Managing Workload Performance on a Private Cloud

A Case Study

Dr. Boris Zibitsker ▸ BEZNext ▸ 7/1/2012

Managing Workload Performance on a Private Cloud

A Case Study

Executive Overview

Cloud based IT infrastructures are rapidly being adopted as the answer to bloated IT costs and poor quality of service. The private cloud concept within a company's own data center is gaining traction as the answer to solving the complexity and cost created by the proliferation of individual and distinct IT infrastructures. The availability of new vendor engineered appliances from vendors such as Oracle, Teradata, IBM/Netezza and Greenplum are often used as the basis for hosting a private cloud infrastructure. Intended to host consolidated workloads, private cloud administrators face risk and uncertainty in these highly interdependent environments where any change to improve the performance of a specific workload can negatively affect the performance of other workloads sharing the private cloud.

In this paper we discuss the major factors affecting workload response time. We review how the use of analytic modeling and predictive analytics can optimize your strategic capacity management, tactical performance management, and operational workload management decision making. To gain a better understanding, we take you through a case study of a consolidation effort on an Oracle Exalogic/Oracle Exadata private cloud to show how modeling answers many of the key capacity management questions.

While workload consolidation on vendor engineered private clouds may offer the promise of faster time-to-value and lower total-cost-of-ownership (TCO), they also present some unique management challenges to insure that all who share the private cloud; consistently meet their response time objectives. Capacity management and predictive analytics enable organizations to set and maintain rational goals so that the promise of the private cloud and lower TCO can become a reality.

Private Cloud Architecture

Private cloud architectures vary in scope and intended usage. Some vendor engineered private clouds incorporate the middle tier application servers and database infrastructure while others are specifically database and storage infrastructures. They also reflect a particular vendor's background and focus. For example, IBM/Netezza and Teradata both come from a very specific decision support background, their appliances are intended for the read intensive workloads found in data marts and data warehouses that support analytic applications. In contrast, Oracle is a vendor that has a much broader approach, encompassing both decision support/analytic and transactional (OLTP) workloads.

All private cloud architectures must embody an attribute known as "elasticity" to enable the infrastructure to scale quickly and predictably. As new and existing application workloads are consolidated to the private cloud they each need CPU, I/O, memory and networking resources. To gain elasticity, private cloud infrastructures are built on clusters of servers. As new compute power, storage capacity or I/O is required; organizations simply add servers or storage (or both) to the cluster. Most vendor engineered appliances add all of these resources via upgrades to larger sized configurations. These larger configurations come at

the cost of hardware, software licenses and maintenance as well as power and cooling.

Private cloud infrastructures often consist of more than just the vendor supplied appliance. They are multi-tiered and virtualized environments that are complex in nature. They have to be because the business of managing any number of mixed workloads effectively and efficiently is complex, simultaneously requiring resource sharing and workload isolation.

One example of a vendor supplied private cloud infrastructure is Oracle's Exalogic Elastic Cloud X2-2, a middleware platform, and Oracle's Exadata Database Machine. Together they provide the building blocks for a scalable and elastic private cloud infrastructure supporting mixed workloads for consolidation and implementation of new applications. Both building blocks come in different sized configurations such as ¼ Rack, ½ Rack and Full Rack within a single cabinet. Multiple cabinets can be clustered via the appliance's InfiniBand based network.

Exalogic provides the application middleware cluster where application servers run application logic in a highly virtualized environment. Oracle Exadata hosts the relational database cluster utilizing Oracle's Real Application Cluster (RAC) technology.

The components, both hardware and software, used by the vendors in building their appliance has an impact on the process of capacity management. As we will discuss later, there are a number of major factors that affect a workload's response time. The ability to anticipate where a particular bottleneck might occur and what tactical performance options are available to the administrator to overcome them can be somewhat vendor specific.

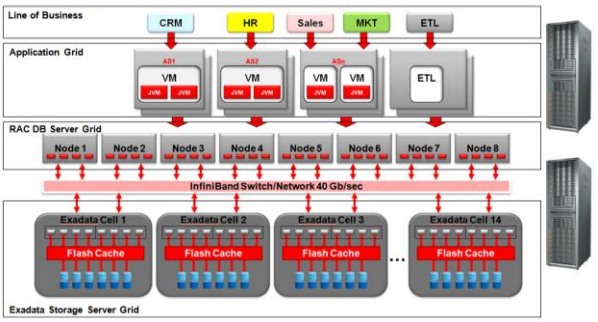


Figure 1: Oracle Private Cloud Architecture

Since Oracle's Exadata is a database machine, one typical bottleneck is I/O. Oracle RAC enables any node in the cluster to access all the data. Oracle Exadata overcomes I/O bottlenecks by employing intelligent storage servers depicted in Figure 1. Each Oracle Exadata storage cell is an actual server with two, six-core processors, 24 GB of memory running Oracle Unbreakable Linux and hosting 12 disks and 384 GB of flash storage. These storage servers actually run database code, pushing SQL predicate processing to the storage level, this is important because it eliminates much of the I/O bottlenecks found in typical database infrastructures. Software automatically places the most accessed data onto the flash storage (typically reserved for transactional data) to further eliminate bottlenecks. Finally, each storage cell works like a shared-nothing server, processing only its snippet of data. Each time another Oracle Exadata storage cell is added, more processing, more storage capacity and more network bandwidth is added as well. Theoretically the Oracle private cloud should scale in a near linear fashion but with so many competing workloads there are always exceptions.

