How to Do Capacity Planning

White Paper

BY JON HILL
How to Do Capacity Planning

This paper explores both scenario-based capacity planning, also known as one-at-a-time or “what-if” capacity planning, and newer, automated predictive analytics solutions designed to meet the needs of today’s complex and rapidly-changing virtualized and cloud environments.
The Evolution of Capacity Planning

IT capacity planning is a process for determining the IT infrastructure that will be required to meet future workload demands. It’s an essential discipline, but for most companies, it’s getting increasingly hard to find people who are capable of doing it. Many of the gurus with “Capacity Planner” as part of their professional title have retired without being replaced. There are fewer and fewer people with the detailed technical knowledge needed to make precise predictions, or even with the experience necessary to make “gut” predictions regarding future IT infrastructure needs.

At the same time that capacity planning expertise has been disappearing, the task of capacity planning has been becoming both more important and more impossible. Capacity planning is more important because organizations are increasingly dependent on IT for success. Modern customers want on-demand services, meaning that companies can’t afford bottlenecks and inefficiency that last even a single minute. Capacity planning is becoming more impossible because IT today is more dynamic and complex. Applications run in rapidly-changing, multi-layered, virtualized, and cloud-based environments – even if there were enough capacity planning experts to go around, there’s rarely enough time anymore for them to do their jobs sufficiently.

Fortunately, the technology exists to automate much of capacity planning, doing even more work than a traditional “guru” would be capable of. It is now possible to continuously apply sophisticated predictive analytics to optimize applications running across thousands of virtualized systems. Automated predictive analytics can help you gain a competitive advantage by maintaining a balance between cost and performance. That means you can efficiently meet service levels and availability requirements, without undue risk.
Benefits of Capacity Planning

- Get a competitive edge
- Avoid waste from over-provisioning
- Save time
- Minimize costs
- Avoid risk
- Prevent bottlenecks
- Increase business productivity

Why Capacity Plan?

Capacity planning is all about optimization – it's where you figure out how to maximize business benefits of IT without overspending, balancing business productivity with IT costs. Any CIO or IT manager with a limited time, money, or personnel budget should be motivated to use capacity planning. Many organizations buy into virtualization and cloud vendor promises of “elasticity” to obviate the need for proper capacity management, but that’s naive. Capacity planning is an important tool for gaining a competitive advantage over companies like these.

Here are just a few of the uses for capacity planning:

- Predicting when your infrastructure will no longer be able to meet service levels
- Determining which infrastructure elements will become future performance bottlenecks
- Comparing private cloud vs. public cloud costs
- Finding and re-purposing under-utilized IT infrastructure
- Determining which workloads can efficiently live together on the same hardware
- Accurately predicting performance for different VM densities
- Finding redline performance limits without costly testing on large, production-like environments
- Preparing for future business workloads, whether forecasted or unexpected
- Predicting the most cost-effective production configuration before rolling out a new application
- Determining optimum virtual and physical machine configurations
- Migrating workloads to new IT infrastructure
- Making sure your disaster recovery solution will have adequate capacity
- Ensuring successful server consolidation

Capacity Planning Methods

The techniques and processes for capacity planning range from doing nothing to actually testing with full-scale workloads applied to production-sized IT infrastructure. However, both of these extremes are rarely – if ever – appropriate. Many cloud and virtualization vendors would like to sell you on the idea that their platform eliminates the need for capacity planning, but that’s marketing, not reality. And testing full-scale workloads on production-sized infrastructure is insanely expensive – spending that much on a single testing session rarely makes practical sense.
Capacity planning is part of a complex discipline called Capacity Management. Good capacity Management is essential for providing optimized IT services so it is crucial that executives and managers understand how well their companies are doing it. That’s why TeamQuest has developed an IT Service Optimization Maturity Model, making it easy for businesses to know where they stand.

The capacity planning techniques an organization uses often mirror that organization’s IT Service Optimization maturity level, with Level 1, “Chaotic” organizations reacting only after service and availability levels have been breached. These organizations handle capacity as a firefighting exercise, dealing with service level breaches after they have already disrupted business. Organizations at Level 5 “Value” in capacity management maturity tend to be more prudent, carefully balancing cost, risk, and business productivity with the help of predictive technologies.

