The Secret to Maximizing Virtualization ROI in the Dynamic Data Center

If IT has a requirement to maximize infrastructure (maximizing use of resources) while the business requires an assurance of performance and sufficient capacity, how can business-critical applications be successfully virtualized? Akorri's BalancePoint™ software provides unique cross-domain analytics and IT Key Performance Indicators (KPIs) that enable virtual infrastructure optimization, or the optimal balance between infrastructure utilization and delivered performance at the lowest cost. This paper reviews these analytics and how to apply them to get the most value out of your virtual infrastructure.

Advancing IT Operations Management

By consolidating low priority applications onto shared resources, virtualization has enabled IT organizations to cut costs and improve efficiency, leverage new IT investment, and improve service.

But as virtualization maturity advances and business-critical applications are virtualized, business application owners become increasingly concerned about assuring continuous performance and future capacity. Many IT organizations lack sufficient capability to optimally manage or even measure their dynamic shared infrastructure to a level that would encourage their business counterparts to trustingly accept virtualization of their critical applications.

If IT has a requirement to maximize infrastructure (maximizing use of resources) while the business requires an assurance of performance and sufficient capacity, how can higher priority applications be successfully virtualized? Built on agentless data collection and delivered on a virtual appliance, *Akorri's BalancePoint® software* provides unique **cross-domain analytics**, metrics, and **Key Performance Indicators** (KPIs) that enable virtual infrastructure optimization, or the optimal balance between infrastructure utilization and delivered performance at the lowest possible cost. This paper reviews these analytics and how to apply them to get the most value out of your virtual infrastructure.

Isn't Virtualization Self-managing?

Server virtualization vendors have marketed their technologies as if they would automatically correct themselves and optimally "self load-balance" to assure great service levels to client applications. But infrastructure performance doesn't depend on virtual servers alone. Total performance depends on end-to-end "cross-domain" resource capacities, configurations and competition (e.g. application, servers, storage, and network). Domain-centric "dynamic resource" features fall short of providing the kind of service assurance that critical business applications require.

To truly achieve the promised ROI of virtualization requires the ability to actively manage the important performance and capacity opportunities provided. While hypervisor managers come with many "knobs and switches", they don't come with the enterprise knowledge or cross-domain visibility that would help optimally and dynamically set those knobs and switches. In fact, worse than not getting the full value out of virtualization is the risk that naive configurations and default policies might actually prove counter-productive.

Optimizing virtualized IT infrastructure requires a new type of analysis that accounts for not only cross-domain components and contention, but also models the impact of resource sharing "entitlement" settings, logical resource pool membership, dynamic re-assignment (migration policies) across clusters of physical resources, and non-linear system performance curves under shared workloads. These data center level metrics are also required to assure optimal performance of the virtual infrastructure before migrating to the private cloud.

Benefits of Cross-domain Analysis

Cross-domain analytics enable IT operations to manage virtualized server and storage infrastructure as a coherent *system*. *System management* is literally evolving as the concept of what a "system" means grows with newly available technologies. In distributed systems, system management initially addressed only physical servers. Then servers with centralized storage. With virtualization, the system became virtualized machines, physical server hosts, virtualized storage, and SANs. IT today requires implementing solutions and processes that enable effective and efficient systems management at the cluster and resource pool of virtual servers.

Cross-domain analytics allow IT to see across silos as well as drill down into elements to analyze and model the interactions between what were previously isolated technology domains in order to:

- Visualize End to End Infrastructure and Troubleshoot Performance Problems
- Optimize Performance and Utilization
- Plan Virtual Infrastructure Capacity
- Manage Service Levels – Meet Business Requirements

IT investment decisions and project evaluations can be made with certainty (no more guesswork), and technical initiatives and efficiency objectives can be measured for success. Organizations can now manage IT more as a business with intelligent, automatic KPIs that can justify, measure and validate IT initiatives like consolidating servers, centralizing storage, or technology upgrades. They can also measure, manage, and report on key virtual service metrics that matter to their application customers, enabling proactive assurance and avoiding unproductive finger-pointing.

Analytics for Dynamic Data Center Management

Using cross-domain analysis, BalancePoint produces a number of unique metrics and reports that provide guidance for IT operations.

Cross-Domain Analytics	Guidance	How to Use
Application Contention Analysis	Shows which applications are competing for resources	**Troubleshoot** – quickly identify bottlenecks, contention
Infrastructure Response Time	Show performance delivered to an app by the total resources assigned to it	**Troubleshoot** – Measure and baseline an application's system Performance
VM CPU Efficiency	Compares guest O/S perspective with actual VM usage	**Troubleshoot** – determine optimal CPU efficiency for each VM
Performance Index	Scores balance between app requirements and infrastructure's ability to deliver *from a physical server perspective*	**Optimize** – confidently determine how many VMs fit on an ESX server
VM Performance Index	Scores balance between app requirements and infrastructure's ability to deliver *from a virtual machine perspective*	**Optimize** – determine optimal size of the VM for each app
VM Resource Entitlement Analysis	Shows actual usage vs. allocation	**Optimize** – enables effective use of VM resource settings; avoid over- or under- allocating critical resources
Virtual Resource Pool Entitlement Analysis	Shows actual usage vs. allocation	**Optimize** – enables effective use of resource pools to manage shared infrastructure
VM Host Resource Contention	Shows CPU and memory in one view	**Plan** - shows resource hogs per VM at any point in time; identify available hosting; reclaim unused resources
VMware Cluster Capacity	Shows cluster usage by server or resource pool	**Plan** – Load balance clusters and resource pools, plan total cluster capacity v. utilization with remaining headroom
Abnormality Analysis	Dynamically thresholds performance against "normal" behavior	**Manage** – predict, alert, and report on performance and capacity
Data Center Reporting	Produce infrastructure scorecards, capacity and performance summaries and forecasts	**Manage** – Prove service level adherence, generate chargeback information

Visualize End to End Infrastructure and Troubleshoot Performance Problems

Application Contention Analysis – Identify and Resolve Bottlenecks and Contention

BalancePoint automatically and agentlessly collects and analyzes performance data from an application's IT infrastructure including virtual and physical servers, SAN's, and storage. A logical data path topology is constructed automatically that includes all the infrastructure resources that each application uses and shares with other applications. Then dynamic performance analysis determines if there are any hotspots or bottlenecks and if so which applications and to what extent each might be contending for critical resources.