Each vendor supplied appliance attempts to avoid the bottlenecks to workload response time. They may use similar approaches but in different ways. Some may employ faster processors or larger memory. Some may use faster network interconnects or solid state drives. All attempt to tailor their software to make optimal use of the hardware infrastructure. All are complex systems and a change of any kind impacts all workloads running on the infrastructure. This is a risky proposition on a shared environment. IT organizations need to justify any changes they make in relation, not just to a single workload, but to the impact of all workloads on the private cloud. As the trend toward cloud based (i.e. rationalized)

infrastructures accelerate, so too will the consolidation of mixed workloads. This is why capacity management has become a required process.

Mixed Workload Management

In a shared private cloud infrastructure, perhaps the most important responsibility an administrator has is workload management. This is the task of allocating the resources of the private cloud to individual workloads. This task is difficult because it involves both a subjective business perception of the importance of a workload and an objective technical response time requirement. The real talent lies in the negotiation between the line-of-business desire for sub-second response time and the IT organization's judgment as to the potential capacity of the private cloud to deliver the desired response time. For the private cloud to be successful, each side (line-of-business and IT) must be working from realistic expectations of what the workload response time can be versus what they would like it to be.

Workload management on a shared private cloud must be a collaborative process. Both sides need a reality check of their available options and alternatives for optimizing workload management. When viewed independently, every line-of-business believes its particular application is the most important one. When considered from a higher perspective (the private cloud), both the line-of-business and IT can better evaluate the relative importance of a single workload. Response time also requires negotiation. If IT can show the line-of-business what the impact (cost) of delivering sub-second response time to their workload they may be happy to accept a three second response time.

Vendor based appliances provide tools that help manage resource allocation among competing workloads on the private cloud. For mixed workload management, Oracle offers CPU Resource Manager, I/O Resource Manager, Parallel Statement Queuing and Runaway Query Management [15].

Oracle Database Resource Manager allocates resources between groups of users (workloads) using consumer groups (group of sessions); it includes a resource plan, that represents a scheme for sharing CPU resources, number of active sessions (Concurrency), degree of parallelism (DOP) and session termination instructions within an Oracle database Instance. It also includes directives on sharing and limits resources between consumer groups/workloads. Administrators can create consumer groups on a per application/workload basis if they choose or group workloads with similar profiles into a single consumer group.

For example a resource plan can allocate 40% of the available CPU resources for transactional workloads (consumer group), 30% for BI Data Warehouse (read-mostly) workloads and 20% for extract, transform, and load (ETL) workloads and the remaining 10% of CPU resources for workloads that are not grouped. Several resource plans can be defined and scheduled to be implemented at different times of the day, different days of the week or perhaps different times of the year such as quarter end or a holiday season when certain workloads take precedence over others.

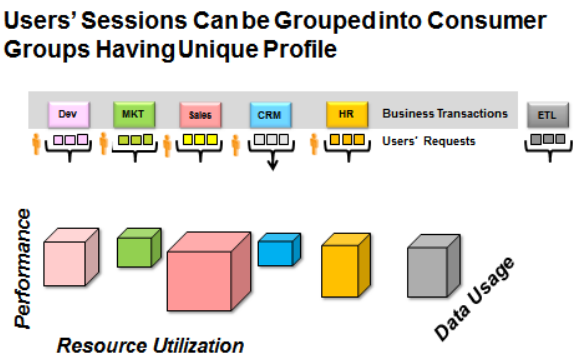


Figure 2: Workload Characterization

All of the appliance vendors offer resource management. One of the goals of workload management is to protect the most important

workloads by reserving enough of the private cloud's resources to maintain the SLO of those important applications. Another goal is to achieve a consistent user experience.

The most important component of a workload's SLO is its response time. The experience a user has is partially dependent upon a consistent response time. If a user experiences sub-second response time early in the morning and thirty second response times after noon, they will adjust their usage pattern to achieve the best response time. If all the users start logging on early to get better response time they could jeopardize the very stability of the infrastructure. It is for this reason that the goal should be to achieve consistency. It would be better if a user always received five second response time than to experience sub-second at some times and one minute response times at others. Consistency equals an improved user experience.

Cloud Response Time Components

The response time of a workload depends on many factors. As we discussed previously, the appliance's architecture and components all impact response time, any capacity management process would have to account for a systems specific architecture.

Figure 3 below illustrates the major components of the typical request in an Oracle private cloud environment. OS statistics and Oracle's Enterprise Manager (OEM) repository contain measurement data characterizing the usage of resources and response times for individual SQL users. OEM also collects data showing applications per node and information about I/O performance and resource utilization within each Exadata storage cell. This data can be summarized by workload, representing the activity generated by a group of users and applications to support each line-of-business. As a result of the workload aggregation and characterization you can see a performance, resource utilization and data usage profile for each workload [6,7]. As part of the capacity management process a predictive model of the Oracle private cloud could be built using the results of this workload characterization to answer different "what if" questions and predict how expected growth and changes will affect the individual components of the response time for each workload [s 1,2,4,9].

