PREDICTING COMPUTING SYSTEM CAPACITY AND THROUGHPUT

Craig A. Shallahamer

Practice Manager - System Performance Group

Oracle Services, Portland, Oregon, USA

1. Introduction

Business managers need trustworthy answers to questions such as, "My business is going to double. Can my current system handle the load?" and "Will the proposed technical architecture handle our production load?" Answering these types of capacity planning questions, with any respectable degree of confidence, is hard. However, the questions still must be answered and that's what this paper is all about.

Capacity planning questions ultimately center around deciding where to best concentrate one's effort to ensure a new or existing system will meet performance requirements. While quit surprising at first, capacity planning questions can be thought of in economic terms. It is extremely rare that an exact match of a proposed system can be prototyped. But does the prototyped system have to be an exact replica? And if not, how close does it need to be? (Figure 1) The economic question then becomes one of directing limited resources in areas that require attention to ensure performance will meet the requirements.

It's much easier to talk about models and mathematics than it is to address the underlining business problems. Our aim in constructing this paper is to help folks begin to tackle complex capacity studies by not only addressing the overall project and technical issues, but by helping folks translate the technical information into meaningful business information.

To achieve our objectives, we first present an overall project plan followed by three proven types of capacity plans. We then drill down into how one approaches the project from a data gathering and modeling perspective. And finally, we end with how one addresses conflicting constraints.

1.1 Capacity and Throughput

The words *capacity* and *throughput* are important concepts used throughout this paper. They are closely related and many times incorrectly used interchangeably.

The concept of capacity is the measurement of physical space. Metrics like mega-bytes, giga-bytes, number of CPUs, and amount of physical memory are all measurements of

physical space. Using this definition, capacity can be used to measure both space capacity and processing capacity. For example, the statement, "The forecasted transaction volume will require 10 GB of disk space" relates to *space* capacity. Whereas the statement, "The forecasted transaction volume will require twenty disks to maintain acceptable i/o response times" relates to *processing* capacity.

The concept of throughput is the measurement of some action occurring within a time interval. For example, transactions per second or invoices entered per hour are examples of throughput.

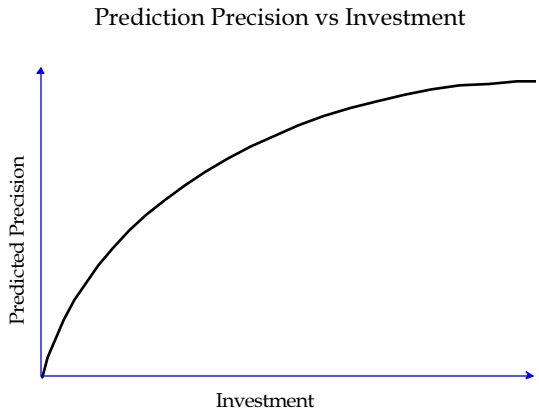

Figure 1. This graphs shows why establishing the minimum level of required precision is so important. Simply throwing more money into a capacity plan will not always increase the prediction precision.

1.2 Case Study Introduction

Throughout this paper the following case study will be referenced. Direct Connect Devices is implementing an 1,200 user Oracle Manufacturing system with users located in both Portland, Oregon and Dallas, Texas. The IT Director knows there is virtually an infinite number of areas where performance could fall short of user expectations. His job is to minimize and direct the possible risk that this could occur to reduce the business impact. The Director has a problem. Only two people in his entire organization are qualified to undertake this study, yet guaranteeing acceptable performance will require a team of ten people over a six month period. The resources and the time are simple not available. He must therefore, find a way to reduce the chance of unacceptable performance by concentrating the teams efforts in specific areas that pose the highest degree of uncertainty and negative business impact.

2. Project Plan

A capacity planning study is no different than other projects in that a solid project plan and approach must be established. The project plan outlined below is structured to directly address and meet the three types of capacity plans discussed the paper.

- Strategy Phase
  - determine goal and scope
  - determine method(s)
  - determine model(s)
- Build Phase
  - characterize workloads
  - model environment
- Execute Phase
  - simulate workloads
  - apply math. model
  - validate everything
- Interpretation Phase
  - analyze results
  - consolidate results
  - present results

Figure 2. Just as with any significant project, a capacity plan needs a solid and proven project plan. This figure shows a project plan shell we have successfully used.

2.1 The Strategy Phase

The theme of the *strategy stage* is to build a foundation upon which the rest of the project rests. The strategy phase consists of three main tasks. They are: determine project goal and scope, determine which method or methods to use, and determine which modeling technique or techniques to use. The various methods and modeling techniques will be discussed in a later section. As you will see, once the methods and the modeling environments have been chosen, it becomes obvious which of the three capacity plans presented in this paper to use.

