

# How to Do 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. This paper describes the fundamental steps for performing capacity planning. Real life examples are provided using TeamQuest® Performance Software.



## About the Author

Enterprise Performance Specialist Joe Rich has been active with performance analysis and capacity planning for many years. He is involved with technical support and field activities for open system performance and has been active in high performance computing and mid-range Open Systems platforms.



## 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.

## Three Steps for Capacity Planning

In this paper we will illustrate three basic steps for capacity planning:

1. Determine Service Level Requirements

The first step in the capacity planning process is to categorize the work done by systems and to quantify users' expectations for how that work gets done.

2. Analyze Current Capacity

Next, the current capacity of the system must be analyzed to determine how it is meeting the needs of the users.

3. Planning for the future

Finally, using forecasts of future business activity, future system requirements 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.

## Determine Service Level Requirements

We have organized this section as follows:

- a. The overall process of establishing service level requirements first demands an understanding of workloads. We will explain how you can view system performance in business terms rather than technical ones, using workloads.
- b. Next, we begin an example, showing workloads on a system running a back-end Oracle database.
- c. Before setting service levels, you need to determine what unit you will use to measure the incoming work.

- d. Finally, you establish service level requirements, the promised level that will be provided by the IT organization.

### Workloads Explained

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 defined and a definition of satisfactory service must be created.

A workload is a logical classification of work performed on a computer system. If you consider all the work performed on your systems as pie, a workload can be thought of as some piece of that pie. Workloads can be classified by a wide variety of criteria.

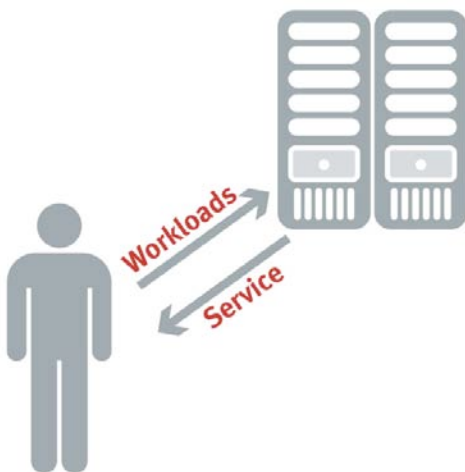
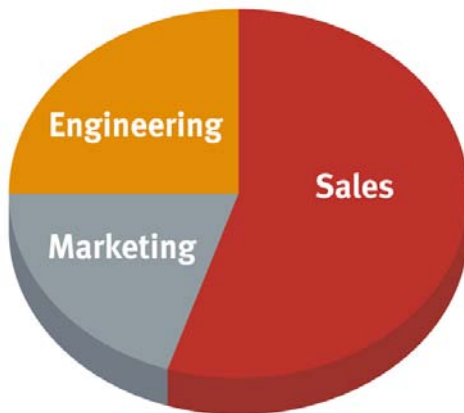


Figure 1  
Workloads and Service

<b>who</b>	is doing the work (particular user or department)
<b>what</b>	type of work is being done (order entry, financial reporting)
<b>how</b>	the work is being done (online inquiries, batch database backups)



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 accounts payable department on a consolidated server than it is to predict the overall increase in transactions for that server.