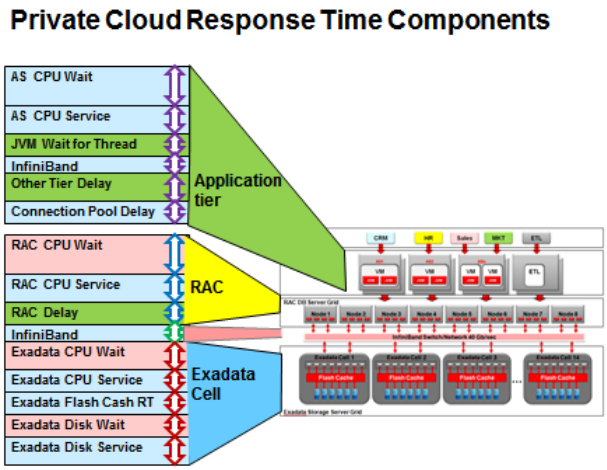


Figure 3: Response Time Components in Oracle Exalogic/Exadata Private Cloud

The Role of Predictive Modeling in Capacity Management Decisions

Response time can basically be represented by the following equation:

Response Time = Service Time + Queuing Time + Delay Time

Service time depends on the complexity of a request and the speed of the hardware. For example, an InfiniBand interconnect is four times faster than a 10GbE (Ethernet) interconnect. Disk based storage will be slower than solid state drives however workload type always matters. A hardware upgrade or simply tuning the DBMS can reduce service time.

Queuing time depends on contention for resources resulting from any type of workload growth such as an increase in volume of data or number of users. A simple software parameter change or tuning a workload's SQL can reduce usage of CPU and reduce queuing time. Changes in workload profile or in software/hardware configuration may impact the queuing time of the individual workloads differently.

Delay time is dependent on several factors. One of them is software parameters that control the level of concurrency. For example, a change in the number of JVM threads can affect delay time as a request will have a shorter wait for available threads or a connection to the DBMS.

There are a number of factors that affect the three major components of response time. Predictive analytics can be used to try and anticipate how these factors will impact workload response times. In this way, the role of predictive analytics as it relates to private cloud capacity management is essentially the same role it plays on the business side. Business analysts build mathematical models of the business. They base the model on available business performance data to represent the current business operation. They then input changes to the model's parameters such as increasing prices by 5% or building 20 new stores. The model shows the business analyst how those changes might impact the business (revenues, profits) if their assumptions become reality. Business leaders use this type of predictive analysis to justify decisions on things such as inventory levels, pricing, expansion, employment levels and more. IT can utilize the same approach to anticipate how changes to workload profiles, software parameters or hardware upgrades will affect the goal of meeting the response time objectives of the workloads running in the private cloud.

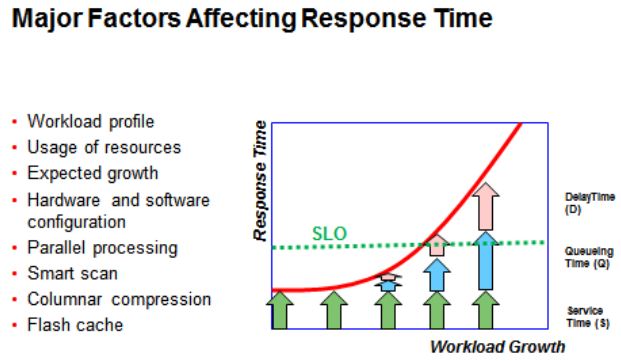


Figure 4: Major Components Affecting Response Time

The process of capacity management can use predictive analytics. The first step is to build a mathematical model that represents the private cloud infrastructure. Each vendor's appliance offers unique capabilities in either hardware of software. The model must represent those capabilities. Figure 4 illustrates some of the unique capabilities offered on an Oracle "Exa" infrastructure that influences components of response time. The goal of the model and using it to apply predictive analytics is to anticipate when the workload will no longer meet its response time objective with the current configuration. From there, different strategies can be tested via the model to see how best to proceed. Using predictive analytics, any number of "what-if" scenarios can be tested rapidly and inexpensively.

Many IT organizations have eschewed the use of modeling and predictive analytics. The reason often stated is that you cannot guarantee 100% accuracy in the model. Some have even gone as far as to say, "Why bother with data collection and workload forecasting when hardware is cheap?"

It is true that a benchmark will be more accurate than a model. It will also be significantly more expensive and time consuming making it cost prohibitive in justifying every possible scenario. Remember that in the private cloud where many workloads have been consolidated, the key attribute is elasticity. This does not apply only to scaling out the physical infrastructure; it implies rapid (and transparent as possible) adaption to change. How can you manage change if you don't know what to expect?

When the business uses analytic models in justifying decisions, they do so knowing that their assumptions and therefore their predicted outcomes will not be 100% accurate. What they have learned is that when faced with a destination/goal it's not as important to know the exact distance to that goal but rather which direction/path is the shortest. Over time, the accuracy of the model improves as well as the accuracy of assumptions and the data collection.

