TEAMQUEST

IT OPTIMIZATION INSIGHTS

Getting Started:
A Manager's Guide to Implementing Capacity Management

White Paper

BY RON POTTER

TeamQuest specializes in IT Service Optimization

TEAMQUEST

Getting Started:
A Manager's Guide to Implementing Capacity Management

You have just been tasked by senior management to develop Capacity Management capabilities for your organization. You're very familiar with managing day-to-day IT infrastructure activities, but when it comes to Capacity Management, you're one of the many professionals with little to no familiarity with the subject.

No matter how IT services are acquired, delivered, or consumed, good Capacity Management is essential for efficiency, reliability, and productivity. With cloud computing, for example, a capacity manager is concerned with:

- The costs that services will incur by virtue of the resources they use, and
- Ensuring that sufficient capacity is reserved to cover potential demand.

And in most IT departments, it is necessary to be able to compare the costs and advantages of in-house solutions vs. public cloud-based solutions. In a DevOps environment, capacity managers have the opportunity to affect the architecture of new applications, and the effects that changes will have on capacity must be gauged

quickly in order to facilitate rapid development cycles. In addition, bimodal IT demands quick comparisons of apples and oranges in order to allow fast decisions about how to pursue capabilities and capacity.

Regardless of the mode of IT that is currently in use, this guide is designed to help those managers who are now looking to assemble a Capacity Management organization and put in place the processes and best practices to make it successful.

What Is Capacity Management?

Capacity Management is a set of work processes associated with the provisioning and management of IT infrastructure resources, such as servers, printers, and telecommunications devices, used to support business processes in a cost-effective manner. Capacity Management work processes include monitoring, reporting, tuning, planning, and predictive modeling.

Extensive, thorough definitions of Capacity Management can be found in plenty of available online resources, including IT Infrastructure Library (ITIL). The effective practice of Capacity Management is crucial to the effciency and effectiveness of any IT organization, and as IT becomes more directly tied to overall business outcomes, its benefits are increasingly being felt across companies. Good Capacity Management results in:

- **Reduced costs:** By allocating your IT resources in the most efficient way possible, Capacity Management reduces maintenance costs and empowers your business to provide better service for less investment.
- **Improved business productivity:** Proper management of IT capacity results in fewer outages, which means less time spent on responding to problems and more time for high-level, value-adding tasks. Work gets done faster and the overall business becomes more agile.
- **Better customer satisfaction:** Downtime or slowed response times will hurt any organization's reputation for consistent service. Implementing a proper Capacity Management strategy will reduce service outages over time, improving customers' experience of your web services and bolstering your reputation.

- **Career advancement:** Improving the cost efficiency and performance of IT operations will increase your value to business leaders and executives.

ITIL version 3 views Capacity Management as comprising three main sub-processes: Component Capacity Management (referred to in ITIL version 2 as Resource Capacity Management), Service Capacity Management, and Business Capacity Management.

Component Capacity Management

Component Capacity Management looks at the use and performance of individual infrastructure components, including servers, telecommunication lines, printers, scanners, routers, tablets, smartphones, and PCs. This is the lowest level approach to Capacity Management — the work only addresses each individual unit's performance and capacity positions. Although very effective in managing individual units, this approach has shortcomings because of its narrow view.

Adding capacity on one infrastructure unit can have unforeseen implications on others. For example, upgrading a telecommunications line can permit more transactions to be performed during a set period of time, potentially flooding a server with unexpected transaction volumes. This possibility can result in poor performing transactions or even worse, cause the server to shut down unexpectedly due to overwork.

Service Capacity Management

Service Capacity Management takes Component Capacity Management to the next level by categorizing individual component information according to the IT services they support. Printing, email, telephone, and online transaction processing are just a few examples of IT services. This approach looks at capacity positions across the IT infrastructure components supporting and sustaining IT services. For example, an online transaction processing service could have server, telecommunications, desktop, and printing components.

