habits or procedures, like staggering work shifts or rescheduling batch jobs.

Look for Opportunities to Downsize and Redeploy. After stabilizing your capacity requirements, you are ready to identify under utilized CIs. Before any purchase, make sure you do not already have “hidden capacity” already available elsewhere. Often you can eliminate, downsize, or redeploy under utilized CIs at significant cost savings.

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.

About itSM Solutions LLC

After becoming one of the leading IT Service Management providers in North America, we decided to focus exclusively on IT Service Management educational services. Drawing extensively on our own experience in the IT world and the best practices in IT Service Management as documented in the IT Infrastructure Library, we assist IT organizations of all sizes in achieving operational excellence through the adoption of a customer-focused, process-oriented, cost-effective approach to IT. Among our educational offerings is the complete suite of ITSM certification and awareness courses. itSM Solutions is accredited by ISEB and EXIN for the delivery of the ITSM Foundation Practitioner and Service Manager. We focus in IT Service Management education. Our goal is to provide you educational services that will guide you on your ITSM journey along with the knowledge base you need to select the right consultant should you need help along the way.

About the Author

Hank Marquis is Chief Technology Officer at itSM Solutions LLC. Hank has 26 years of experience in IT. He was an early proponent of IT best practices and process such as ITIL. He has vendor and practitioner experience, most recently he was Chief Technology Officer at a service management software firm. His practitioner experience comes from managing service desk, incident, and problem resolution activities at large telephone companies; as a Sr. Systems Engineer in the banking and brokerage industry; and as a telecoms technician. His instructional design and adult education experience comes from his work as an advisor on the Baccalaureate panel for NYC Technical College, Rochester Institute of Technology and Northeastern University. He has authored many training programs, written several books, and published dozens of articles. He holds ITSM Service Manager (Masters) Certification with Distinction in Service Delivery and is an ISEB accredited Course Director.