The issue is that some workloads are I/O bound while others are CPU intensive. Some require sub-second response time while others are fine waiting minutes for a response. Many workloads have peak periods or different profiles during different times of the week or month. As discussed earlier, appliance vendors supply workload management capabilities. It may be relatively easy to create or change workload management rules, what is difficult is picking the correct rules or changing the correct software parameters that will satisfy the SLOs of all the workloads on the private cloud. This is where the role of modeling and predictive analytics becomes clear.

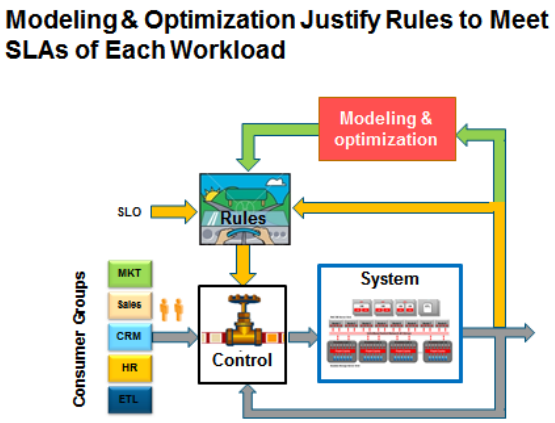


Figure 5: In House IT people analyze and use measurement data (yellow line) to set up rules and policies controlling performance of the system (grey line). The role of modeling and optimization (green line) is to find optimum operational Workload Management, tactical Performance Management and strategic Capacity Planning decisions.

Capacity management is expected to be an ongoing process of modeling, testing assumptions, getting predictions, validating predictions and refinement of the model and assumptions. Let us examine a specific case study to illustrate the power of predictive analytics when managing the private cloud.

Case Study

Let's use a case study to demonstrate the value of using a predictive analytics to justify various management decisions for a private cloud infrastructure. In this example we will use a Oracle Exalogic /Oracle Exadata based infrastructure running a mix of workload profiles.

Different modeling tools that incorporate *queuing network modeling* technology can be used to model Exalogic and Exadata. A model is an abstraction of the physical computing infrastructure. A queuing network model represents the physical computing infrastructure as *a network of queues* that can be evaluated analytically. Essentially a queue represents a component of the physical system where users or transactions that makeup a workload might wait for resources. Simplistically, all transaction time is made up of either time *waiting for service* and *time being serviced*. Using existing monitors that most IT organizations have acquired over the years, data can be gathered to provide information like average CPU time, average response time per user, average number of concurrently active sessions to name a few. This becomes the basis for parameters of the model. Since modern compute infrastructures are complex, modeling software manages any number of parameters that can be manipulated. Since different workloads have different characteristics, the model supports multiple classes of users or transactions. Figure 6 illustrates a basic queuing network model.

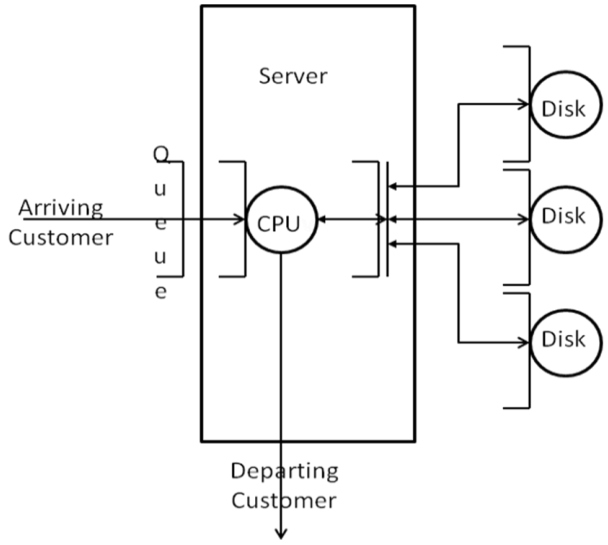


Figure 6: Example Illustration of a Queuing Network Model

In this case study we review several examples addressed using modeling software developed by BEZNext capacity management experts [14] to illustrate how predictive analytics can be used to justify management decisions and support SLOs of different workloads running on the private cloud. Below are some of the questions that can be asked and answered via modeling and predictive analytics.

1. What will be the impact of the expected growth and planned changes?
2. How to set realistic SLO?
3. How to change workload's priority to meet SLOs?
4. How to set workloads' concurrency level to met SLOs?
5. How to justify tuning measures to meet SLOs?
6. How to predict new application implementation impact?
7. What is the minimal hardware upgrade required to meet SLOs?
8. How to compare actual performance with expected?

Let's examine some examples of how modeling and predictive analytics works in a practical way.

What will the impact of expected growth be?

If successful, the private cloud usage will increase over time. New applications will be added, existing applications will gain additional users and all applications will add new data. How well the IT organization manages this expected growth is dependent on their ability to anticipate when the expected growth will impact the SLOs of individual workloads.

The first step in predicting the impact of expected growth is the need to build the analytic model of the private cloud. Measurement data of the existing system is critical to this step.