BalancePoint provides a fully navigable visual topology that is color-coded red/yellow/green to quickly find and resolve any cross-domain resource hotspot and contention issues. Because BalancePoint analyzes across IT domains, it can find and analyze deeply buried contention that domain-centric tools simply cannot see.

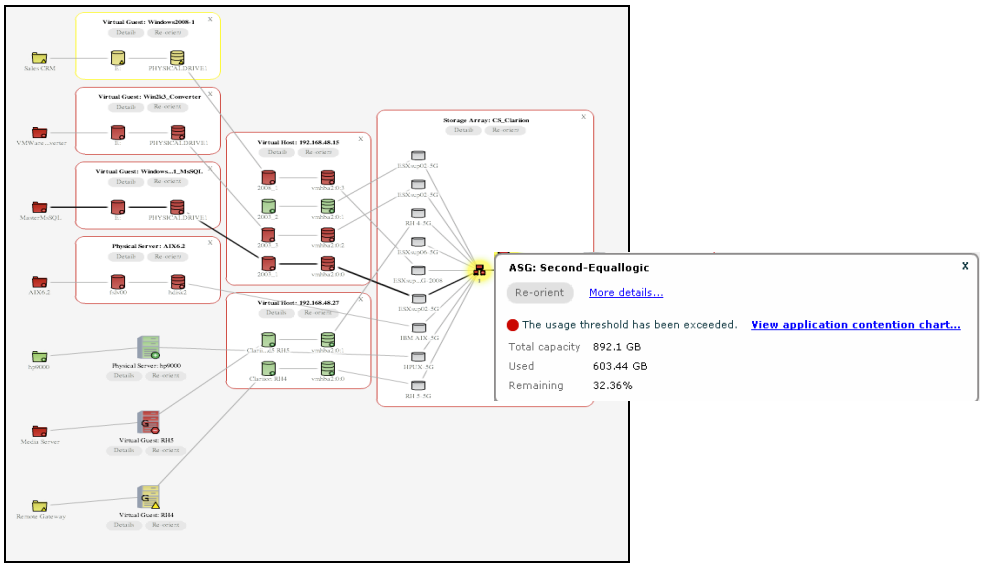

Figure 1 – See potential contention spots across virtual servers, host servers, SAN and storage then double-click to drill down and view the application contention analysis

In addition, BalancePoint generates alerts with informative analysis text and supporting tables that provide an operator with detailed information on resource performance issues, enabling quick and often proactive troubleshooting and remediation.

Infrastructure Response Time (IRT) - Measure and Model an Application's System Performance

Defined as the time it takes for the system to perform work submitted by an application, IRT directly shows the system's infrastructure responsiveness. It is a component of the application end-user's response time that includes IRT along with application code and UI specific latencies, remote network and desktop delays, and other factors outside of the IT infrastructure. For business reporting, IRT can be statistically correlated to end-user business application transaction volumes.

IT shops with responsibilities for *end-user application transaction monitoring* often utilize application performance management (APM) and other user-side *service assurance* technologies like synthetic transactions, application-derived metrics, logfile post-processing, and instrumentation of device drivers, middleware, or web page code. While it is important to assure actual end-user performance, these technologies treat the infrastructure as a "black-box" meaning they do not provide any insight into troubleshooting resource contention, infrastructure optimization, or planning future capacity. These solutions

can measure performance from the end-user perspective but they don't help manage the performance and capacity of the underlying infrastructure.

BalancePoint's IRT can be used to quickly show if poor end-user performance is due to infrastructure issues. If IRT is poor then the end-user's experience will likely be poor, but if IRT is good when the end-user's experience is poor, there is likely a problem with a non-infrastructure component of performance. Tracking IRT is a great way for IT infrastructure owners to prove that they are providing good infrastructure service to application owners, and can be used as a base measurement for service level agreements (SLA's). Because application owners often reside outside of IT, IRT is also a good IT management KPI, or in other words, a measurable way to tell if IT is doing a good job delivering service to the business.

IRT is expressed as seconds per transaction. A lower number indicates faster performance. Transactions are averaged units of work performed by each application (e.g. X CPU seconds, Y I/O's, Z Mb's memory). IRT values can be compared over time and across systems independent of the application-specific user transactions being performed. This is exactly the perspective of application utilization that an IT infrastructure owner can use to optimize total performance and plan future capacities.

In BalancePoint's IRT chart, as shown below, performance problems show up visually as color-coded areas of large response time.