Figure 2  
Workloads by  
Department

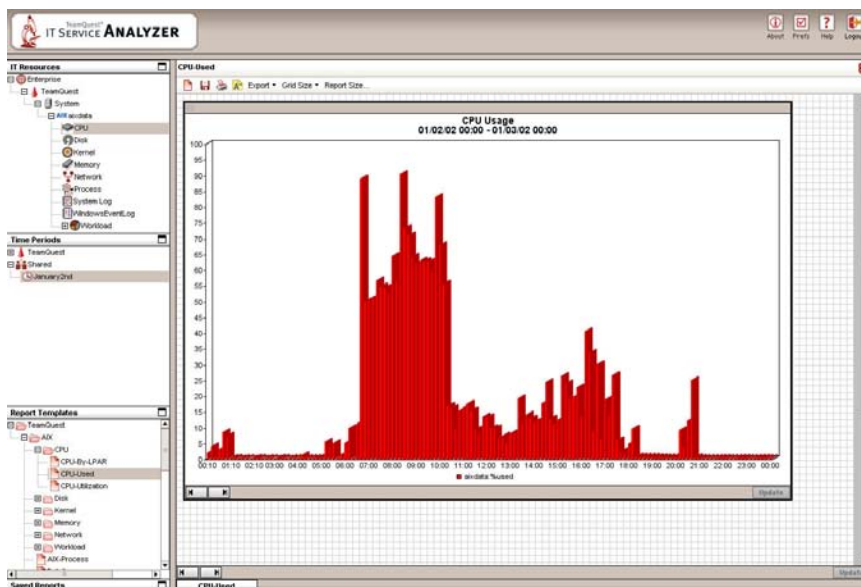


Figure 3  
CPU Utilization

### An Example Using Workloads

TeamQuest Analyzer® chart, left, shows 24 hours of CPU utilization on an IBM F50 PowerPC system. The chart is useful, but it provides a bird's eye view of performance, at best.

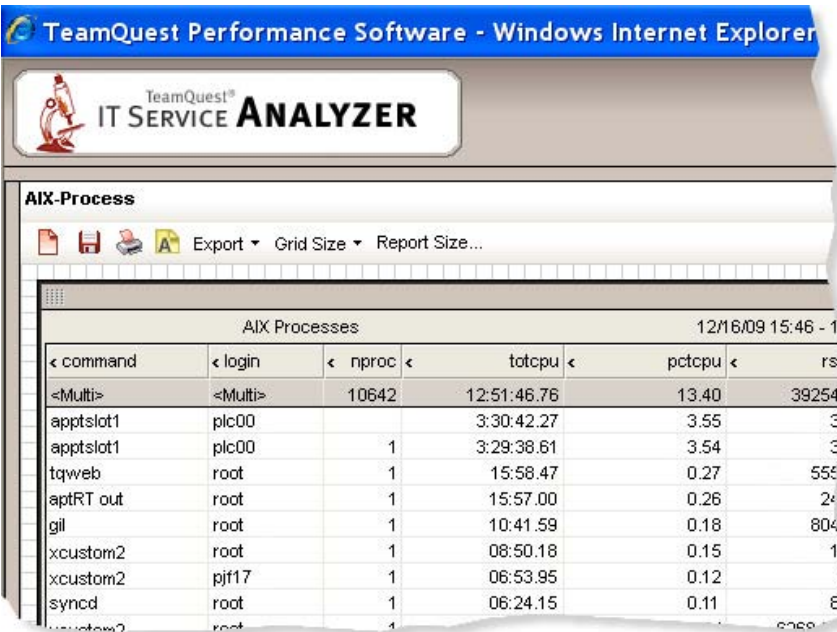


Figure 4  
Process Table

TeamQuest Performance Software how to determine what resource utilization goes with which workload. This is done on a per process level, using selection criteria to tell TeamQuest Performance Software how to determine which processes belong to which workloads.

In Figure 4, left, the “Process Table” chart reveals that during the same 24 hour period, 10,642 individual processes ran on this system. All of the utilization information for all of those processes was displayed together in our CPU utilization chart. Wouldn’t it be nice if we could show a similar chart, but displayed utilization based on the major functions being formed on this system? Using TeamQuest Performance Software, we can do just that, by going through a process called workload characterization.

We will leave the detailed instructions for performing workload characterization to another paper, or you can refer to the TeamQuest Performance Software documentation. In a nutshell, workload characterization requires you to tell

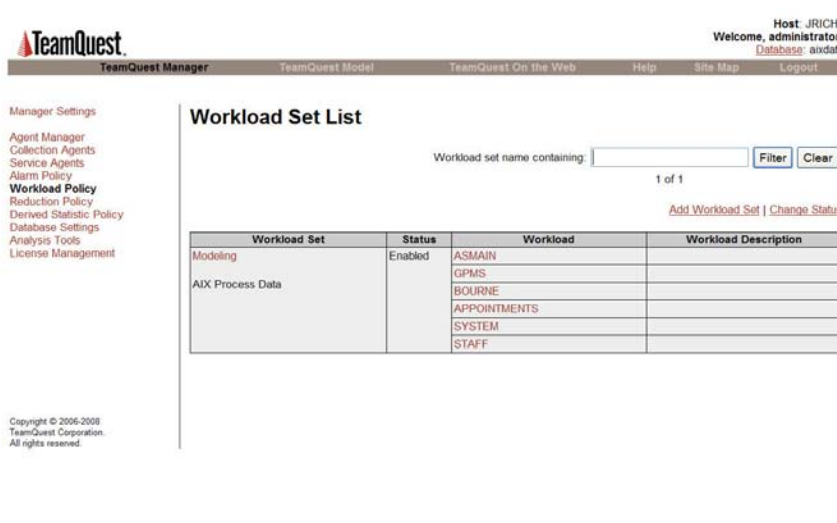


Figure 5  
Workload Definitions

data “falls through the cracks” simply because it didn’t match any of the defined workloads.