Managing from a service-centric perspective ensures that upgrades to any one component do not adversely impact the others. However, the shortcoming of this approach is that while those in IT understand the work and its value to the organization, the business often cannot relate to the technical terminology. As IT becomes more integral to business, executives and company leaders will become more familiar with its most important concepts — still, there are usually communication breakdowns between technical and business staff.

Business Capacity Management

This is the highest level approach to Capacity Management. Instead of aggregating usage and performance into IT services, Business Capacity Management organizes them into business processes. This approach looks at capacity positions across modes of IT that support and sustain business processes. Sales order entry, claims adjudication, customer relationship management, and general ledger are just a few examples of business processes.

For example, a Sales Order Entry system could have telephone, server, telecommunications, database, desktop, and printing components, just to name a few, or it might be implemented in the cloud using IaaS, PaaS, or SaaS. There is a lot of value in employing a business-oriented approach -- results are expressed in business terms, so all relevant parties can understand. Conversations are transformed from ones that cover expense to ones of business investment. Reaching this level should be any organization's long-term goal.

Managing Capacity Management

Contrary to popular belief, you don't need a doctorate in mathematics to manage a Capacity Management team. Understanding queuing theory is a plus, but also not required. You do need to understand the work that must be performed. Your challenges will be to develop credibility in your work and demonstrate to the business

how important their participation in the process will be to business success. You will need to understand business processes, IT processes, and how they interact. However, your job is primarily centered around relationship-building and the facilitation of good communication throughout the organization.

You are the gatherer of business planning information to support capacity planning efforts. You are also the bearer of news resulting from those efforts — sometimes it's good and sometimes it's bad. That being said, in medium and large organizations you can have more financial impact on the organization than in any other IT department, and your ability to relate IT services to business processes will be invaluable to the organization.

Getting Started

Now that you have a high level of understanding of the work, we need the organizational resources to perform the work. As in any business endeavor, people, processes, data, and tools are needed to ensure success. The other unspoken needs are senior and business management buy-in and ongoing support. Since Capacity management has so many touch points across the entire organization, lack of management support will result in limited value to the organization or outright failure.

Business and IT management support is crucial to the success of any effort to implement a Capacity Management practice. That means more than just getting management's blessing to proceed with the work, or encouraging cooperation where there might otherwise be none. Business strategy, planning information, and business process flows are also needed.

Also necessary are strong relationships and frequent dialogues to better understand business needs so that the right IT resources are in the right place when the business needs them. Building relationships also ensures you get involved early-on in a project, permitting you time to review and plan instead of reacting to a project at the last minute.

People

Staffing will depend upon the size of the IT organization. In smaller sites, a few generalists are in order. Larger sites should rely on specialization, as more detailed work can be completed in shorter periods of time. Senior people are preferred because things rarely act in the same manner as they do in the college laboratory.

I once hired a mathematician as a capacity planner. He had a great understanding of the mathematics behind the work and created wonderful statistical documents. The problem was that he didn't have any experience with the equipment or software, so he couldn't judge the reasonableness of his work. Remember that we previously mentioned that this work is part science, part experience, and part intuition. You need all the parts to be successful.

On the whole, I have found that system programmers and system administrators make the best capacity planners, especially those with considerable operational experience. It doesn't seem to make a difference if they have a college degree. Most are intelligent and can quickly absorb the mathematics and processes. The biggest challenge will be to get them out of the details. Since capacity planners deal with averages, working at too fine a level of detail can bog down analyses and have little or no improvement on the precision of the work.

Training is an important part of keeping staff current and effective. Technology changes at a frightful rate and your staff needs to keep abreast of it. Newer technology may provide efficiencies that can substantially improve operations and lower costs. If you and your people do not continue to improve your knowledge, you cannot hope to keep your business ahead of its competitors.