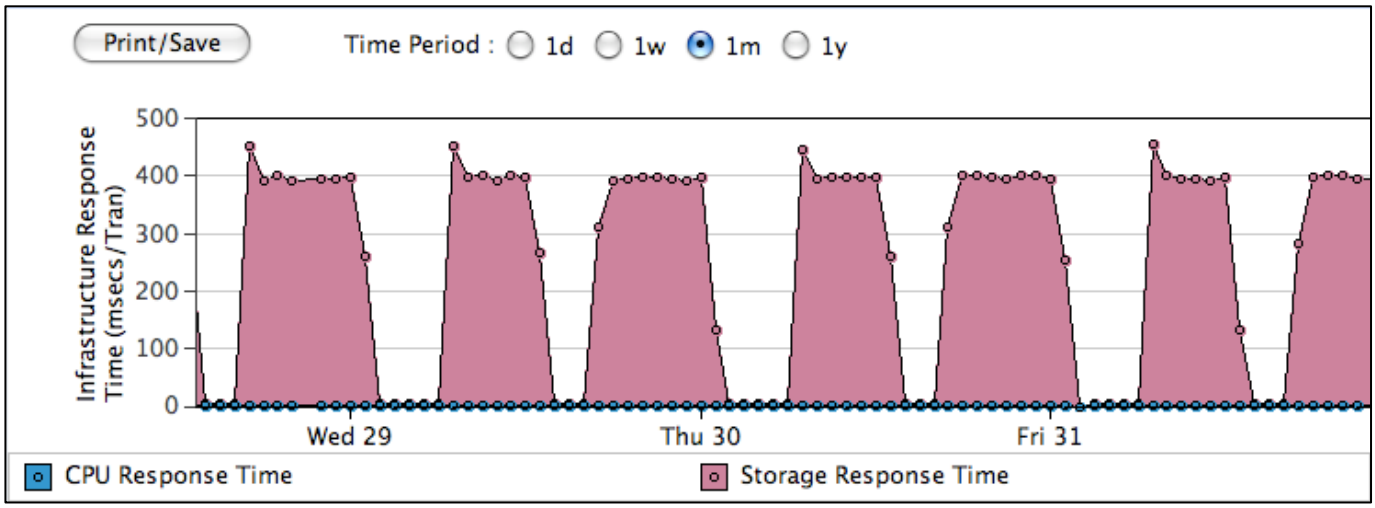

Figure 2 – In this example, Infrastructure Response Time (IRT) shows a large recurring storage bottleneck.

BalancePoint's transaction volume chart provides additional insight. If the transaction volumes are extremely low then high IRT values may not be cause for alarm. However, highly variable, large ramp-up, or sudden spikes in volume may indicate that research into application behavior and usage is warranted before further troubleshooting infrastructure. Realistic capacity plans can be developed by identifying seasonal patterns and periods of peak usage.

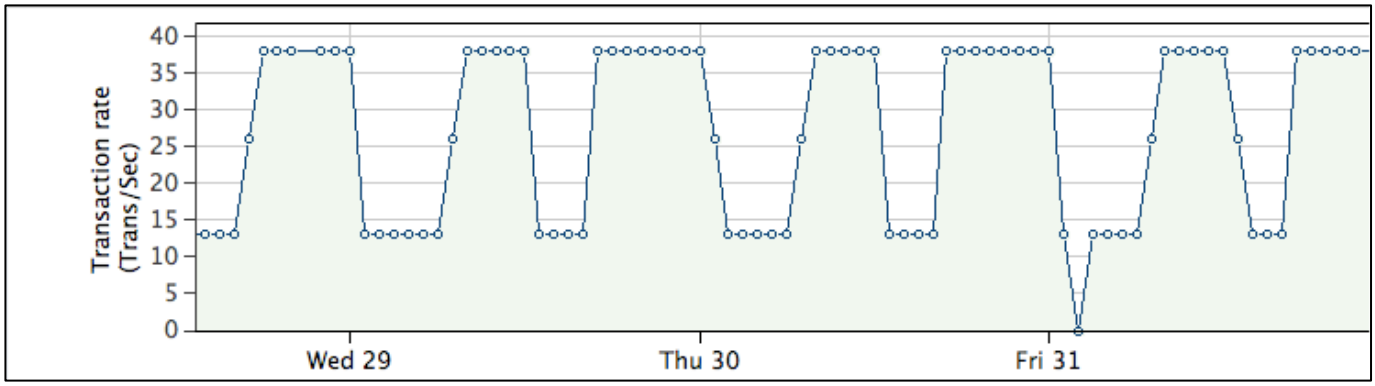

Figure 3 - Transaction Rate - Note the peaks in transaction volumes correlate to the IRT degradation.

Virtual CPU Efficiency – Quickly Determine Resource Contention or Configuration Performance Issues

Virtual CPU Efficiency is a direct VM-level index that indicates when a guest VM is not getting all the resources it wants when it wants them. A virtual CPU efficiency less than 100% indicates a resource "access" bottleneck.