Figure 5, left, shows a list of workloads that have been characterized so that the work of each the 10,642 processes is attributed to one of seven workloads. These workloads are defined according to the type of work being done on the system.

If you look carefully you will see six explicitly defined workloads in Figure 5, but we said there were seven. The reason is that there is always an “OTHER” workload in addition to the explicitly defined workloads. Any resource utilization that does not match the characterization for any of the explicitly defined workloads becomes associated with “OTHER.” This ensures that no performance

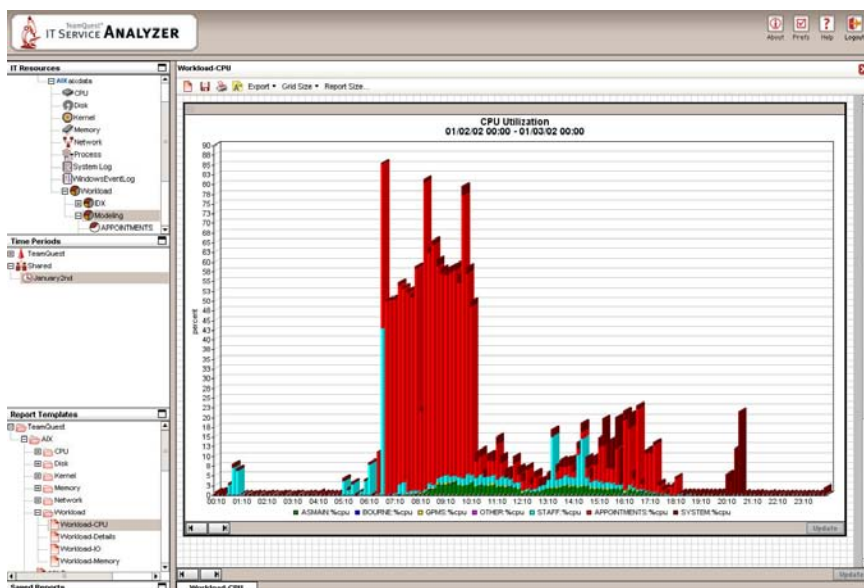


Figure 6  
CPU Utilization by  
Workload

In Figure 6, left, “pink” bars show utilization that did not match the characterization criteria for any of our defined workloads. There is little or no pink in the chart, demonstrating that we have done a good job of explicitly characterizing most of the work done on this server.

All we did was define workloads based on the type of work being performed on this server, but notice how much more useful the information is that is provided in our new chart. Workloads can be very powerful.

## Determine the Unit of Work

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.

To understand the difference, consider measuring the work done at a fast food restaurant. When deciding on the unit of work, you might consider counting the number of customers served, the weight of the food served, the number of sandwiches served, or the money taken in for the food served. This is as opposed to the resources used to accomplish the work, i.e. the amount of French fries, raw hamburgers or pickle slices used to produce the food served to customers.

When talking about IT performance, instead of French fries, raw hamburger or pickle slices, 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.

The examples given in this paper use a server running an appointment scheduling application process, so it seems logical to use a “calendar request” as the unit of work. A calendar request results in an instance of an appointment process being executed.

## Establish Service Levels

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.

In the case of our appointment scheduling application, we might establish service level requirements regarding the number of requests that should be processed within a given period of time, or we might require that each request be processed within a certain time limit. These possibilities are analogous to a fast food restaurant requiring that a certain number of customers should be serviced per hour during the lunch rush, or that each customer should have to wait no longer than three minutes to have his or her order filled.

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, but not surprisingly people frequently resort to setting service level requirements like, "provide a response time at least as good as is currently experienced, even after we ramp up our business." As long as you know how much the business will "ramp up," this sort of service level requirement can work.

If you want to base your service level requirements on present actual service levels, then you may want to analyze your current capacity before setting your service levels.