A capacity planning expert should review your strategy before any building begins. If the strategy is not realistic or does not map the required data inputs with what data is actually available, the strategy stage will have to be repeated.

2.1.1 Determine Goal and Scope

The project *goal and scope* determine exactly what the deliverables will include and what the deliverables will not include. For example, "The goal of this Limited Performance Assurance Test is to determine if the current computing system in its current configuration will support the projected workload requirements. The results will not include expected response times but rather approximately where the system will fail to support the required workload."

Establishing what the study will and will not state is paramount. People tend to push predictions to a degree of precision that the study was not structured to provide. The key parties that will be judging the project's success must agree on exactly what you will be telling them and to what degree of precision. If this is not established during the strategy phase, the worthiness of the entire study will be questioned.

2.1.2 Determine The Method

After the project goal and scope have been agreed upon, a *method* (sometimes called the *approach*) must be chosen to ensure that the project goal will be reached, and to determine how and what statistics need to be gathered. Later in the paper, five different methods will be discussed. The capacity planning committee must approve of your approach. If they do not agree on your approach, your results will be in question and the study deemed of no value.

2.1.3 Determine The Model

All of our capacity planning studies use models. They may be simple arithmetic models, queuing models, or simulations. Determining which model or models to use is dependent upon which method you have chosen and what data inputs are available. For example, the summation *method* requires real production data to be captured which feeds nicely into a regression *model*.

2.2 The Build Phase

The theme of the *build phase* is to construct the framework to execute upon. A key component of all capacity planning studies is characterizing the workload and constructing a mechanism to create a production environment (if production data is not available) .

2.2.1 Characterizing the Workload

Characterizing the workload is simply understanding what work will be performed and when. For example, do we want to characterize month-end or quarter-end? Do we want characterize typical day-time usage, night-time usage, or if the mirrored disk subsystem is resyncing during any of the above situations? Oracle Applications programs can be classified as either OLTP or batch. And each process consumes disk i/o, memory, CPU, and possibly network resources. Combine this with a variety of workload scenarios and it becomes very difficult to construct an environment to model the "real" environment. Skilled capacity planners are experts in determining the minimum number of required workloads that will provide real value and minimize project investment.

One of the most misunderstood concepts of planning capacity is the answer to the question, "At what load should we plan for capacity?" A careful analysis of this question leads us to some very profound yet simple conclusions.

Let's take a step back and ask ourselves what our objective is when planning capacity. What do our customers want from their computing system? The answer is really quite simple. They want a computing system that will allow them to meet their business needs.

When architecting a system to meet business needs, we need to ensure the needs will *always* be met. This leads us to ask the question, "What are the needs we must accommodate?" An example need may be "Ninety-five percent of the time, this query must complete within five seconds." Designing a system that meets documented require-

ments seventy-five percent of the time is simply not acceptable. With these thoughts in mind, we can answer some fundamental capacity planning questions.

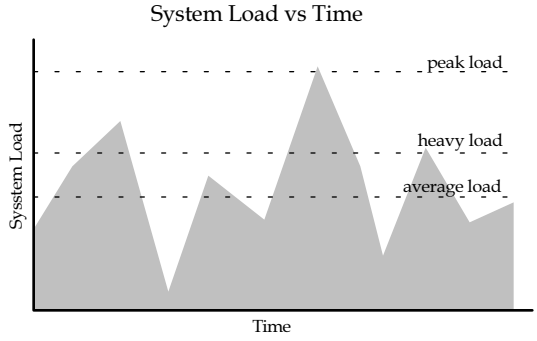

Figure 3. To ensure the computing system can always meet business throughput requirements, computing system capacity should be planned for peak or worst case scenarios, not average or "typical" usage.

Should we simulate and model typical performance? No. There is no such thing as *typical* in the capacity planning realm. We could ask the question, "What is the typical system load at 4:30 PM on the first Monday of each month?" But so what! If the system is more heavily loaded at 3:00 PM on the first Monday of each month, we should then plan our capacity to meet the 3:00 PM requirements. Ruminating a bit, it is clear we need to plan capacity for peak usage. Planning capacity during peak usage ensures business can continue at an acceptable throughput regardless of the day or time.

How many times have you had to bite your tongue when a user said, "The system is so slow today. I can barely get any work done." and you quickly responded with, "Well it's month end." Systems are designed to empower people to do their work and to empower a business. Just because the system is operating at peak load is no excuse for unacceptable performance.

Properly and honestly characterizing the workload can be very difficult, especially during pre-production implementation. While it may be difficult, the capacity planning committee must agree on the workload characterizations. When the final study is presented, one of the first place folks look to invalidate the study will be related to the workload characterization.