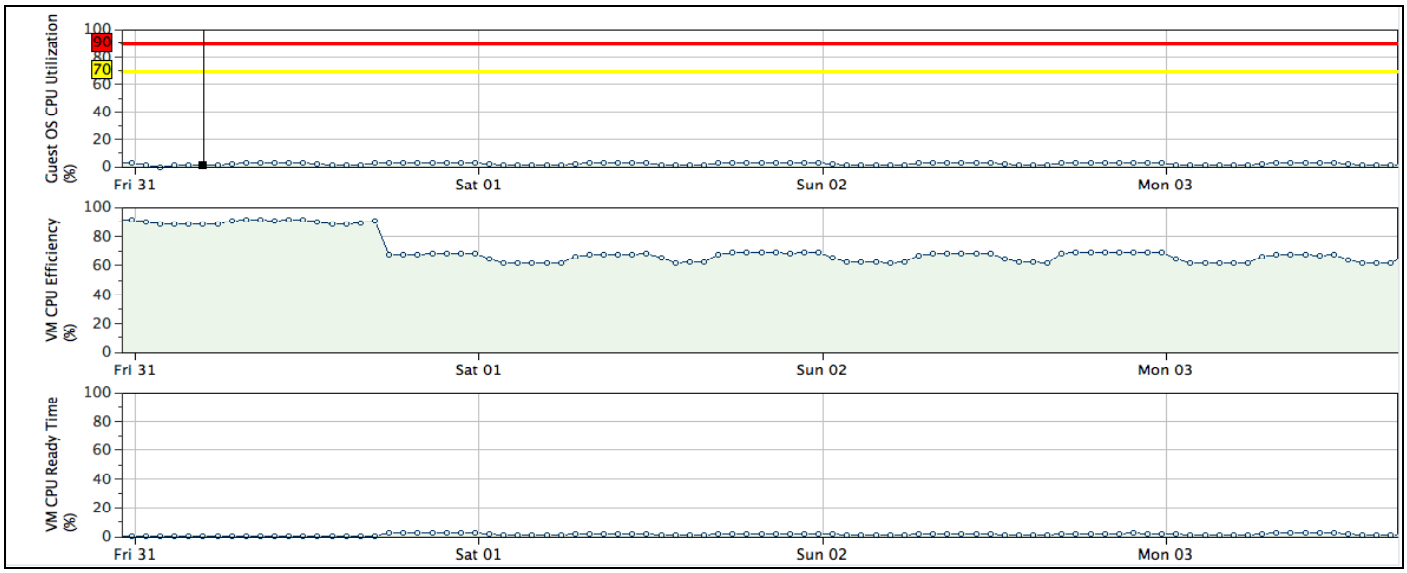

Figure 4 - Virtual CPU Efficiency, indicating when a guest VM is not getting all of the resources it wants when it wants them, is also shown with Guest OS CPU and VM CPU Ready Time in the BalancePoint GUI

If the VM CPU Efficiency drops significantly below 100%, it means that for a large percentage of time that the guest was ready to run, the system did not give it the resources it wanted. By examining the guest O/S utilization and the actual VM CPU utilization (versus its entitlement described later in this paper) it can be determined directly if there is a resource entitlement or contention issue. If not, then there is likely a misconfiguration in the number of threads/processors per VM on that specific server host (with a limited

number of actual physical processors and some number of competing VM's). The raw VM CPU Ready Time can also be compared and consulted for validation.

Optimize Performance and Utilization

Performance Index – Optimize VM Densities

While IRT is a fundamental performance metric, the Performance Index (PI) is a unique KPI produced for each application that can be used to quickly and easily optimize usage of a system by balancing capacity utilization with good performance.

PI is based on scoring the underlying IRT model relative to the system's Optimal Point (patent pending) that determines the limit of good system operating behavior. The Optimal Point represents the natural service level threshold of the system taking into account all the resources, components, and current contention from application workloads. Visually, the optimal point occurs just "beneath" the knee of the system's non-linear performance curve. Loading a system beyond the optimal point will result in an exponential increase in IRT.

PI is constructed so that it equals 100 at the Optimal Point because this is where the system can be highest utilized while still delivering good performance. Below the Optimal Point, from 0 to 100, the PI is a measure of *effective system utilization*. The distance from PI to 100 is the remaining capacity *headroom*. In this range (0-100), PI can be used as a *"% effective capacity"* for capacity planning calculations.

A low number for PI, perhaps below 20, indicates under-utilized resources that could be recovered or re-allocated (or workloads consolidated). A PI that exceeds 100 means that the current system is unable to satisfactorily handle the workload, and indicates that the work should be redistributed or tuned, or the system upgraded. A PI above 100 measures the *"% performance degradation"* (e.g. A PI of 200 indicates performance is twice as slow as at the Optimal Point).

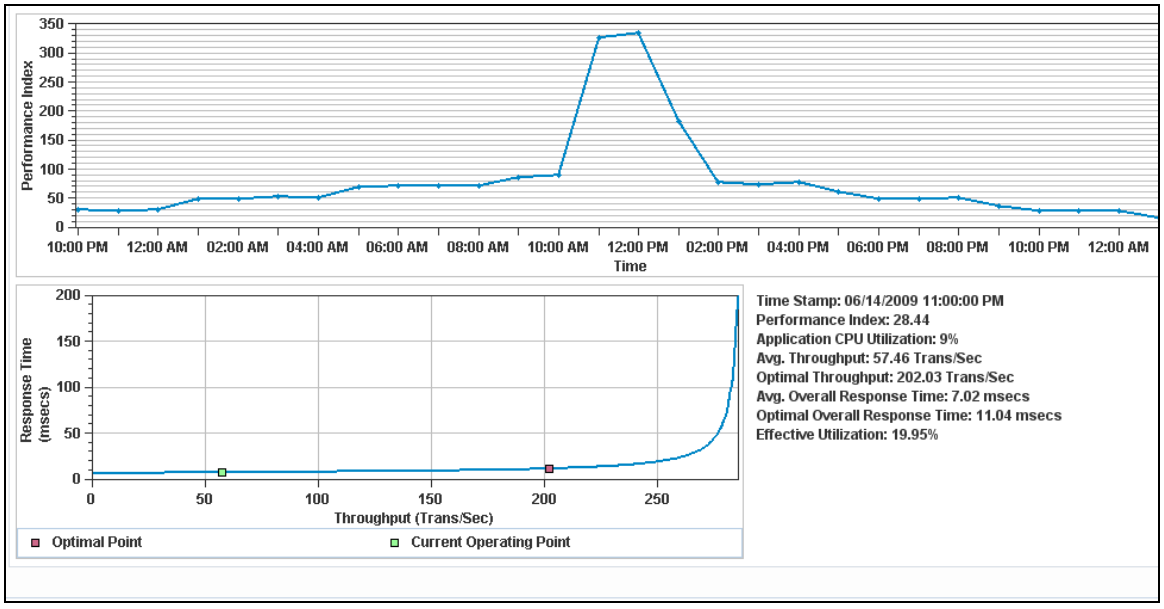

Figure 5 - The top chart shows the Performance Index over time, the lower shows the non-linear modeled performance curve with the current (green) and optimal point (red) for a specific interval.