## Analyze Current Capacity

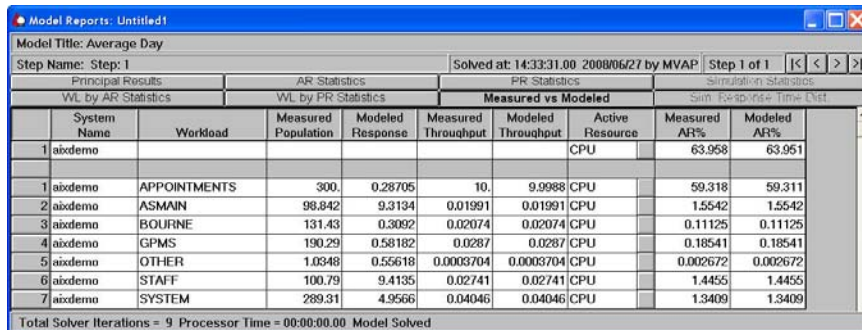
There are several steps that should be performed during the analysis of capacity measurement data.

- a. 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.
- b. 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.
- c. 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.
- d. 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.



## Measure Service Levels and Compare to Objectives

TeamQuest Predictor® includes a tool that can help us check measured service levels against objectives. For example, after building a model of our example system for a three-hour window, 7:00 AM–10:00 AM, the display



Model Reports: Untitled1  
Model Title: Average Day  
Step Name: Step: 1  
Solved at: 14:33:31.00 2008/06/27 by MVAP Step 1 of 1

Principal Results		AR Statistics		PR Statistics		Simulation Statistics	
WL by AR Statistics		WL by PR Statistics		Measured vs Modeled		Sim. Response Time Unit	
System Name	Workload	Measured Population	Modeled Response	Measured Throughput	Modeled Throughput	Active Resource	Measured AR%
1 aixdemo						CPU	63.958
1 aixdemo	APPOINTMENTS	300.	0.28705	10.	9.9988	CPU	59.318
2 aixdemo	ASMAIN	98.842	9.3134	0.01991	0.01991	CPU	1.5542
3 aixdemo	BOURNE	131.43	0.3092	0.02074	0.02074	CPU	0.11125
4 aixdemo	GPMS	190.29	0.58182	0.0287	0.0287	CPU	0.18541
5 aixdemo	OTHER	1.0348	0.55618	0.0003704	0.0003704	CPU	0.002672
6 aixdemo	STAFF	100.79	9.4135	0.02741	0.02741	CPU	1.4455
7 aixdemo	SYSTEM	289.31	4.9566	0.04046	0.04046	CPU	1.3409

Total Solver Iterations = 9 Processor Time = 00:00:00.00 Model Solved

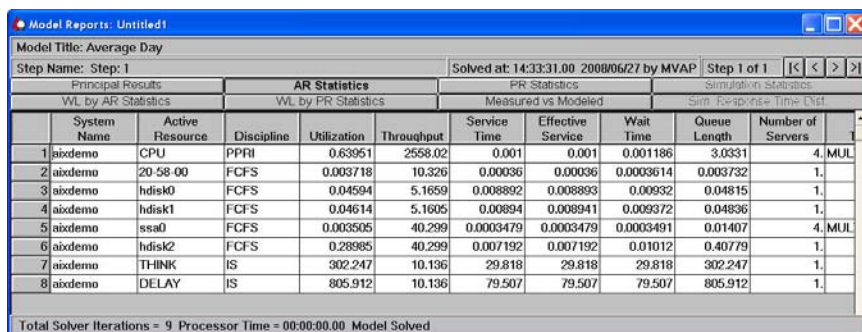
Figure 7  
Response Time and Throughput

An active resource is a resource that is made 100% available once it has been allocated to a waiting process. In this case, the active resource is CPU.

Because no changes have been made in the system configuration, modeled response time and throughput for each workload should closely match reality. In our example, the response time means the amount of time required to process a unit of work, which in the case of our application, is an appointment request process. So this report provides us with an appointment request average response time for each of our workloads that we could compare with desired service levels.

## Measure Overall Resource Usage

It is also 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. If your goal is throughput rather than response time, utilization is still very important. If you have two disk controllers, for example, and one is 50% utilized and the other is swamped, then you have an opportunity to improve throughput by spreading the work more evenly between the controllers.