2.2.2 Building the Models

Modeling is an integral part of planning capacity. So integral that we use multiple models in our engagements for validation purposes. Models can be broadly classified

into either simulation models or mathematical models. Mathematical models require input from either a production system or a simulated production system. If mathematical models are being used, unless you are observing a production system, some form of workload simulation must occur to "feed" the mathematical models.

Constructing simulation and mathematical modeling tools can be very difficult and time consuming. We use generic model templates which we tailor for each client. By doing so, we save time and money while continually improving our models. Further modeling detail is covered in a later section.

2.3 The Execution Phase

The theme of the *execution phase* is use the items built during the build phase. The execution phase is generally shorter than folks suspect. The build phase and the analysis phase usually take longer then running the simulations, gathering the data, and enter the data into the mathematical models. At this point, a rigorous quality assurance check must be performed to validate the models. Even simple mathematical mistakes can make a tremendous difference in what your model predicts.

2.3.1 Simulate the Workload

If the workload has been properly characterized and the build phase was executed properly, simulating the workload should be very straightforward. Simulating a production environment is serious business. Very significant organizational, budgetary, and purchasing decisions are based, in part, on the results of the simulation. Since the goal of a simulation is to simulate reality, at a minimum the below items should be investigated.

- Ensure nothing but the simulation is occurring on the machine.
- Understand the affect of your monitoring tools.
- Ensure there is no unusual Oracle contention. Usually this occurs when there is not enough data or the simulation repeatedly uses the same data. Latch contention, a very high data block cache hit ratio, and low i/o activity are all signs that the simulation is flawed.
- Tune slow processes. If you discover a process will not be acceptable in production, then to model the processes is inappropriate. Have the processes tuned before the simulations begin.

2.3.2 Apply the Mathematical Model

After the model has been built, the workload simulated, and the statistics gathered, you are ready to input the statistics into the mathematical models. This is another time to have your progress checked by an expert.

2.3.3 Validate Everything

Books have been written about validating predictive mathematical models and simulations. For this paper, we want to stress the importance of validation before anyone begins making predictions. We listed above a few simulation checks that can be performed. In addition, basic "sanity checks" should be performed and questions like, "Is this what you expected?" should be asked and answered. While the next phase is the interpretation phase, at this point you should be confident your model is accurately making predictions based upon your input.

Mathematical model validation can be as simple as plugging in the appropriate data gathered from the simulation and checking if the model predicts what actually happened in the simulation. But let's suppose the mathematical model does not predict what the simulation results where. Does one quickly alter the mathematical model? Don't quickly do anything. Take a step back and think about what could have caused the discrepancy. Then investigate both the mathematical model and the simulation. The mathematical model may be predicting satisfactorily but the data gathering mechanism may be functioning incorrectly. It could be that the model forgot to take into account some type of processing overhead. There are many things that can contribute to a simulation and mathematical model mismatch.

Once the mathematical model has been validated, you have a powerful predictive tool that can be used long after the capacity planning study is completed. As long as the inputs are updated and the mathematical model validated whenever possible, the model output is a strong predictive tool.

2.4 The Interpretation Phase

The theme of the *interpretation phase* is to understand the simulation and model predictions well enough to discuss them from a business and technical perspective. The interpretation phase will take a different shape depending on the study scope. Once the analysis is complete, all the information should be available to consolidate your thoughts, make recommendations, and then construct the final study document and/or presentation materials.

3. Types of Capacity Plans

When is enough, enough? How much should I invest in a capacity planning effort? Is it wise to invest in a three month study when all I want to know is if there appears to be significant risk in not meeting the performance requirements? Probably not. However, it may be a wise investment to in a three month study if you need to know what type of response time degradation will occur when 1,200 users are hard at work trying to desperately ship as many products out the door as possible during quarter-end close.

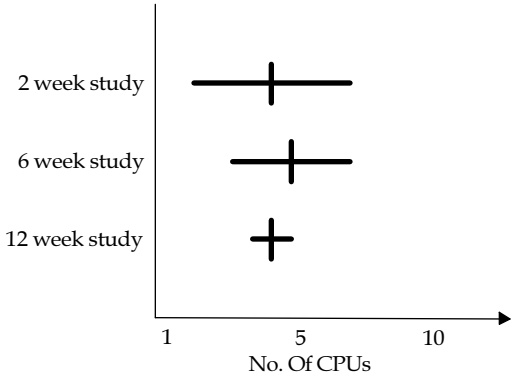

Figure 4. Different capacity plan types yield results with different precision. This figures shows the two week study predictions four CPUs will be required, plus or minus four CPUs. The six week study predicts five CPUs, plus or minus three CPUs, and the twelve week study predicts four CPUs, plus or minus one CPU.

These types of questions relate to the allocation of limited resources towards providing the highest return on investment. To address different levels of required precision, our team provides three main capacity planning study types. They are the *Initial Performance Readiness Review*, the *Limited Performance Assurance Test*, and the *Classic Performance Assurance Test*.