Deciding whether to tune or migrate an application or to upgrade or re-architect a system is part of the art of performance management and capacity planning. Virtualization provides an easy resource "upgrade" path to growing applications, but do not make a determination about the appropriateness of doing so. BalancePoint provides the necessary intelligence to make optimal decisions.

To optimize the infrastructure, maintain PI numbers close to but below 100 by a capacity margin negotiated with the business for handling workload variance and expected peaks. Application owners could negotiate (and pay for) extra capacity headroom as risk insurance for unexpected peaks if desired. A benefit of virtualization is that unexpected peaks can be serviced out of the unused portion of the shared resource pool, usually at no extra cost to the application owner.

Both IRT and PI are application workload metrics. They reflect the responsiveness and capacity of a system delivering service to an application. In this manner they are also service metrics quantifying the service delivered by IT to the business application owners and can be summarized and reported on IT-business dashboards. For sophisticated users, the PI chart notes provide an additional level of detail that can be used to study service vs. wait times, maximum throughput and loading limits, and system-level efficiency and agility.

PI for virtual servers (VMware ESX) is calculated for the relationship between the physical host server and the sum of the resident VM's as a total workload. This host server PI enables managing the host server capacity in terms of average VM's - how many more will fit, should VM's be migrated or vmotioned elsewhere (and to where), and enables performance-based loading comparisons with other ESX servers (for intelligent resource scheduling).

Virtual Machine Performance Index – Baseline Performance & Determine the Optimal Size of the VM for each Application

A Virtual Machine Performance Index (VM PI) is produced for each virtualized application running in a virtual machine, enabling performance optimization and capacity planning at the guest level. VM PI enables comparison baselining and continuous management of infrastructure performance for applications as they migrate from physical to virtual (P2V) hosting. IT management can measure and baseline the IRT and PI for a physically hosted application and then continue to report the IRT and PI as the application is virtualized.

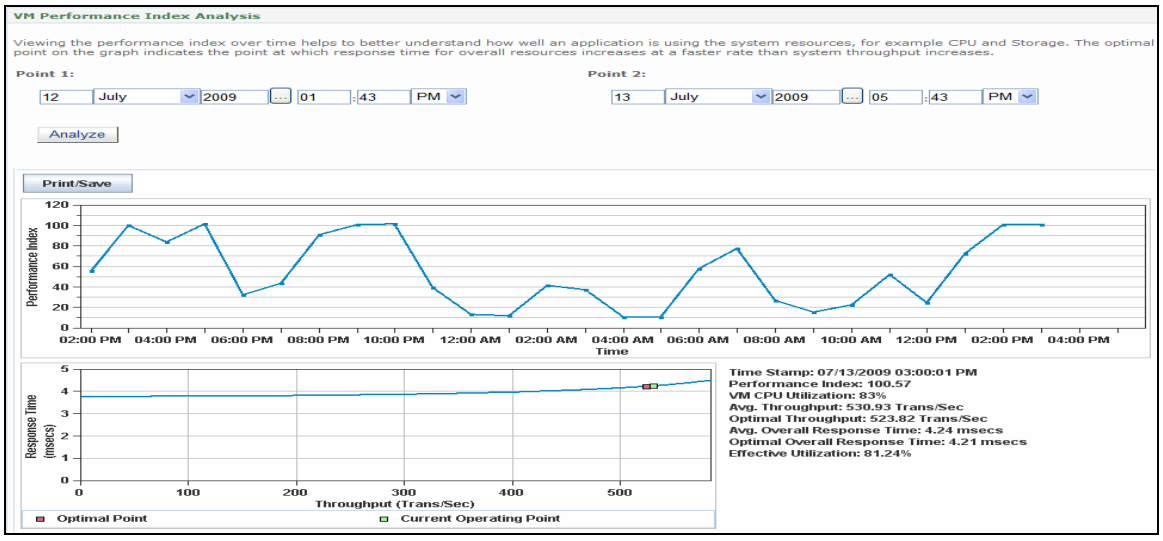

Figure 6 - VM PI determines the optimal size of the VM for each application.

The VM PI enables IT to virtualize performance-critical business applications while assuring the application owners that they are receiving as good or better performance and capacity (headroom) as when hosted on physically dedicated infrastructure. Producing and reporting on VM PI provides a transparency that fosters a working trust-based relationship between IT resource managers and business owners with revenue-impacting performance objectives.

Where PI from the ESX server perspective helps optimize the total VM load or density on the server, the VM PI helps optimize the performance and resource allocations of each virtualized application. Producing the VM

level PI requires "nested" queuing modeling that accounts for all the cross-domain resources in the system infrastructure and then also evaluates the contention caused by all competing VM applications sharing those resources.

The VM PI is interpreted the same as PI. When the VM PI is under 100 it tells how much remaining capacity headroom a VM guest application has at a good performance service level, or when over 100, how much unnecessary performance degradation the VM guest application is suffering. Virtualized application owners can rely on this one number without having to be knowledgeable about the actual IT architecture "behind" the virtual server (or private cloud) curtain.

VM Resource Entitlement Analysis – Resolve Capacity and Configuration Issues

What if the VM PI is bad? A first step in troubleshooting virtual infrastructure would be to check VM Resource Entitlement (VM RE) analysis. This metric shows the net resources each VM is entitled to based on its settings for reserves, shares, and limits, its resource pool membership (and the pool's entitlements), and any competition from other VMs. This metric enables better management of resource allocation settings for each VM including leveraging unused shares, assuring performance when adding new VMs to the resource pool, and appropriating allocating critical resources for cost savings.