Processes

Both traditional ITIL and DevOps approaches to IT use processes to deliver meaningful, repeatable, and reliable results. Processes ensure you come up with the same answer every time you analyze the same pieces of data. Meaningful, repeatable, and reliable results generate credibility — something Capacity Management needs in order to survive.

Processes describe the work that needs to be accomplished. They identify the inputs and outputs; the inputs being the information coming into the process that needs to be acted upon and the outputs being the deliverables that result from the work. Processes

then identify the recipients of the results of the work performed. Well-written processes ensure the people executing them completely understand their roles and responsibilities. They also permit people to understand the importance of their work and how their work supports business success.

Example ITIL and DevOps processes can be obtained from a number of published sources. Many consultancies can also provide samples. You may be able to find sample processes on the Internet. In all cases, be prepared to tailor them to your needs. Few organizations exactly match the sample scenarios.

ITIL v3 was designed to account for ascendant philosophies like Agile Methodology and DevOps — its discussion of "Continual Service Improvement" encourages teams to ceaselessly align ITIL processes with emerging business needs. However, some DevOps/ Agile proponents believe that ITIL's reliance on processes still puts too much emphasis on planning and not enough on experimentation.

While many people feel DevOps is perfectly compatible with ITIL processes, those who wish to focus their Capacity Management efforts on cutting down on waste in the development process should ensure they're monitoring both the performance and the usage of their applications. The rapid experimentation and testing that's integral to DevOps requires a proper understanding of Capacity Management, and therefore, should grade iterations of new products and applications based on their efficient use of IT resources.

Data

The type and amount of data you need to collect depends upon the mission of your Capacity Management organization. The most commonly collected and stored types of data involve performance, transactions, and usage. For servers, you will collect data such as:

- Processor busy
- Disk read/writes
- Memory statistics
- Job/process execution statistics
- Database related statistics

Networking capacity planners commonly look at bandwidths, transmission delay, router, firewall, and packet statistics. The types of data you'll collect will depend on your Capacity Management goals. In most organizations, too much data is collected during the initial phases, then reduced and refined as the Capacity Management organization matures.

Data should be captured in the same intervals across all infrastructure components, cloud-based or not, for which you are responsible. It is difficult to analyze the interactions between a transaction server and a database server if one is collecting at one-minute intervals and the other at five-minute intervals. Spikes seen at one-minute intervals will be unnoticeable on the five-minute intervals, making work correlation difficult, if not impossible. You need clocks to be synchronized as well.

Data archiving is very dependent on the organization. Retention times depend on the amount of change in the environment and business cycles. For example, if applications and business processes change frequently, storing data for long periods of time may be counter-productive, as last year's data will have little or no relevance to current operations. Business cycles also determine retention periods. A business cycle with one busy period that lasts one or two months per year may wish to keep between three and four years of data. Another organization with monthly peaks may choose to save only 13-18 months of data.

Tools

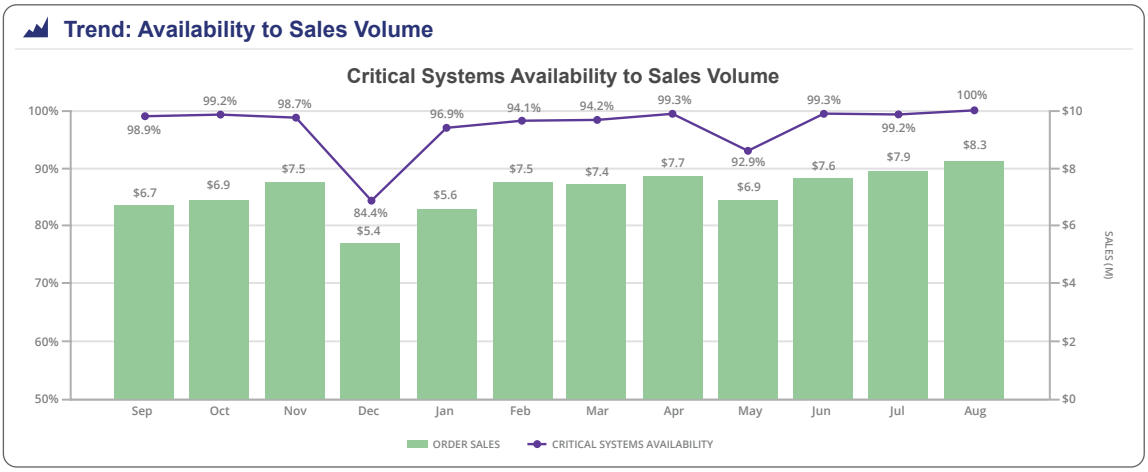
Tools allow you to trade cost for productivity. Yes, you can do the work with spreadsheets, but it will be difficult, and your people will spend more time gathering and processing data than they will analyzing it. Modern tools automatically collect, organize, and archive data, as well as provide analysis and modeling capabilities. They free up your staff's time from mundane, repetitive tasks to work on more important analysis, planning, and prediction work.