The different predictive study types all revolve around the concept of *precision* and *statistical confidence* (Figure 4). *Precision* is a relative term that is used to describe how much one should trust the predictive power of a study. *Confidence levels and intervals* are terms used to statistically describe precision. For example, a high precision study (e.g., Classic Performance Assurance Test) may predict an average of 12 CPUs are required at a confidence level of ninety-percent and a confidence interval of two CPUs (i.e., plus or minus one CPU). Whereas a lower precision study (e.g., Limited Performance Assurance Test) may also predict an average of 12 CPUs are required at a confidence level of ninety-percent but at a confidence interval of six CPUs (i.e., plus or minus three CPUs).

A low precision study (e.g., Initial Performance Readiness Review) is not bad or good in itself. Just as with purchasing a car, when determining the required precision, the objectives are to match your requirements (e.g., access, speed, mileage, etc.) with your constraints (e.g., financial situation, time, etc.). If you require a very expensive car (i.e., BMW M3) and you are relatively unconstrained (i.e., rich and wealthy[1]) you will attempt to purchase one. The same is true with capacity planning studies. Your objective is to determine the correct strategy to match the required study precision.

---

[1] At this particular moment (July 12, 1995 at 6:15 PM), according to United States congressman Richard Gephardt, being "rich and wealthy" begins at an annual salary of $60,000. The current "rich and wealthy" figure will most likely be different at the time of your reading.

Many times we have heard someone say, "We can't use this technique because it makes bogus predictions." What the "confidence-lingo-challenged" individual is saying is, "The confidence interval this capacity plan produces does not meet our precision requirements." In many cases, low precision predictions (i.e., large confidence interval), are all that are necessary, or all that are possible. Performing a high precision (i.e., small confidence interval) study can be very expensive.

Study Type	Purpose	Deliverables
Initial Performance Readiness Review	Identifying where to best invest people's time to ensure performance meets requirements	• Resource allocation matrix • Mathematical models
Limited Performance Assurance Test	Understanding workload characteristics and how this will affect the production system.	• Workload characterization • Key performance parameters • Workload balance equation • Mathematical models • Resource allocation matrix • Business impact
Classic Performance Assurance Test	Understanding how workload characteristics affect user response times and the business	• Response time degradation • Workload characterization • Key performance parameters • Workload balance equation • Resource allocation matrix • Business impact • Mathematical models

Figure 5. Types of Capacity Planning Studies.

Capacity studies are no different than other projects in that they must make good business sense. It is irresponsible to undertake a study without first considering the investment versus the return in terms of business value. The three different capacity plans presented below require a different level of investment and returns a different value to the business. The trick is maximizing the value while minimizing the investment. The information presented below should help you make a better investment choice.

3.1 Initial Performance Readiness Review

While unfortunate, we have seen many companies avoid a simple performance readiness review only to have unidentified areas of potential risk become major issues when the application was placed into production. With so many different vendors involved in today's complex application implementations, it is a wise investment to perform a

low cost performance readiness review. The Initial Performance Readiness Review is targeted towards identifying where allocating resources for future study will reap the greatest rewards.

An Initial Performance Readiness Review (Review) can be performed relatively quickly and is guaranteed to liven up any meeting. With only a handful of questions answered regarding the proposed system, experience, a reference site, and some very simple arithmetic models, a quick review can be completed. When presenting the results, folks involved with the areas targeted for further investigation quickly want a more precise study performed. And that's exactly what we want: to identify areas of possible significant risk, concentrate our efforts in those areas, and then to drill down, i.e., produce a more precise study.

The Review produces a very simple and straightforward "Resource Allocation" matrix. The matrix simply shows how and where to concentrate the available resources to maximize their impact in terms of ensuring performance requirements will be met. Referring to our Direct Connect Devices case study, it was discovered the implementation team expects twenty simultaneous batch processes will be required to run during quarter-end close. Based upon reference site information, experience, and the proposed technical architecture, the best use of people's time would be to further investigate server CPU and then memory capacity.

Initial Performance Readiness Reviews are performed everyday. It's amusing that folks who denounce low confidence predictions are usually the most the dedicated practitioners. Have you ever heard someone say, "You mean to tell me you can not predict how many users my system?" and then they will say, "Yeah, I'd say we could handle another fifty users on this system." It happens all the time! What this person did in their head was a quick mental calculation to "ballpark" the requirements versus the capacity. Put another way, they quickly isolated areas that warrant attention, i.e., a higher confidence level study. This is exactly what an Initial Performance Readiness Review is all about: identifying where to best invest people's time to ensure performance will meet the user's requirements.