Model Reports: Untitled1  
Model Title: Average Day  
Step Name: Step: 1  
Solved at: 14:33:31.00 2008/06/27 by MVAP Step 1 of 1

Principal Results		AR Statistics		PR Statistics		Simulation Statistics	
WL by AR Statistics		WL by PR Statistics		Measured vs Modeled		Sim. Response Time Unit	
System Name	Active Resource	Discipline	Utilization	Throughput	Service Time	Effective Service	Wait Time
1 aixdemo	CPU	PPRI	0.63951	2558.02	0.001	0.001	0.001186
2 aixdemo	20-58-00	FCFS	0.003718	10.326	0.00036	0.00036	0.0003614
3 aixdemo	hdisk0	FCFS	0.04594	5.1659	0.008892	0.008893	0.00932
4 aixdemo	hdisk1	FCFS	0.04614	5.1695	0.00894	0.008941	0.009372
5 aixdemo	ssa0	FCFS	0.003505	40.299	0.0003479	0.0003479	0.0003491
6 aixdemo	hdisk2	FCFS	0.28985	40.299	0.007192	0.007192	0.01012
7 aixdemo	THINK	IS	302.247	10.136	29.818	29.818	302.247
8 aixdemo	DELAY	IS	805.912	10.136	79.507	79.507	805.912

Total Solver Iterations = 9 Processor Time = 00:00:00.00 Model Solved

Figure 8  
Overall Resource Usage

of CPU utilization that was shown earlier in Figures 3 and 6.

No resource in the report seems to be saturated at this point, though hdisk2 is getting a lot more of the I/O than either hdisk0 or hdisk1. This might be worthy of attention; future increases in workloads might make evening out the disparity in disk usage worthwhile.

the display left (Figure 7) shows the response time and throughput of the seven workloads that were active during this time.

By looking at the top line of the table, you can tell that the model has been successfully calibrated for our example system, because the total Measured AR% and Modeled AR% are equal. “AR” stands for “Active Resource.”

The table on the left (Figure 8) shows the various resources comprising our example server. The table shows the overall utilization for each resource. Utilization for the four CPUs are shown together treated as one resource, otherwise each resource is shown separately.

Notice that CPU utilization is about 64% over this period of time (7:00 AM - 10:00 AM on January 02). This corresponds with the burst

## Measure Resource Usage by Workload

Model Reports: Untitled1									
Model Title: Average Day									
Step Name: Step: 1									
Principal Results		AR Statistics		PR Statistics		Simulation Statistics			
WL by AR Statistics		WL by PR Statistics		Measured vs Modeled		Sim Response Time Dtd			
System Name	Workload	Active Resource	Discipline	Utilization	Throughput	Service Time	Effective Service	Wait Time	Total Service
1 aixdemo	APPOINTMENTS	CPU	PPRI	0.59311	2372.43	0.001	0.001	0.001185	0.23
2 aixdemo	APPOINTMENTS	20-58-00	FCFS	0.0004263	1.1841	0.00036	0.00036	0.0003614	0.000000
3 aixdemo	APPOINTMENTS	hdisk0	FCFS	0.005268	0.59238	0.008892	0.008893	0.009321	0.0005
4 aixdemo	APPOINTMENTS	hdisk1	FCFS	0.00529	0.59173	0.00894	0.008941	0.009373	0.0005
5 aixdemo	APPOINTMENTS	ssa0	FCFS	0.0004019	4.6209	0.0003479	0.0003479	0.0003491	
6 aixdemo	APPOINTMENTS	hdisk2	FCFS	0.03324	4.6209	0.007192	0.007192	0.01012	0.003
7 aixdemo	APPOINTMENTS	THINK	IS	294.13	9.9988	29.417	29.417	29.417	29.
8 aixdemo	APPOINTMENTS	DELAY	IS	2.9996	9.9988	0.3	0.3	0.3	
9 aixdemo	ASMAIN	CPU	PPRI	0.01554	62.167	0.001	0.001	0.001173	3.1
10 aixdemo	ASMAIN	20-58-00	FCFS	0.0008294	2.3037	0.00036	0.00036	0.0003614	
11 aixdemo	ASMAIN	hdisk0	FCFS	0.01025	1.1524	0.008892	0.008893	0.00932	0.51
12 aixdemo	ASMAIN	hdisk1	FCFS	0.01029	1.1512	0.00894	0.008941	0.009372	0.51
13 aixdemo	ASMAIN	ssa0	FCFS	0.0007819	8.9901	0.0003479	0.0003479	0.0003491	
14 aixdemo	ASMAIN	hdisk2	FCFS	0.06466	8.9901	0.007192	0.007192	0.01012	3.2
15 aixdemo	ASMAIN	THINK	IS	0.98839	0.01991	49.651	49.651	49.651	49.
16 aixdemo	ASMAIN	DELAY	IS	97.668	0.01991	4906.29	4906.29	4906.29	4906.
17 aixdemo	BOURNE	CPU	PPRI	0.001112	4.45	0.001	0.001	0.001267	0.21
18 aixdemo	BOURNE	20-58-00	FCFS	0.000005706	0.01585	0.00036	0.00036	0.0003614	
19 aixdemo	BOURNE	hdisk0	FCFS	0.00007051	0.007929	0.008892	0.008893	0.009321	0.003
20 aixdemo	BOURNE	hdisk1	FCFS	0.00007082	0.007921	0.00894	0.008941	0.009373	0.003