Workload characterization is a key input for building the model because it shows us how each workload utilizes the various resources in the private cloud and therefore what each one contributes to the overall system workload. This step is crucial to workload management as it can illustrate opportunities for trimming waste or rescheduling workloads to insure all workloads get the resources they need. Most importantly, it begins the collaborative process of capacity management because it can show the business how their business processes actually use their computing resources.

Figure 7 below shows the workload characterization output from our case study.

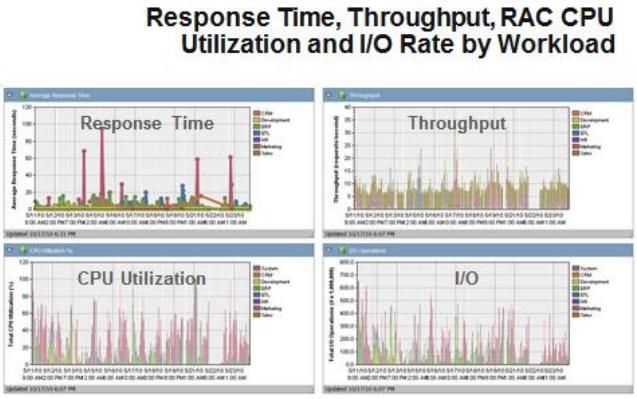
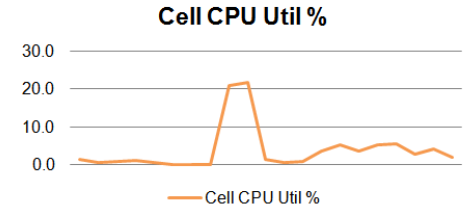
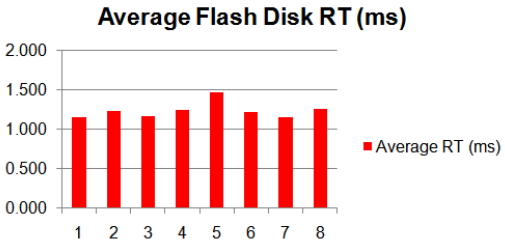


Figure 7: Example workload consolidation on Oracle private cloud

One unique aspect of Oracle Exadata is its intelligent storage subsystem. We capture performance information at this tier as well because it is crucial in building an analytic model that it represents the physical infrastructure as accurately as possible. Each Oracle Exadata storage cell contains CPU, flash storage and disk storage. The following graphs depict the performance of each of these components.





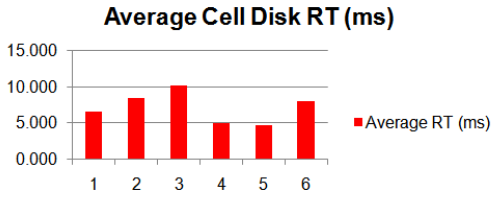


Figure 8: Exadata storage cell performance

The information in figure 8 helps us to understand how the components of the Oracle Exadata storage cells work. In this case study we found that the average storage cell CPU utilization was very low (5-20%) and that the flash storage read time averaged 5-10 times faster than the hard disk performance.

So performance data can for the basis for the parameters of the model. As we change parameters such as more users, the model mathematically calculates the impact. This step is where we ascertain what the assumptions for future growth might be. We call this workload forecasting and in this step we document what the expected growth will be in user activity and/or data growth. Typically this is based on historical trending along with some estimation based on business plans.

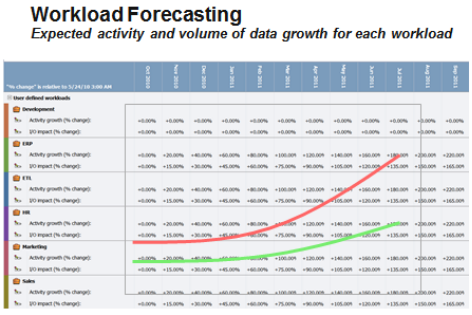


Figure 9: Forecast Spreadsheet

Workload forecasting is a crucial step in the success of capacity management and the use of predictive analytics. While workload characterization provided business stake-holders and IT staff a view into how the existing workloads use the infrastructure,

agreement on future growth is just as important. Many companies may have little experience in workload forecasting especially if capacity management has not historically been a process that has been followed. Over time, assumptions will improve as both IT and the business stake-holders iterate through the capacity management process several times. Each time they learn more about their business, their workloads and their private cloud infrastructure capabilities. Assumptions improve via a collaborative effort and so too will the accuracy of the predictions produced by the model. By extension, the quality of the management decisions improves and the risk of change is reduced as IT is better able to justify decisions that deliver a consistent user experience.

The goal of capacity management and the use of predictive analytics is to meet SLOs. Documenting service levels for each individual workload is an important step. Just as documenting assumptions in a collaborative manner was important, documenting service levels and reaching agreement is just as crucial. After all, we cannot measure success or failure unless we have all agreed upon what success is.

Typically service levels are negotiated between the business owners of the workload and IT. Often we find that organizations may have no formal service levels in place. In those situations we recommend that the workload characterization data be used as the basis for determining success. This can form the basis for informal success criteria. For example, we may agree that despite expected growth, the maximum degradation of response time of any workload should not exceed two times and throughput should not be less than 20% of current levels. Getting this success criteria agreed upon by both the business user and IT is crucial to justifying future decisions based on the model's predictions as well as determining success overall when we compare assumptions, goals and predictions to actual outcomes at the end of a capacity management iteration.