3.2 Limited Performance Assurance Test

Most often the Initial Performance Readiness Review does identify areas that warrant future review. This is where the Limited Performance Assurance Test comes in. This "test" is focused upon understanding how people will be using the system (characterizing the workload), which parameters dramatically effect performance (e.g., the number of simultaneous batch jobs), determining the workload balance options (e.g., one batch job equals twelve OLTP users), and if necessary, which area(s) should be the focus for a more detailed study (e.g., Classic Performance Assurance Test).

One way of looking at capacity planning is that it is a quest for truth. We have found forming a capacity planning committee is one of the key aspects of a successful capacity planning study. Performance assurance tests dig deep into a company's organizational, operational, and many times political infrastructure. This may sprout unusual obstacles

which a committee dedicated to pursuit of truth can help overcome. It's better to deal with uncomfortable situations before a system is placed into production.

A significant difference between the Initial Performance Readiness Review and the Limited Performance Assurance Test is that the latter requires a prototype system to simulate a production load so actual statistics can be gathered.

Once the statistics have been gathered (know as the "method" and discussed later in this paper) they can be directly analyzed and also applied to a variety of predictive models. Simulating many different scenarios is usually not possible due to time constraints. However, a model's parameters can quickly altered. Once the predictive model has been validated against the actual statistics gathered, the model can be used to quickly make predictions under many different workload scenarios. Later in this paper we will discuss various modeling techniques.

3.3 Classic Performance Assurance Test

Reality is the best predictor. The Classic Performance Assurance Test (CPAT) is aimed at simulating reality. As we will discuss, while a CPAT can be very costly, in terms of predictive power it is the one of the best ways to ensure acceptable performance without a real-life production system.

A substantial advantage of the CPAT over the Limited Performance Assurance Test is the CPAT can provide response time degradation figures. To put this into perspective, the Initial Performance Readiness Review can roughly predict if the technical architecture is a good fit, the Initial Performance Assurance Test will predict where the technical architecture will break, and the CPAT will also provide details about what kind of response time users can expect before the system "breaks."

The best solution, which many large implementations now undertake, is to perform repetitive Initial Performance Readiness Reviews using different technical architectures and then performing a few Limited Performance Assurance Tests. When a suitable architecture becomes apparent, the Classic Performance Assurance Test is performed.

Today's implementations are complex. Suppose we complicate our case study by introducing a client machine, fifteen users will be using the system from Hong Kong, and the business has distinct seasonal characteristics. Now the situation looks like;

- multiple peak processing loads; quarterly, yearly, and seasonally,
- users are spread throughout three different physical locations and three different time-zones, and
- 1,215 concurrent OLTP users.

In situations like this, having real people simulate their proposed workload may simply be impossible. The resources required (E.g., information, people, time) to perform a real-life test can be massive and may not be available at any cost.

As we stated before, reality is the best predictor. CPATs simulate real-life production by using, for example, Remote Terminal Emulators (RTEs) to simulate (sometimes called

benchmarking or stress testing) the forecasted production load. Remote Terminal Emulators are not free and the time involved to implement a real-life workload mix can take weeks or even months. However, once the simulation has been setup, many different workload scenarios can tested that are probably not logistically possible with people. It boils down to risk, cost, and benefits. And that means meetings, negotiations, and honest dialog. That's hard.

As with the Limited Performance Assurance Test, combining actual statistics with predictive models will allow many business scenarios to be run. Plus, when the system goes into production the model can be validated and future growth predictions can be modeled.

A word of warning. A "high confidence" prediction rapidly weakens when one pushes technology to new limits. The models presented in this paper will not predict an algorithm break down. For example, some deep dark algorithm may not scale well when stressed in a certain way. These type of situations are extremely difficult, if not impossible, to model. Certainly with the modeling techniques we use, predicting algorithmic break downs would be inappropriate.

4. Methods or Approaches

The purpose of the *method* or *approach* is to realize the project goal and scope, to determine what statistics will be gathered, and how the statistics will be gathered. Scope asks the question the methods and models must answer. In addition, it states to what precision (i.e., confidence level) the question must be answered.

Gaining method, model, and ultimately study prediction confidence from the capacity planning committee is extremely important. So much so, that during the strategy phase one of the first tasks is to form a capacity planning committee consisting of management, users, software vendors, hardware vendors, and MIS folks. A straightforward and clearly presented method will help ensure both the technical and the non-technical personnel understand your approach. If people do not understand your methods, then they must have an overwhelming faith in your abilities or they will doubt your conclusions.

The next few sections will detail five different methods for scoping a project, and in general, how the data will be gathered. As stated above, each method requires different inputs and produces output at various confidence levels.