Between Level 1 and Level 5 you’ll find capacity managers keeping historical data regarding performance and using it to plot trends that help predict future performance. Trending is better than nothing, but it is far from optimal for most capacity planning situations. That’s because trending assumes that workloads are increasing at a steady rate, or that they’re changing according to a statistical formula that is often not accurate. Trending falls down when adding a new workload, for example, or consolidating servers. And it ignores the fact that performance scales non-linearly. To perform accurate capacity planning, you need something better.
But first, let’s explore why performance doesn’t scale linearly. The simple answer is queuing. Once a CPU, controller, or other device has more work to do than it can perform, the systems managing that work will put new requests on a waiting list or queue. When this happens, response time begins to increase exponentially.

Planners who use simple trending avoid dealing with the nonlinear nature of capacity planning by always keeping utilization levels far from 100%, which helps them avoid the knee in the performance curve. While this tactic sometimes works, it almost always amounts to over-provisioning – these planners are wasting resources because their techniques lack precision. This all might sound lazy, but lazy is the wrong word: with the current technology available, you can actually get much more accurate predictions without expending any special effort.

Two techniques for capacity planning that take queuing behavior into account are discrete event simulation and analytic queuing network.
solvers. Both techniques are often referred to as “performance modeling” by experts.

Event simulation is extremely flexible for modeling very diverse infrastructure configurations, but each simulation can be time-consuming to set up and run. An analytic queuing network solver, on the other hand, is easy to set up, highly accurate, and incredibly time-efficient. It makes sense, therefore, to use an analytic queuing network solver whenever possible, falling back on simulation in those rare cases where the extra flexibility is necessary.

Want to know where your company stands on TeamQuest’s IT Service Optimization Maturity Model? Take our free Maturity Assessment – answer 20 yes or no questions, get your results back immediately, and find out how you can improve your Capacity Management processes and start providing better, more proactive, and more valuable IT service today.
Scenario-Based Capacity Planning vs. Continuous Automated Predictive Analytics

Traditionally, capacity planning has always been handled one performance model at a time, but with the right software (such as TeamQuest Predictor), the process can be automated, making it possible to run thousands of models at a time. That said, there is still a need for both one-at-time modeling and automated modeling or automated predictive analytics.

One-at-time modeling is useful for answering “what-if” capacity planning questions for a limited number of infrastructure elements or configuration scenarios, which is why it’s widely referred to as scenario-based planning. What-if analysis can answer, for example:

- What will the response time amount to if the incoming workload for this application doubles?
- How will all of my applications perform after this new application is added to production?
- How will service levels be affected if servers are consolidated?
- How many VMs running this workload can safely run on each physical server?
- What is the cheapest infrastructure configuration that will still meet service levels?
- How will changes to I/O devices, network bandwidth, and the size and number of CPUs affect daily operations?

This kind of capacity planning also allows for experimentation with different scenario-based situations or models. By running predictive...
analytics against extreme scenarios, like a sudden doubling in workload or the loss of a critical server, capacity planners can find the optimal configuration for minimizing risk without undue cost.

Automated predictive analytics, on the other hand, are what capacity planners will find most useful in their day-to-day monitoring of applications and systems. Automated predictive analytics are useful for keeping a continuous eye on the predicted performance of large numbers of infrastructure elements serving applications and business services. For example, you'll be able to confidently answer each of the following questions:

- Which of my applications are in danger of failing to meet service levels within the next six months?
- Where will my future bottlenecks be?
- What specific component will be responsible for future bottlenecks in my complex, multi-tiered environment?
- How long will it be before my current configurations will fail to meet service levels?
- What will response time be for application xyz next month?
An Example of Scenario-Based Analysis

Looking at the Oracle workload, we can see in the chart below which infrastructure elements will be responsible for the lion's share of the response time.