Total Solver Iterations = 9 Processor Time = 00:00:00.00 Model Solved

Figure 9  
Resource Utilization  
by Workload

Figure 9, shows the same period again, only now resource utilization is displayed for the APPOINTMENTS workload. Note that this particular workload is using 59% of the CPU resource, nearly all of the 64% utilization that the previous table showed as the total utilization by all workloads. Clearly, the APPOINTMENTS workload is where a capacity planner would want to focus his or her attention, unless it is known that future business needs will increase the amount of work to be done by other workloads on this system. In our example, that is not the case. Ramp-ups in work are expected mainly for the APPOINTMENTS workload.

The previous charts and tables have been useful for determining that CPU Utilization is likely to be a determining factor if the amount of work that our system is expected to perform increases in the future. Furthermore, we were able to tell that the APPOINTMENTS workload is the primary user of the CPU resources on this system.

This same sort of analysis can work no matter how you choose to set up your workloads. In our example, we chose to treat appointment processes as a workload. Your needs may cause you to set up your workloads to correspond to different business activities, such as a Wholesale Lumber unit vs. Real Estate Development, thus allowing you to analyze performance based on the different requirements of your various business units.

## Identify Components of Response Time

Next we will show how to determine 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 you should concentrate your efforts to optimize performance. Using TeamQuest Predictor we can determine the components of response time on a workload by workload basis, and you can predict what the components will be after a ramp-up in business or a change in system configuration.

A components of response time analysis shows the average resource or component usage time for a unit of work. It shows the contribution of each component to the total time required to complete a unit of work.



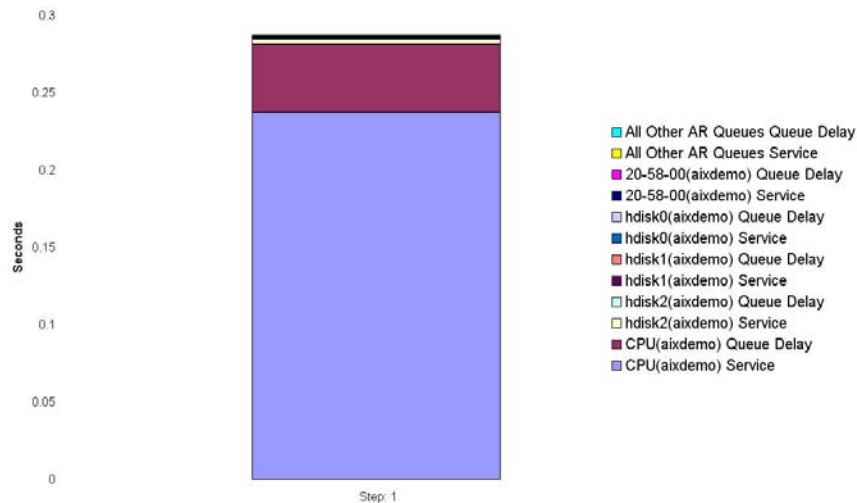


Figure 10  
APPOINTMENTS  
Components of  
Response Time

Figure 10, left, shows the components of response time for the APPOINTMENTS workload. Note that CPU service time comprises the vast majority of the time required to process an appointment. Queuing delay, time spent waiting for a CPU, is responsible for the rest. I/O resources made only a negligible contribution to the total amount of time needed to process each user call.