4.1 Estimation Method

Everyone has used the *estimation method* at some point in their career. All one must do when using the estimation method is ask a few good questions, have applicable experience, and possess a good estimation model. The Initial Performance Readiness Review typically uses the estimation method. When computing systems are sold, someone asked questions (hopefully), shoved the numbers in a spreadsheet, and exclaimed, "This is the box you should buy!" While this example has been simplified, this type of situa-

tion frequently occurs and we can learn quite a bit about when the estimation method should and should not be used.

The beauty of the estimation method is it requires a very low level of effort. Various predictions can be produced within a few days or even a few hours. This is an excellent way to "ballpark" the system capabilities to ensure performance meets the user's expectations.

Suppose your memory estimate predicts 500 MB of RAM is required and your machine is configured with 1,500 MB of RAM. In this case, the risk of being undersized is very low and a more detailed memory study may not be necessary. However, if your estimate predicts 1,400 MB of memory, the risk of being undersized is substantially higher. In this case, it may be wise to perform a higher precision investigation to assess if the identified risk presents a significant risk to the project's success.

Another benefit of the estimation method is it does not require your application to be in production. Because of its inherent low output confidence, you can use information from your vendor, use your past experience, and use the experience from those who have measured the resource requirements from another similar production system.

4.2 Summation Method

The *summation method* can be used when a question such as "How many CPU seconds will it take to process 800 orders?", "How many disks will be required to store five million journal entries?", or "My business is going to double. Can my current system handle the load?" must be answered and the associated production system is available for data gathering.

Weather your gathering data regarding memory, CPU, I/O throughput, or disk space requirements, the summation method requires high level throughput metrics and the computing resources required to achieve the metrics to make predictions. For example, a high level throughput metric may be orders processed per day, order lines processed per day, journal entries per day, invoices entered per day, the number of Oracle database blocks required per journal entry, or orders shipped per day. Standard computing resource data (CPU, memory, disk I/O, and network) need to be gathered. Once you gather the information, you can perform a regression analysis to produce a fairly high confidence model.

As mentioned above, a production system is required. The summation method is a terrific way to reduce a very complex computing system down to a few mathematical equations. If production load data can not be gathered, the predictive capacity of this method diminishes and predictive risk dramatically increases. With all this considered, the summation method can be a very accurate method with minimal effort required.

4.3 Process Methods

The next set of methods to characterize the workload are operating system process or transaction processing related. This is in contrast to the estimation and summation

methods where the system was treated as a single unit. The processing methods can be used for both throughput and disk sizing predictions. The process methods have the advantage of allowing one to manipulate the inherent workload characterization risk by combining the benefit of each method to arrive at a hybrid method. For example, if you know of a key business process, you can factor that process directly into your workload along with other large processes. Using these techniques increases prediction confidence and allows you to simulate and model very complex systems more simply.

I will discuss three process related methods. They are the *key process*, the *large process*, and the *all process* methods. Each method has an underlying theory or emphasis behind it which allows it to stand on its own. However, as mentioned above, the power of the process related methods is they can be used together to produce higher confidence predictions.

The *key process* method is based upon the assumption that if the key business processes are modeled then business will perform within acceptable service levels. An underlying assumption is that other non-key business processes do not constitute a significant resource drain and therefore do not need to be modeled. Practically speaking, both key processes and large process data must be gathered. The important point is if key business processes are not modeled, while other application areas may perform well, if the key business processes do not perform well, the application system may not be successful. It should be clear that while the key process method has its place, it should be used with another method to ensure completeness.

The *large process* method is based upon the assumption that if the large (i.e., heavy) computing system resource processes, modeled then business will perform within acceptable service levels. An underlying assumption is that other processes do not constitute a significant resource drain and therefore do not need to be modeled. The large process method is typically combined with the key process method to ensure completeness.

As you might suspect, the *all process* method takes all operating system processes into account during modeling. While this technique produces very accurate results, unless the appropriate model is applied, it is extremely time intensive and pushes models to their own capacity limits. In nearly all cases, unless cluster analysis (discussed later) is the applied model, the *all process* method costs more to perform than buying additional hardware. In addition, the system must be in production to gather the information since it is usually impractical to estimate every process and its associated resource demands.

5. Modeling

Modeling computing systems has been and is the pursuit of many university professors. Our objective is to present the modeling techniques our group has successfully used when predicting capacity. I have no doubt additional models will be developed and used as they become available and are appropriate for our needs. As mentioned earlier, modeling can be separated into the broad categories of mathematical models and simulation models.

There a number of modeling techniques available today. However, for our purposes we only need to concern ourselves with six different modeling techniques. They is simple arithmetic, regression analysis, queuing theory analysis, cluster analysis, and simulation. Each of these modeling techniques requires different inputs and produces outputs with varying degrees of confidence. A detailed discussion on each of these techniques is out of scope for this paper. However, a few words on each technique is worth mentioning.