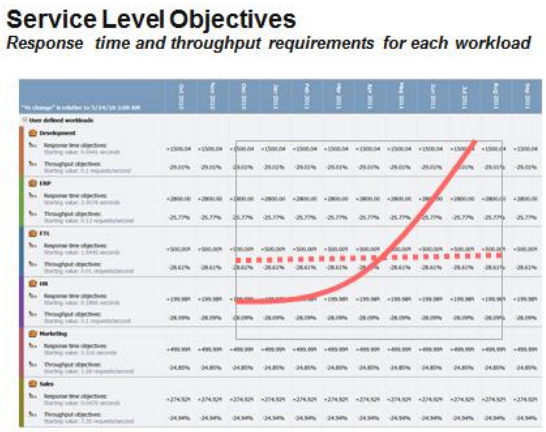


Figure 10: SLOs (Response Time and Throughput) Documented

Now we are ready to run our growth assumptions through the model to see how the expected growth will impact the performance of each of the workloads on the private cloud. Since each workload has a different profile, each will have a different sensitivity to the expected growth.

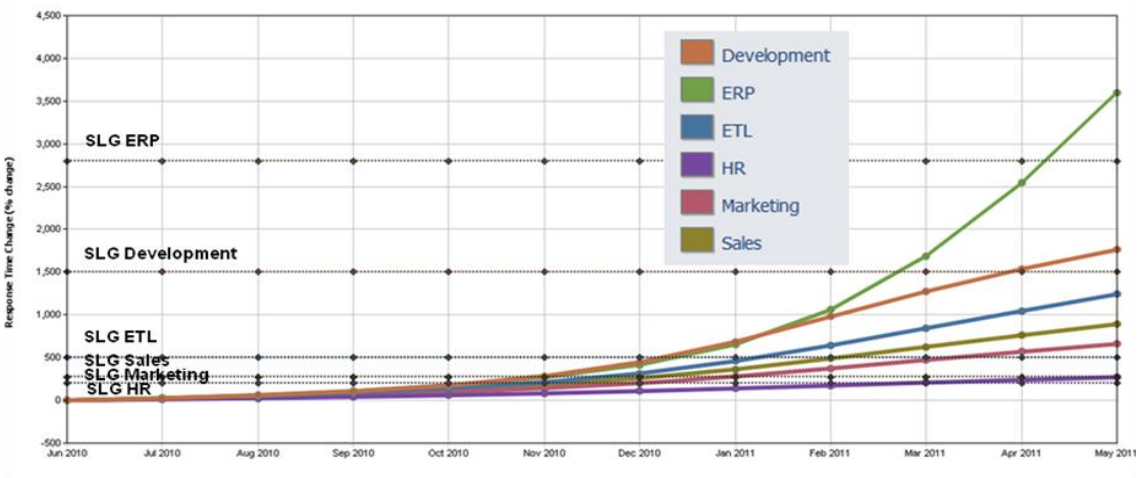


Figure 11: Predicted impact of workload and data growth on individual workload performance. Response time includes service time, queuing time and delay time at every Oracle Exalogic and Oracle Exadata tier

Predicted response time for each workload is compared with the corresponding SLO to determine when the private cloud will no longer meet the expected response time objective. Here we identify individual workloads that will be most impacted by the change and then evaluate options to proactively change workload management, identify performance tuning needs and any potential hardware upgrades that may be required.

Figure 12: The Sales workload will not meet its SLO and CPU wait will be the largest component of Response Time

According to the performance prediction results the Sales workload will be the first that will not meet its response time objectives. This will happen around mid-year. We predict that the main culprit will be CPU wait time on the Oracle Exadata machine's RAC tier. Knowing this provides us with our first clue as to what can be done to avoid this future degradation of performance. If we can identify which workloads will

utilize the most CPU resources on the Oracle Exadata RAC tier after our expected growth we can begin to formulate a plan to fix the problem.

According to the modeling results, ERP and Marketing will be using the maximum amount of CPU resources.

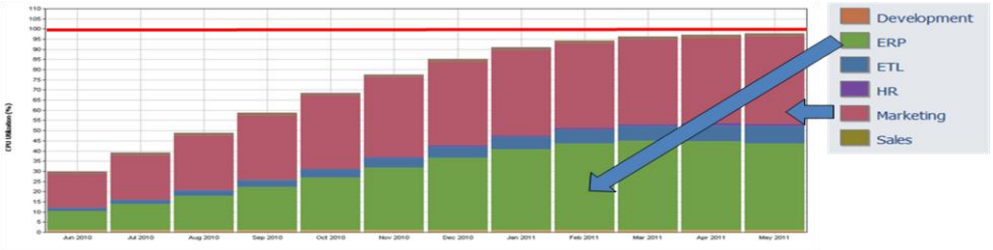


Figure 13: Marketing and ERP workloads will use the majority of CPU resources

Now that we know that the Sales workload will have a problem and that ERP and Marketing will cause the problem we have several options available to us related to workload and performance tuning.