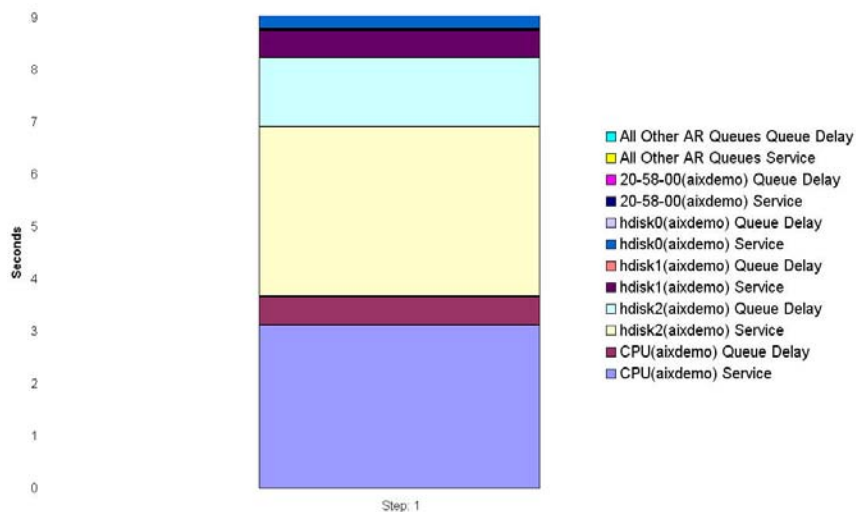


Figure 11  
ASMAIN Components  
of Response Time

The ASMAIN workload, shown in Figure 11, is more balanced. There is no single resource that is the obvious winner in the contest for the capacity planner's attention (however, make note of the queue delay for hdisk2.)

## Plan 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.



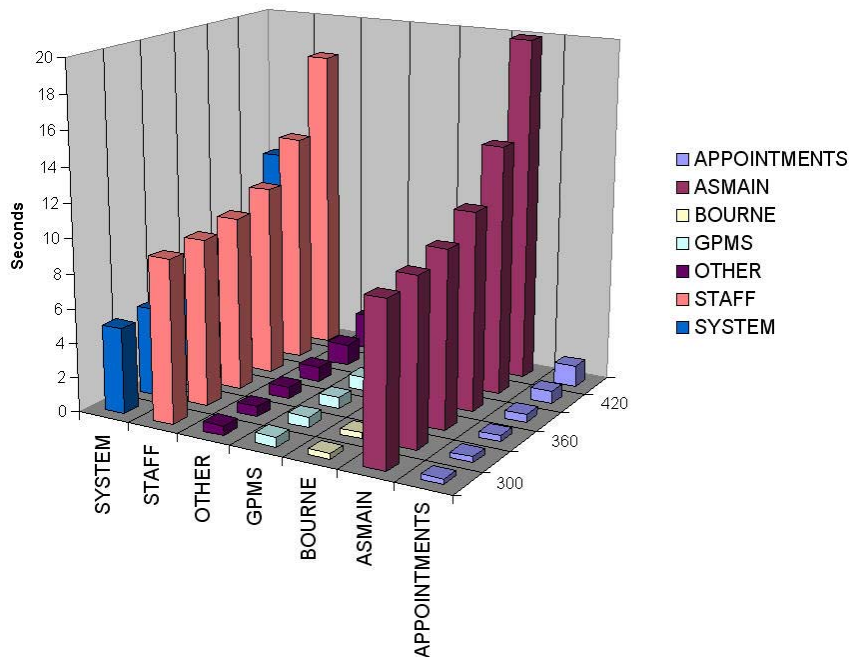


Figure 13  
Predicted Response  
Time

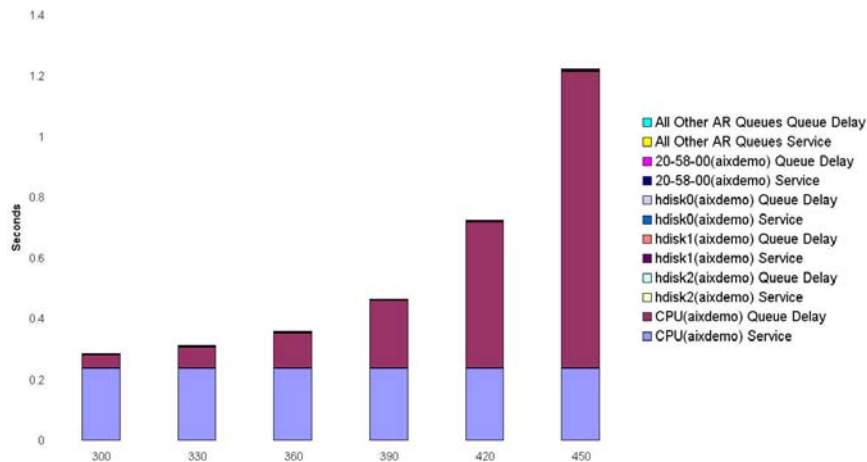


Figure 14  
Predicted Response  
Time

TeamQuest Predictor will predict performance of the current system configuration for each of the steps we have set up. Figure 13 is a chart generated using TeamQuest Predictor showing the predicted response time for each workload.