When the Resource Entitlement chart shows that a VM is operating over its current entitlement, the virtual administrator can show that resource pooling and sharing is working - the VM is effectively scavenging unused resources from other VM's that are not currently using their entitlements. However, that VM is at *high risk for future performance problems* if it consistently expects or requires the extra resources above its entitlement. At any point in the future, if the other VM's are using their entitlements, the hungry VM could easily be limited to only its entitlement.

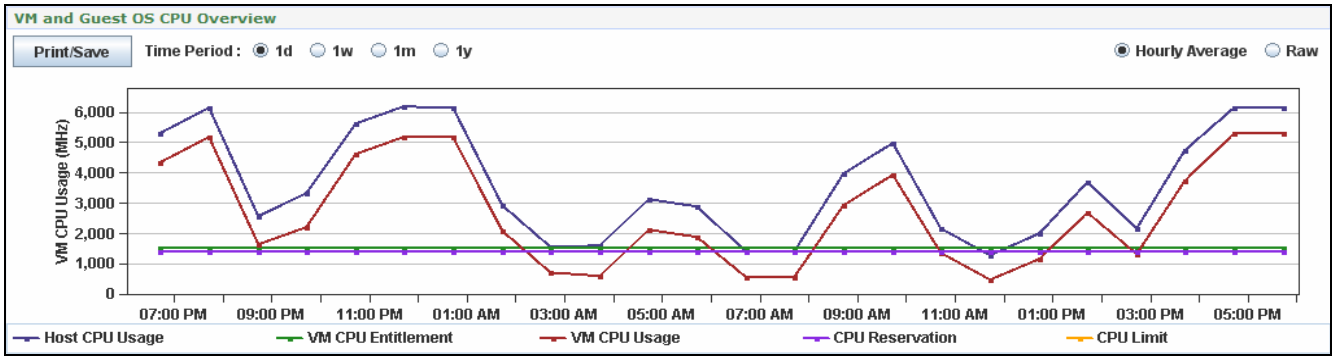

Figure 7 – Use VM Resource Entitlement Analysis to compare actual utilization (red) with the total entitlement (green).

The Resource Entitlement chart shows the reserve and limit as well as the overall host utilization. Optimally, each VM would have a reserve that met its minimum requirements, a limit that protects other VM's from application problems, and a share that sets the resulting total entitlement for optimal operation at peak periods.

When a VM is consistently operating below its entitlement and especially below its reserve there is an opportunity to tighten the VM's entitlements freeing up resources for allocation to other VM's. Blindly reserving excessively large chunks of resources results in the same kind of waste that P2V projects are trying to fix.

Virtual Resource Pool Entitlement – Determine Optimal Sharing in Resource Pools

Virtual machines may belong to an identified resource pool. Many virtual server implementations have not yet taken advantage of the benefits of defining finely grained resource pools. Dividing up a large server or a server cluster into smaller resource pools enables assuring optimal resources to groupings of applications.

Each pool's member VMs effectively share the pool's resource entitlement. (A VM's entitlement settings are evaluated within the pool's entitlement).

Resource pools make it easy for IT to allocate resources to specific business units that might run many virtual machines. Each business unit can then decide for themselves how to best share their resource pool among their own applications through VM level settings. This method guarantees resources to each business unit and avoids forcing all the VM's in a large cluster to compete head-to-head. Resource pools can insulate business units from those that tend to "hog" resources or guarantee resources to departments that are funding them without having to dedicate the resources in isolation.

Other useful resource pool strategies may group applications by type, by function, or by other relevant attribute (e.g. user geography/timezone, administrative control, business priority, or service level objective). Think of resource pools as a set of "school playground" rules for virtual machines, dividing up the playground into different spaces for each class to use as they see fit while allowing for common (or currently unused) areas to be shared according to some "fair" rules.

BalancePoint not only analyzes and charts each resource pool's overall utilization by VM, but shows the resource pool's utilization against the pool's entitlement.

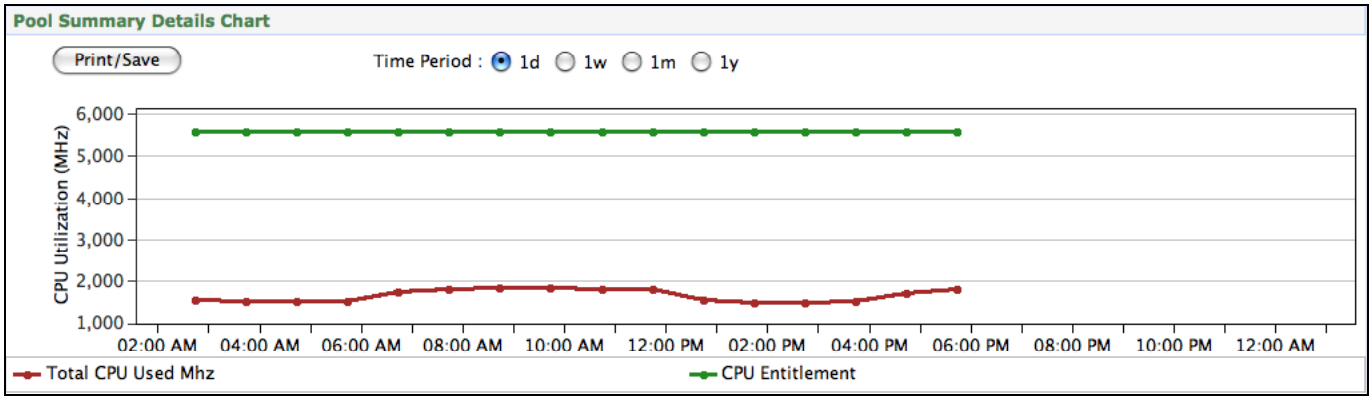

Figure 8 - Assure optimal resources to groupings of applications using Resource Pool Entitlement

If a resource pool is consistently operating above its entitlement, it may indicate a need to examine its member VM's actual requirements and entitlement settings. If a resource pool is operating significantly below its entitlement (and specifically its reserve), there may be opportunity to reallocate or "better share" underlying resources.