- Increase the priority of the Sales workload
- Reduce the concurrency of the ERP and Marketing workloads
- Tune the Marketing and ERP workloads
- Upgrade/Add hardware

Each of these options can in turn be model evaluated by the model to ascertain the effectiveness and unintended consequences of each individual option on other workloads. In this way, we can fully justify our best course of action.

Are Service Levels Realistic?

One mistake IT organizations make in the capacity management process is to focus on solving the problem of meeting service level objectives. This sounds counter intuitive but the important first question to answer is not "How can I meet the service levels of my customer's workloads?" Instead, the first question should be "Are the expected workload service levels realistic?"

The private cloud is a shared, finite resource. It is constrained by the technical limitations of the appliance's architecture and that of the individual components that go into building the appliance. It is constrained by the ability of the IT personnel to effectively manage it. Of course, the private cloud is constrained by the company's budget, the dimensions of the data center and the availability of cheap and reliable power among other things. This realization is important because we must first view IT from the perspective of its impact on the business. If the service levels are too "relaxed", it negatively impacts the number of business transactions that can be completed which directly impacts the company's bottom line. If the service level objectives are too aggressive, your IT costs will rise significantly and that will impact the company's bottom line. So the real focus of the IT organization should be to balance these two realities. Modeling results can be used to organize a collaborative effort between the business consumers of IT services and the IT personnel that provide those services. This collaboration provides the basis for developing and maintaining an IT infrastructure that regulates incoming work based on the business's priorities and nothing else. With that in mind, IT can manage the private cloud's capacity as efficiently as possible.

Figure 14: Performance prediction results provide an opportunity to organize a collaborative process of evaluating business demand and configuration required to support business needs. SLO affect price/performance. Aggressive SLOs are expensive to meet, but relaxed SLOs can negatively affect Business

How will changes in workload priority affect performance?

Previously our modeling predictions had indicated that the Sales workload would experience performance degradation due to the growth in user activity and data volume. We identified the potential bottleneck as being CPU wait and we identified potential options we could pursue to alleviate this issue. One of those options was to change the priority of the Sales workload so that it would not have to wait for CPU resources being consumed by the Marketing and ERP workloads.

It's important to model our options because as we have stated throughout, the private cloud is a shared infrastructure which means that any change we make to improve the performance of one workload (Sales) may have a negative impact on other workloads. This ability to continuously iterate through the model and use predictive analytics to quickly see potential outcomes of our management decisions is what makes this approach so compelling.

Unfortunately in this case study the option of changing the priority of the Sales workload is predicted to help the Sales workload but that change will negatively impact the performance of other workloads. The model considers that the proposed change must be evaluated in context to all the other workloads. In this case the CPU wait times will lengthen for other workloads after changing Sales' priority. The goal is to find an option that will enable all the workloads to continue to meet their performance objectives. This option is not predicted to meet that goal.

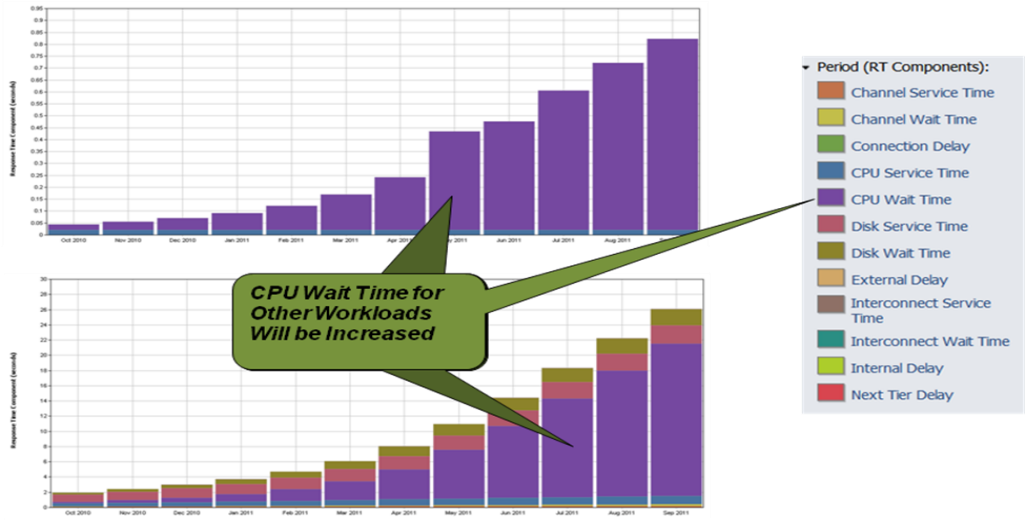


Figure 15: Priority for Sales Will Elongate Response Time for Other Workloads

How will a change concurrency level affect performance?

Another option for improving the performance of the Sales workload is to adjust the concurrency levels of other workloads. By doing this we could reduce the resource contention and hopefully reduce the wait-time for the Sales workload. The trick is to make the changes without forcing those workloads to miss their performance objectives.