Components of Response for Oracle

You can see clearly in the chart above that there is a bottleneck at on the Saturn system's CPU beginning to form at a point three months in the future. This bottleneck is predicted to get much worse over the next few months. A typical follow-up to a model like this would be to predict the change in this response time graph if the CPU were upgraded. Below is an example.
It is clear in the graph above that CPU Queue Delay has been eliminated as a major contributor to response time and that CPU Service, the amount of wall clock time a CPU spends crunching away on a problem, has been reduced as well. The future problem can be avoided by the modeled CPU upgrade. It might even be possible to go with a less capable upgrade. That scenario could be modeled next, to determine which upgrade will provide the best cost/performance balance.
Examples of Automated Predictive Analytics

With the advent of TeamQuest Predictor, it is now possible to continuously predict performance by automatically solving performance models across your entire data center. You can get regular reports showing you which of your systems are likely to violate service levels in the future, as well as when and why this will occur.

Below is an example of just such a report. (You can zoom in with Acrobat or Preview to see the details.)
The report above uses TeamQuest Predictor’s AutoPredict to forecast and color-code future service status. This gives capacity planners advance warning of systems that are likely to exceed performance limits as defined by “TPI,” or the TeamQuest Performance Indicator. TPI is a single, easy-to-understand, but sophisticated metric that summarizes server, service, or application health, and it’s more accurate than other KPIs because it’s based on an analytic performance model. TeamQuest software can automatically and continuously calculate TPI across your entire infrastructure, making it possible for you to detect and correct performance problems before they occur.

*Every modern business should be available to customers 24/7, but in the finance industry, every second offline could mean breaking the trust of your customers, as well as some serious government regulations. Find out how Interactive Data Corporation uses TeamQuest to stay one step ahead of performance constraints to ensure its hosted financial applications are always running smoothly.*
This Predicted Server Health report was produced by the TeamQuest Surveyor and shows color-coded TPI predictions for an entire year, giving you plenty of time to head off potential bottlenecks or outages. You can even drill down to get a predicted components of response time analysis to reveal why server health is going to suffer in the future, so you'll know what to change in advance in order to head off the problem.

**Want help with capacity management, capacity planning, or need advice on how to improve your organization’s IT Service Optimization maturity? Contact us at our offices in the U.S., Mexico, Hong Kong, or one of our three European locations.**
Conclusion

The state of the art of capacity planning has evolved. Scenario-based, one-at-a-time analysis techniques have now been supplemented with highly-automated management software. As a result, the process no longer requires immense amounts of up-front capacity planning time or expertise. You can be informed of future problems well in advance of their occurrence, with sufficient detail to know how to avoid those problems entirely. With new, automated predictive analytical techniques, you get continuous capacity analysis, resulting in greater efficiency and decreased risk.

With automated predictive analytics, there is no longer a need to triage which services, applications, or servers you should model – you can analyze all of them automatically and with ease. Capacity monitoring is now possible, enabling continuous IT Service Optimization. You can avoid embarrassing bottlenecks and stop wasting resources on firefighting, and because it’s all automated; you don't need a capacity planning guru to make it all happen.

ABOUT THE AUTHOR

Jon Hill has been working with TeamQuest since its inception in 1991. He currently participates on the product management and marketing teams at TeamQuest, helping to keep the company in touch with industry, market, and competitive trends.
How to Do Capacity Planning

TeamQuest, the TeamQuest logo, VITYL and all other TeamQuest trademarks are trademarks owned by TeamQuest Corporation. All other trademarks listed or referenced herein are the property of their respective owners.

NO WARRANTIES OF ANY NATURE ARE EXTENDED BY THE DOCUMENT. The only warranties made, remedies given, and/or liability accepted by TeamQuest, if any, with respect to the products described in this document herein are set forth in a separate such license agreement. TeamQuest cannot accept any financial or other responsibility that may be the result of your use of the information in this document or software material, including direct, indirect, special, or consequential damages. You should ensure that the use of this information and/or software material complies with the laws, rules, and regulations of the jurisdictions with respect to which it is used. The information contained herein is subject to change without notice. Revisions may be issued to advise of such changes and/or additions.

U.S. Government Rights. All documents, product and related material provided to the U.S. Government are provided and delivered subject to the commercial license rights and restrictions described in the governing license agreement. All rights not expressly granted therein are reserved.