5.1 Basic Arithmetic Models

Many situations do not require complex models. This is typically because no queuing is involved or detailed statistics are not available and/or not required (E.g., Initial Performance Readiness Review). When this is the case, simple worksheets using simple *arithmetic models* can be built. For example, modeling memory usually only requires a simple spreadsheet.

5.2 Linear Regression Models

When queuing is not an issue, performing mathematical *regression analysis* can provide one with high confidence results. For example, suppose the number of disk i/o's per day is captured along with the number of orders processed per day over a period of sixty days (Figure 6). Regression analysis can be performed on these data points producing an equation which shows the relationship between disk i/o's and the number of orders processed.

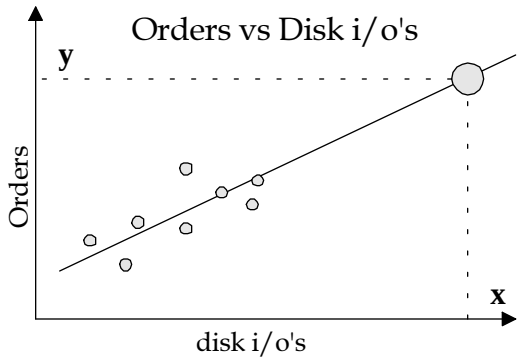

Figure 6. By gathering data points information one can perform regression analysis to predict capacity requirements or maximum throughput. In this example, we can predict that ***y*** orders processed will require ***x*** disk i/o's.

5.3 Queuing Models

Queuing (Figure 7) can become an issue when predicting CPU requirements and disk requirements. There are two basic modeling techniques that can be used when queuing is involved. There are various *simulation* models and *queuing theory* models. We have successfully used Monte Carlo simulation techniques implemented in BASIC and the

M/M/*m* queuing theory implemented in an Excel workbook. When the method is properly chosen, these modeling techniques can produce output at high levels of confidence with a minimum of effort.

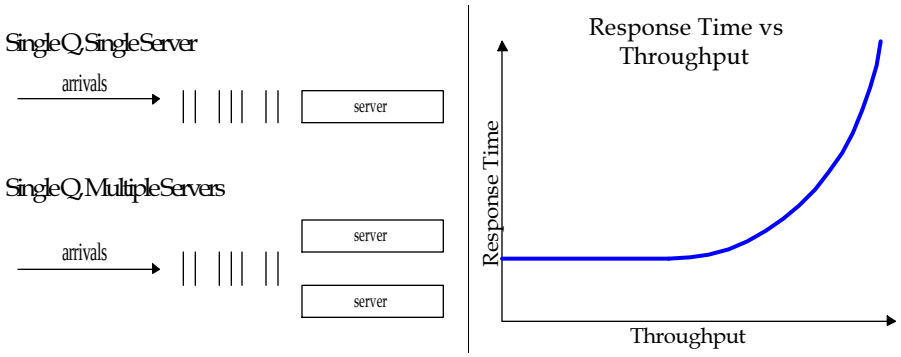

Figure 7. The figure on the left shows multiple requests arriving at different intervals to be serviced by one and two servers. This situation can be modeled by various simulation models and queuing theory models. The figure on the right is the classic response time versus throughput curve. As with all queuing systems, at some throughput queuing begins to rapidly increase response time.

5.4 Cluster Analysis

The *cluster analysis* model is a mathematical dream come true. Data is gathered for each and every process and then categorized into a handful of groups. For example, one group may characterize processes which consume on average 500 KB of RAM, five seconds of CPU time per minute, and read/write 750 KB of data from/to the I/O subsystem. Once the processes are grouped into their respective categories, simulation or queuing theory models will produce a very high confidence result. To perform cluster analysis in an acceptable time frame, very powerful models (E.g., queuing theory) and data gathering processes (process clusterization) must be developed.

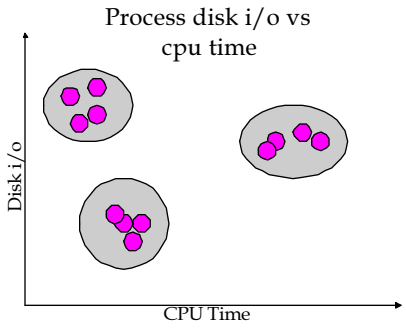

Figure 8. The cluster analysis modeling technique allows the "all process" method to be a viable. Data is gathered from each and every operating system process and grouped into categories by their common characteristics.

5.5 Simulation Models

Referring to our Direct Connect Devices case study, how would one organize a real user, full-blown performance assurance test? While it is possible, the realities of time, money, and logistics do not make it practical. We can, however, create fake users and have them perform their daily duties just like real users.