Plan Virtual Infrastructure Capacity

Virtual Host Resource Contention – Identify Resource Hogs

Virtual Host Resource Contention produces a unique visualization that presents a valuable averaging interval summary of actual host resource utilizations by VM. The chart is simple to interpret - the bigger boxes show the biggest VM users of CPU and Memory. It shows which VM's are using the lion's share of actual resources, or if there is a large chunk of unallocated or unused resource going to waste.

Figure 9 - - Find heavy VM users over averaging intervals using Virtual Host Resource Contention Analysis

This chart is interactively clickable, and enables quick examination of the resource entitlements and resource pools of the biggest VM "hogs" which may be candidates for possibly having a new "limit" set. Another use would be to check this chart when examining a host server or resident VM PI over 100 to see who might be the biggest users of or competitors for server resources.

VMware Cluster Capacity – Load Balance Clusters and Resource Pools

BalancePoint provides VMware ESX cluster level metrics to plan and track cluster capacity utilization. BalancePoint shows cluster utilization stacked by server to quickly examine physical load balancing, or by resource pool to analyze relative pool utilizations.

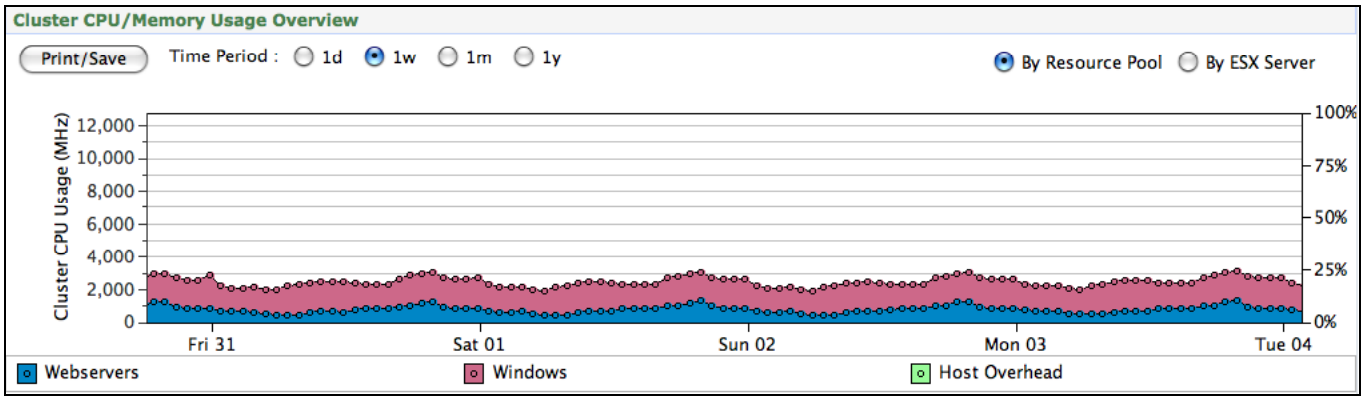

Figure 10 – Examine cluster CPU usage by resource pool to plan and track cluster capacity utilization

The server or pool "Ghz" is shown on the left y-axis, and the total cluster capacity "%" is shown on the right. Total cluster capacity can be used to judge when new resources might be needed. Depending on resource pool definition, pool utilizations can show how departments, users, or other groupings are actually using the virtual environment.

Many useful secondary statistics can be created by combining cluster utilization with average server utilizations, server or VM PI, VM application IRT, or CPU Efficiency metrics. By comparing migration policy

"vmotions" with cluster utilization and average VM PI, it can be determined if policies are becoming more active or ineffective as capacity headroom declines.

Manage Service Levels – Meet Business Requirements

Abnormality Analysis – Predict and Manage Infrastructure Service

It can be difficult for IT organizations to set a hard threshold on response times for monitoring with event management solutions. Each application's IRT will have widely different acceptable values and ranges. Since IRT is an IT "system" or infrastructure metric, any hard service level thresholds would be rather arbitrary and would not be easy to manually discover, validate or support. One labor-intensive method might be to observe the IRT over an interval of time where there was known good user experience, and to set a threshold statistically from that "acceptable" baseline.

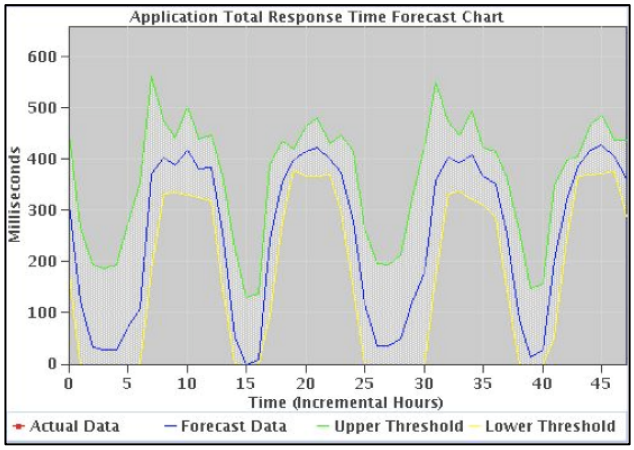

Figure 11 – Abnormality Analysis showing "normal" Infrastructure Response Time (IRT) in the light gray region between the green and yellow thresholds.