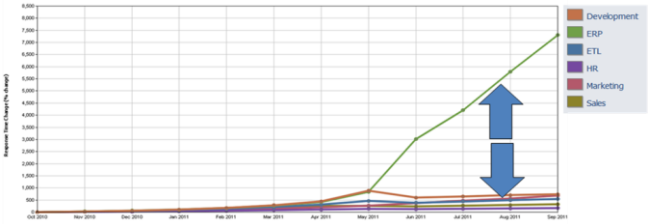


Figure 16: Reducing the concurrency for ERP will have a positive impact on Sales workload performance

Modeling and predictive analytics is once again used to prove the efficacy of this tuning option. As with changing workload priority, limiting concurrency on one or more workloads can have a very different impact on all other workloads. For example, if we reduce the number of JVM threads in the application server for one workload it will limit the resource consumption by that workload, but increase consumption of resources by other workloads using a different JVM. The adjustment may also move the bottleneck from the application server to the DBMS server. In our case study we focused on the ERP workload since it had been identified as one of the two largest consumers of CPU time. We wanted to model what the impact would be of changing the concurrency level of ERP and therefore its consumption of CPU resources on the Sales workload.

By limiting the concurrency rate of the ERP application we reduced its overall CPU consumption. This change did improve the performance of the Sales workload and several other workloads. While the average response time for ERP users did increase, the end user response time actually became more consistent, improving the average user experience. Of course determining what the optimal level of concurrency should be for the ERP application is a complex problem and almost impossible to get correct by manually making the change and then observing it impact. This type of optimization problem is well suited for the use of modeling and predictive analytics. We can easily evaluate all combinations of changing workload priorities and concurrency levels until we find the proper set of parameters that will enable all workloads to meet their SLOs.

What is the minimal hardware upgrade required to support existing SLOs?

Every company would like to avoid a hardware upgrade for as long as possible. The total cost of a hardware upgrade when you factor in man hours, hardware, software licenses, power, and floor space adds up quickly. For many companies, the biggest concern is the risk inherent in disturbing a stable system.

Sometimes a hardware upgrade is required. When changing workloads priorities and concurrency levels still does not satisfy the SLOs of every workload, adding more physical resources is our only option. Through the use of modeling we can predict when the upgrade will be required and how much more capacity will be needed. This allows management to budget for this requirement and gives the IT administrative staff ample time to plan for the upgrade. The risk level is even higher when the upgrade is being made to a shared infrastructure like the private cloud, exposing many more business processes to outages, failures or performance issues than in a standalone infrastructure. Having strong advanced notice of the upgrade need is a powerful risk reduction tool for any company.

Previously we identified CPU wait time within the Oracle Exadata RAC tier as the primary bottleneck. At some point it is determined that workload user activity and data growth will overwhelm the current Oracle Exadata infrastructure regardless of other changes. A hardware upgrade to Oracle Exadata is required but should the company move to the ½ rack or the full rack configuration?

The use of modeling enabled the company to evaluate all of the workload priorities and concurrency levels along with the impact of a hardware upgrade. It was determined that the hardware upgrade would be needed by August to continue to maintain all of the individual workload's SLOs. It was determined that the company wanted to limit the need for another hardware upgrade to be no sooner than once annually. This enabled them the balance their desired for reduced risk resulting from such a big change without wasting resources sitting idle.

When assumptions were extrapolated out twelve months, the model predicted that an upgrade to the Oracle Exadata full rack, which added an additional four RAC nodes and seven more Oracle Exadata storage cells would be more than sufficient to meet the needs of all the workloads for the next twelve months.

Figure 17: Predictions justify the upgrade to a full Exadata rack

Conclusion: How do you compare predictions with actual results?

Modeling and the use of predictive analytics can form the basis for organizing a continuous, proactive performance management process. It should be expected that there will be divergence between predicted performance results and actual results. This doesn't mean that the modeling effort was a failure, indeed it is the very basis for determining the success of the capacity management process.

Without modeling and prediction the notion of the success of a project is subjective. Some lines-of-business might be happy while others were not. The problem is that there existed no quantitative and objective basis for their opinions. The capacity management process's use of modeling and predictive analytics invites collaboration between IT and the line-of-business stakeholders. Both participated in developing and approving assumptions as well as reviewing and approving which options to pursue. When divergence between predicted and actual results occurs it becomes another opportunity for both groups to collaborate.

Armed with the information gathered during modeling, the assumptions agreed upon, and the actual results, both sides can concentrate on root-cause analysis. We can begin to answer why the difference occurred. From this we can improve the accuracy of the model and/or the assumptions so that the next iteration is more accurate than the prior one. We can suggest new options for workload management and performance tuning.

Cloud based infrastructures and the consolidation of workloads they enable; place new management challenges on the IT organization. While many have talked about the need for IT and business stakeholders to be in alignment, consolidated platforms like the private cloud make this notion of collaboration an imperative. We have discussed how modeling and predictive analytics can help foster this needed collaboration as part of a continuous capacity management process. We have illustrated examples in the case study that show how predictive analytics can be a powerful tool for companies to reduce the risk of future change at lower cost. We have demonstrated the need to evaluate every possible outcome quickly as important given the interdependence of each individual workload running on the private cloud. When operating in such an environment, how can any IT organization possible proceed with any change if they don't know what to expect? Finally, if you don't have agreed upon expectations to compare with actual results, how do you determine the success of any change? It's time to manage the IT infrastructure and not allow it to manage us.