I can't tell you the right answer for your organization. You know your cost and business drivers better than I do. However, I have formed a number of Capacity Management teams over my career, and my biggest successes have come from using automated analytical tools to manage the data and free my people's time to do the analysis necessary to improve service and reduce costs.

Common tasks performed by Capacity Management tools are data collection and archival, analysis, modeling, and reporting. Data collection and archival frees up a lot of time for staff — save for a few tweaks now and then, the process basically runs itself after the initial setup. Analysis tools permit you to correlate events and drill down into details to find causes of performance problems or to identify performance improvement opportunities. Modeling tools enable you to look at very accurate "what-if" scenarios that you can use to predict future requirements, permitting you to more effectively use and manage IT resources. Reporting tools can help you set up automatically generated reports. You develop the report once, set up a publishing schedule and it does the work. You only get involved with exceptions or changes.

Business value dashboards will also be crucial, especially when it comes to communicating reduced costs and justifying IT budget to company leadership. Being able to communicate effectively with management is key, but no matter how effective you are at shifting between IT and business vocabulary, nothing beats a visual element. Spend time coordinating with your staff and with business leaders to

ensure that IT and business goals are aligned with one another and are all being properly measured. Learn to use dashboards to show the value of IT to managers by demonstrating instances in which IT intervened to prevent service outages or over-provisioning.



Above: the dip in critical systems availability in December represents a point at which an IT team was able to quickly intervene and reallocate resources to avoid an impact on overall revenue.

The Work of Capacity Management

Good Capacity Management requires performance management and capacity planning together with business-aligned financial analytics. Those higher-level functions are enabled by necessary foundational tasks.

Monitoring

Monitoring permits you to see how a particular infrastructure component or application is performing at that moment in time. Monitoring can help you diagnose a customer complaint of poor performance or, if used proactively, help you find performance problems before they negatively impact your customers' experience.

As more and more organizations come to rely on these DevOps and Agile methodologies, the goal of Capacity Management has become the real-time monitoring of demand and capacity. Without strong monitoring tools, development teams are essentially "flying blind" as they roll out new application changes — no analysis means no clue as to how those changes will affect the utilization of CPU, memory, and disc I/O. While some bugs can be caught by implicit testing, waiting for service demand to spike or for a customer complaint limits the speed at which problems can be resolved.

But these monitoring capabilities must go beyond performance: an IT department with only a performance monitoring tool in place will get blindsided by over-provisioning costs. To make effective server purchasing decisions and determine where their resource provisions are likely to fail, companies also need usage monitoring and management. When automated, those processes will continually

alert project managers to weaknesses within their applications and infrastructure as they relate to both performance and cost.

The objectives of companies hoping to implement Capacity Management processes into a DevOps/Agile environment should be able to effectively manage both usage and performance and to translate that into cost savings for business leaders. Both of these objectives are benefitted by the use of visualization, especially through a business value dashboard.