As we can see, the response time starts to elongate after 390 users.

Figure 14 is a chart showing predicted response time using a stack bar chart that also shows the components of response time. Notice the substantial increase in CPU wait time after the number of users reaches 360. It seems that the performance bottleneck is CPU resource.

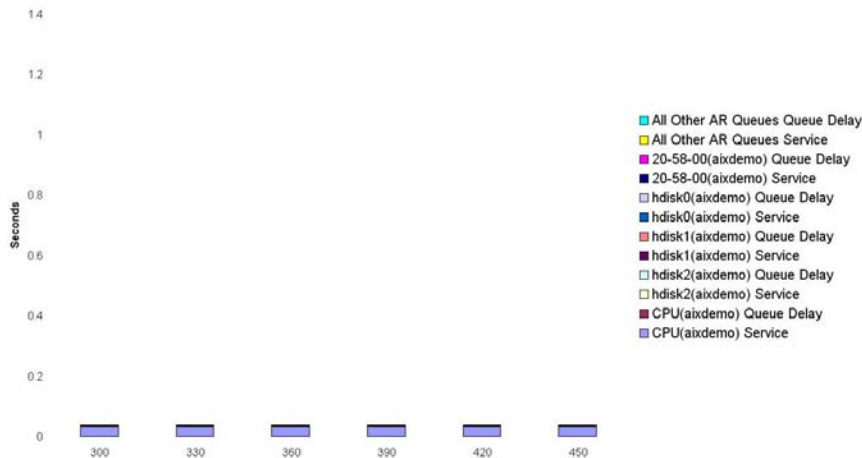


Figure 15  
Predicted Response  
Time After Upgrade

Figure 15 shows the same stack bar chart, but this time TeamQuest Predictor was told to predict performance if the system involved was changed to a p670 1100Mhz 4-CPU system. Clearly, the newer, faster architecture not only allows us substantial growth, but reduces our overall response time to a more realistic level and still allows us the headroom to experience additional growth if needed.

## Capacity Planning Process

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

1. Determine service level requirements
  - a. Define workloads
  - b. Determine the unit of work
  - c. Identify service levels for each workload
2. Analyze current system capacity
  - a. Measure service levels and compare to objectives
  - b. Measure overall resource usage
  - c. Measure resource usage by workload
  - d. Identify components of response time
3. Plan for the future
  - a. Determine future processing requirements
  - b. 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.

# TeamQuest Corporation

**[www.teamquest.com](http://www.teamquest.com)**

Follow the TeamQuest Community at:

## **Americas**

One TeamQuest Way  
Clear Lake, IA 50428  
USA  
+1 641.357.2700  
+1 800.551.8326  
[info@teamquest.com](mailto:info@teamquest.com)

## **Europe, Middle East and Africa**

Box 1125  
405 23 Gothenburg  
Sweden  
+46 (0)31 80 95 00  
United Kingdom  
+44 (0)1865 481 424  
Germany  
+49 (0)69 6 77 33 466  
France  
+33 (1)40 90 30 93  
[emea@teamquest.com](mailto:emea@teamquest.com)

## **Asia Pacific**

35/F Central Plaza  
18 Harbour Road  
Wanchai  
+852 2824 8510  
[asiapacific@teamquest.com](mailto:asiapacific@teamquest.com)

**Copyright ©2010, 2013 TeamQuest Corporation  
All Rights Reserved**

TeamQuest and the TeamQuest logo are registered trademarks in the US, EU, and elsewhere. All other trademarks and service marks are the property of their respective owners. No use of a third-party mark is to be construed to mean such mark's owner endorses TeamQuest products or services.

The names, places and/or events used in this publication are purely fictitious and are not intended to correspond to any real individual, group, company or event. Any similarity or likeness to any real individual, company or event is purely coincidental and unintentional.

NO WARRANTIES OF ANY NATURE ARE EXTENDED BY THE DOCUMENT. Any product and related material disclosed herein are only furnished pursuant and subject to the terms and conditions of a license agreement. The only warranties made, remedies given, and liability accepted by TeamQuest, if any, with respect to the products described in this document are set forth in 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 be very careful to 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.