There are many simulation tools commercially available today. However, depending on your budget and the workload characterization complexity, you may be able to develop your own simulation tools. For example, the Oracle Application's regression tester can be used to simulate users. The Oracle Application's **aiap** program has the ability to save and replay keystrokes. However, there is a danger of using "home grown" simulation tools. The statistical gathering mechanism will have not passed the same rigorous QA process and the impact of the simulation tool will probably place a noticeable processing burden on the system. Since there are many simulation possibilities and not every simulated environment is the same, it is best to learn about your options and get hands-on experience.

5.6 Modeling Disk Space Capacity

This paper has been focused on throughput capacity not space capacity. However, predicting space capacity is important for budgetary and immediate hardware justification reasons. All the models presented in this paper can be used to predict space capacity. The difference is in where to gather the required statistics to begin your analysis.

While throughput related statistics can be gathered from the Oracle **v$sesstat, v$filestat, v$session** fixed tables, the **aud$** table, the UNIX accounting tools, the UNIX **ps** command, and several vendor specific tools, space capacity information is based around Oracle's segment related data dictionary views.

When predicting Oracle data capacity (i.e., physical space requirements), keep in mind, that other information management systems space requirements will not map exactly to Oracle's. Variable length fields and indexes are two common examples.

Fortunately, there are number of options to help gather space requirements. As you might suspect, the more precise the results, the longer it takes and the more difficult it is to retrieve the information. The **dba_segments** and **dba_extents** data dictionary views, **vsize** function, and the **analyze** command all provide varying levels of size granularity. Just as when deciding about *methods* and *models*, understand your precision requirements and then choose the appropriate data sizing method.

6. Satisfying Conflicting Constraints

Predicting computing system capacity does not leave us with a perfect solution. As with most decisions in life, concessions must be made.

Suppose your disk space predictions call for twenty disks and your disk I/O throughput predictions call for forty disks. Which one should you use? It's simple if you want to ensure all constraints are met. By implementing forty disks, both the disk space and disk I/O throughput requirements will be met.

A more difficult situation is when throughput can be achieved with six CPUs, response time requirements will not be met unless their are eight CPUs, and you only have the budget for a four CPU system. This is when skilled negotiations and creative minds must work together to meet user requirements. The solution could be as simple as implementing Oracle's hot backup facility versus a cold backup (which shuts down the Oracle system) to allow more batch processes to complete at night which were planned to be run during peak OLTP hours.

To summarize, predictions normally do not present perfect solutions and must be translated into a form so business decisions can be made. To construct a system to meet all the predicted and stated requirements, additional capacity may be required, or business and technical processes may have to be changed.

7. Conclusion

Thank you for taking the time to read this paper. We trust it has been a worthwhile endeavor. Our goal will have been reached if, in the future, you will better allocate and direct scarce capacity planning resources. By understanding how to structure the project, the three capacity planning study types, the importance and how data can be gathered, the variety and appropriateness of modeling techniques, and finally how to address conflicting constraints, we hope you will be able begin your own capacity planning studies or at least understand the importance of planning capacity.

8. Acknowledgments

This work was supported by Oracle Corporation and was conducted while the authors were working (and still are) for the Oracle Services' System Performance Group. A spe-

cial and sincere thanks to Dave Cook, Tom Corrado, Dominic Delmolino, Ellen Dudar, and Cary Millsap for their critiques and stimulating discussions.

9. About the Author

Craig Shallahamer is a Practice Manager of Oracle Services' System Performance Group. The team is responsible for building new tools and capabilities like the ones described in this paper for Oracle and its customers. The System Performance Group provides technical architecture services including capacity planning, and performance management services to customers worldwide.

Since joining Oracle in 1989, Mr. Shallahamer has worked at hundreds of clients around the world. His specialization is performance management and capacity planning related research and on-site engagements that have resulted in a number of published papers at EOUG, OAUG, IOUW, Asia Pacific OUG, and *Oracle Magazine*. Craig is based in Portland, Oregon and can be reached at 503/220-5122 or on the internet at **cshallah@us.oracle.com**. Craig's papers are available via the World-Wide Web at **http://www.europa.com/~orapub**.

BIBLIOGRAPHY

Jain, R. *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, 1991. ISBN 0-471-50336-3

Menascé, D.; Almeida, V.; Dowdy, L. *Capacity Planning and Performance Modeling*. PTR Prentice Hall, Englewood Cliffs NJ, 1994. ISBN 0-13-035494-5

Millsap, C. *Designing Your System To Meet Your Requirements*. Oracle Corporation White Paper, 1995. http://www.europa.com/~orapub

Chatterjee, S.; Price, B. *Regression Analysis by Example*. John Wiley & Sons, 1991. ISBN 0-471-88479-0

Levin, R.; Kirkpatrick C.; Rubin, D. *Quantitative Approaches to Management*. McGraw-Hill Book Company, 1982. ISBN 0-07-037436-8