Analysis & Reporting

Data by itself has no intrinsic value to the business. To be valuable, it needs to be analyzed, correlated, and converted into information with which leaders can take actionable steps. Analysis is the process by which we turn our data into information. Telecommunications bandwidth, server memory, data channel busy, and server utilization are just a few of the resource types that are frequently analyzed.

Within IT, there are two commonly used analysis types: reactive and proactive. Reactive analysis is used when there are issues in service delivery. The goal is to find the cause of the issue, determine ways to mitigate it, and do the work necessary to restore service to previous levels. Proactive analysis employs some of the same processes, but its goal is to find potential or impending issues so they can be mitigated before impacting customers and business staff. Reactive analysis will always be with us, but the more proactive work that is accomplished, the fewer reactive events will occur.

Performance Management

Better-performing applications use fewer resources. Tuning applications to perform better improves service while reducing costs — a truly win-win situation. DevOps methodologies can help to ensure performance by allowing experimentation and testing of performance as part of each iteration.

Licensed software and hardware infrastructure components are delivered with configuration parameters directed for the "average" data center. If these are not tweaked to reflect your actual operating conditions, you may not get the most from your investment. The tuning process ensures the proper configuration parameters are in place.

To that effect, it's good to have a flexible infrastructure environment that can be easily adjusted to ensure costs are kept low when service spikes. Elastic infrastructure from the cloud is helpful here, but as noted earlier, IT and business units should collaborate closely to identify the infrastructure pieces they need to ensure that data is stored and transferred efficiently.

Capacity Planning

This is the planning function of Capacity Management. Capacity Planning takes business planning information, translates it into IT infrastructure resource needs, and predicts what resources need to be put in place at a particular time to satisfy the business needs.

Capacity Planning also ensures that the costs associated with the service are within the business' ability to pay. If they are not, Capacity Planners looks at alternatives that best fit the need. Reducing service levels could be an alternative where support costs are too expensive.

Part of what Capacity Planners do is look at historical growth. They determine periods of time representative of normal business operations and analyze usage and performance information during those time periods. Using this data they develop a point-in-time "picture" of infrastructure.

This is called a “Baseline.” All growth projections, whether positive or negative, are applied to the Baseline to develop predictions of future infrastructure needs. For example, a call center application may be evaluated from 10am to noon on Mondays and Tuesdays because those call volumes are representative of the levels that the business wants to provide good service.

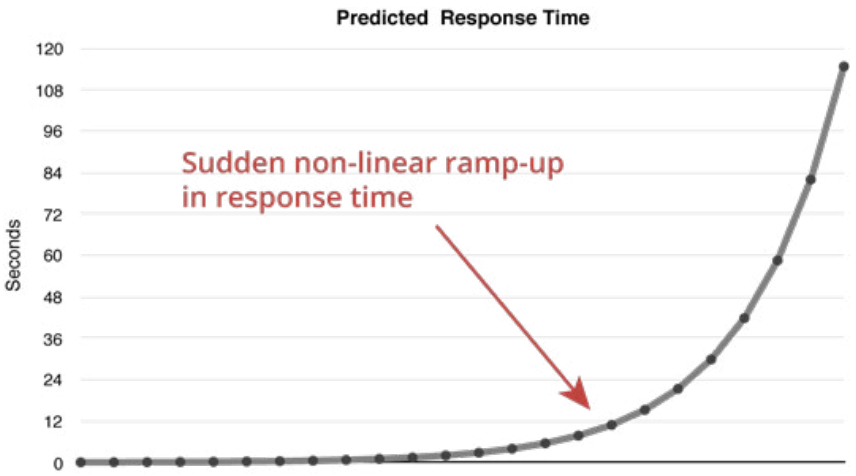
Capacity Planning work looks at all aspects of the IT infrastructure — addressing more than hardware and software. Capacity Planners need to understand impacts to facilities, both IT and business. They also need to understand the capacity of the staff to perform the work to add or reduce infrastructure components. Without those considerations, projects can fail because there are not enough people to do the work or there is insufficient floor space to house the new equipment. Mistakes such as these can be costly in a fast-moving business environment.

Capacity Planning time frames should address immediate needs and also look several years out. Capacity Planning may also be performed on a sprint or development iteration timeframe, taking the predicted effects of each new iteration on target systems into consideration. One view of Capacity Planning should be aligned with the budget cycle. The business and senior IT leaders are already managing to the budget cycle so understanding of the implications of the planning work is more easily understood. A longer-term view provides additional input to the decision making process. If long-term direction varies from immediate needs, a different approach may be taken than if short- and long-term directions are similar.

Predictive Modeling/Simulation

Many organizations use trending to predict future IT infrastructure component performance. Trending involves graphing capacity positions for a period in the past, drawing a line that depicts the growth trend, then projecting the line into the future based upon business growth plans.

The problem with this approach is that computer systems do not perform in a linear manner. As a threshold capacity of a device is approached, requests for that resource start to queue (wait for service). In a very short time, transaction queuing causes performance to degrade at a rapid rate as depicted in the graph below.



In order to more accurately predict computer system performance, Analytical Modeling or Event Simulation techniques must be employed. Event Simulation techniques are the most accurate, but are also the most time consuming. Event Simulation does what the name implies; it uses computer software to simulate the running of all the work on the infrastructure components. In order to do this work, modelers need to identify each individual transaction, process and/or network packet and enter it into the model. Depending on the size of the model, it could take weeks or months to complete the work.

On the other hand, analytical modeling uses mathematical equations to predict future performance. It can be very accurate as well; results usually fall within 5% of actual conditions. Because the software has mathematical background, you can generate many more models in the same period of time. In fact, analytical modeling is so efficient that automated processes can run models on large numbers of systems or services in order to predict where and when future capacity issues will occur and what costs future workloads might incur. With analytic

modeling, it is also much easier to run a variety of what-if scenarios. For these reasons, most people who model capacity employ analytical modeling software.

Financials

Capacity Management needs to be concerned with financials. You can have the best provisioning plan in the world but it is of little value if the business cannot afford to implement it. Money is the language of business, so all work needs to be done with that aspect in mind. There are several aspects of financials that you need to consider:

- **Budget:** This is the financial process associated with planning expenditures and capital improvements over the planning period — whether yearly, quarterly, or monthly, this planning process must be aligned with the organization's business cycle.
- **Total Cost of Ownership (TCO):** This is a methodology to determine the actual costs of implementing and maintaining a specific IT project over a specific period of time, usually five years. The process looks at all aspects of costs, such as usage charges for the cloud and acquisition, installation, ongoing maintenance, and facilities costs for in-house IT. Organizational overhead and staffing costs should also be considered. With short- and long-term project costs better understood, management can make more informed decisions on whether to proceed with a project.
- **Capex vs. Opex:** An important consideration when evaluating the financial requirements for providing capacity is whether or not the expense will be incurred as a capital expense (Capex) or an operational expense (Opex). Typical in-house computing systems are accounted for as a capital expense, where an upfront investment in infrastructure is amortized over its lifetime. Cloud services are typically thought of as Opex, though many cloud services allow businesses to reserve capacity in advance, which would then be accounted for as a capital expense. When infrastructure is provided via Capex, users might

still have their budgets sapped by usage (see "Chargeback," below), but a capital outlay is still necessary in order to get the ball rolling. The decision to make the investment will depend on the competitive advantages of having the infrastructure in-house (or reserved in the cloud) and how the total cost of ownership over time compares with providing the same services using infrastructure that is provided on a pay-as-you-go basis, i.e. Opex.

- **Chargeback:** Business has already learned that anything free is likely to be abused. Almost every organization has controls on office supplies and travel, making each manager responsible for what they use. IT resources are no different. Chargeback systems permit IT resource usage to be attributed to a particular user or cost center. By asking managers to plan and account for usage, more effective use of computer systems has been experienced.
- **Expressing the value of your work:** In order to communicate the value of your team's work, you need to express results in financial terms. If you don't, management will not understand your value. For example, will business better understand the value of what you do if you say that you removed 200 servers from the infrastructure through consolidation or that you saved the organization $7 million per year by combining servers? Again, this communication would be greatly aided by the use of an effective business value dashboard that will translate dense IT data into metrics that business leaders can not only understand, but find compelling.

Easing Into It

Now that we know the work that needs to be done and what is needed to do it, where do we start? Again, each organization has its own set of priorities and goals so where to start is dependent on them. Here are a few areas where successful organizations have started.

- **Mission critical resources:** These transactions and jobs are critical to business success, so any improvements in service or reduction in costs gets immediate attention from senior IT and business leaders. Starting with these can more quickly develop credibility in your work; making it much easier to gain cooperation in other areas.
- **Low hanging fruit:** Your technical staff probably already knows which applications and services are poor-performing. In many cases, they already know the solution; they just haven't been able to garner support for corrective activities. Choosing to start here allows you to address performance problems more rapidly. Being able to quickly resolve some nagging problems with the business will go a long way towards developing credibility within your organization.
- **More frequently used transactions or jobs:** Savings from small, frequently used transactions can have bigger impact than a single large saving on a seldom used job. I have seen simple performance tweaking of a heavily used transaction drop a large server's utilization from 32 processors to 12. We were able to satisfy several years' growth on that server plus provision several new applications, saving considerable hardware and software expense.

What's Next?

Keeping focus on the start-up tasks – learning the basics, getting your processes in place and maturing them – is important; however you need to keep an eye to the future. Once your Capacity Management organization starts to stabilize, there are several things to consider to take it to the next level...

I can't overemphasize the importance of building better business relationships. Better relationships means business leaders come to you in the planning stages of a new project. It means they come to you to alert you to the new marketing campaign which, if successful, could over-run your servers. Relationships also mean that your business partners understand the value of your work to the success of the business, making it much easier to sell the needed work.

Finally, let's talk a little about Sales and Marketing. You may be doing a great job of improving performance and reducing costs, but if you don't tell anyone, all that great work may go for naught. To paraphrase a Forrester analyst, "Shout your successes to the rafters."

In addition, start to put yourself in position to change conversations from that of expense to that of business investment. Showing graphs of server growth or Total Cost of Ownership spreadsheets visualize cost, thereby making it easier to understand. If you can express results in terms of IT cost per increment of business work, such as an order or workflow item, the business can see the value to the investments, especially if the unit costs decline as a result of economies of scale.

In Closing

Good luck with your work to establish a Capacity Management organization. I think you will find it as rewarding as I have. I hope this guide has helped you understand the work from a high level and reduced the uncertainty of beginning.

ABOUT THE AUTHOR

Ron Potter is the Best Practices manager for TeamQuest Corporation. Ron's background includes more than 20 years in the IT industry, spearheading a successful ITIL implementation with a Fortune 500 insurance company, and discussing ITIL topics as a presenter at several conferences and trade shows.

TEAMQUEST

connect with TeamQuest:

WORLDWIDE HEADQUARTERS

UNITED STATES

Executive Offices
TeamQuest Corporation
430 N 1st Ave
4th floor
Minneapolis, MN 55401

Development Lab
TeamQuest Corporation
One TeamQuest Way
Clear Lake, Iowa USA 50428

OTHER LOCATIONS

SWEDEN
GERMANY
UNITED KINGDOM
CANADA
MEXICO
HONG KONG

With resellers in many additional countries.

CONTACT US

info@teamquest.com
teamquest.com/about-us/contact-us/