BalancePoint's Abnormality Analysis automatically calculates/updates a "normal" IRT profile over time that can be used as a predictive dynamic threshold. These IRT dynamic thresholds generate alerts when current IRT deviates significantly from the expected normal profile (by two standard deviations, or 95%).

Automated 360 Degree Reports and Scorecards on VM's, Servers, and Storage

BalancePoint collects and produces a significant amount of both cross-domain configuration data and historical performance and utilization data rolled up into dozens of reports including chargeback, storage space forecasting, VM and ESX host scorecard reporting, and best practice recommendations. There is even a pre-defined set of IT infrastructure scorecards profiling the whole data center that can be produced with push of a button.

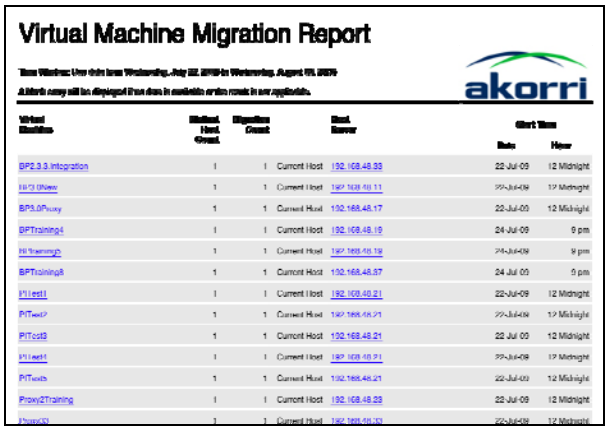

Figure 12 - Virtual Machine Migration Report records and reports on VM migrations.

In addition to the dynamic thresholding of performance service levels, BalancePoint can also provide uniquely valuable service level history through Application PI reporting. Each application PI chart immediately shows where the application received good or bad service from the allocated infrastructure and how much the application might be consuming remaining headroom capacity. For an application owner, this is exactly the information needed to decide where new efforts or investments may need to be focused without requiring a technical understanding of either the application or the actual IT infrastructure.

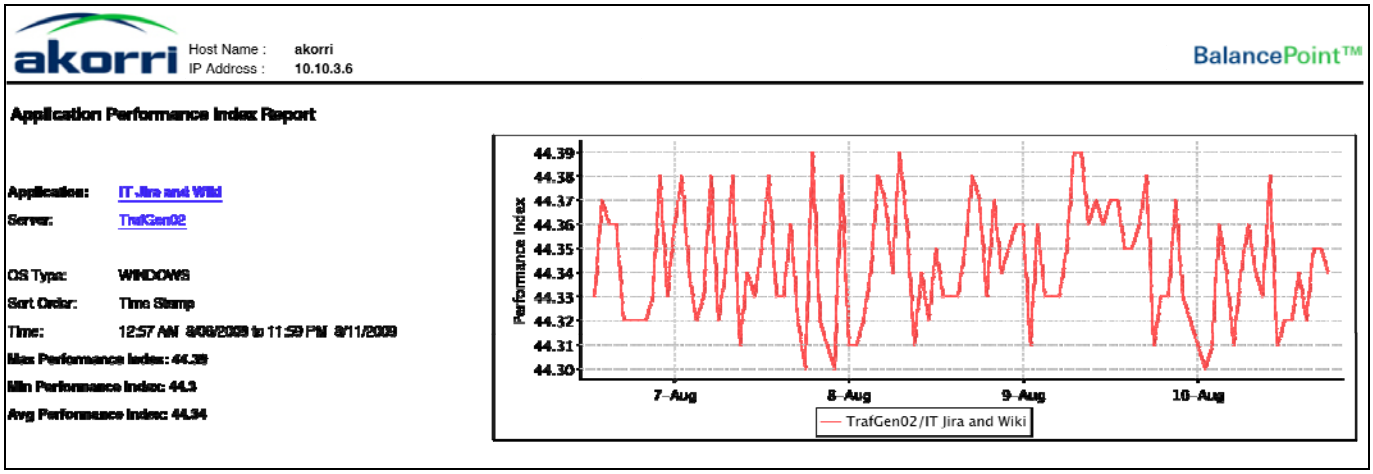

Figure 13 - Application PI Report - Manage Performance and Capacity Service Levels with one chart

Any of the previously discussed performance and capacity metrics can be analyzed against the pattern of virtual machine migrations to either validate or identify deep issues with dynamic resource management policies and automatic migration engines. BalancePoint can be used to identify VM "thrashing" where a VM is moved from host to host without actually addressing the real performance bottlenecks, sub-optimal resource leveling where physical servers are either over or under-used, and other policy-induced operational problems.

Conclusion

Using Akorri BalancePoint's cross-domain analytics and unique performance management metrics previously described in this paper, IT organizations can quickly advance their virtualization IT operations to the next level. BalancePoint's agentless appliance deploys quickly and is immediately useful. Customers typically realize ROI within two to four months leveraging BalancePoint as an independent IT advisor to:

- Increase VM densities with performance assurance
- Reclaim server and storage assets
- Forecast hardware purchases
- Improve staff visibility and productivity

Today, being able to optimally manage the performance of virtual infrastructure is a key stepping stone towards optimizing the dynamic data center and thus enabling effective private cloud computing.

ABOUT AKORRI

Akorri develops analytical software solutions that optimize performance and utilization in the dynamic data center. BalancePoint virtual infrastructure management software provides automated, infrastructure-wide visibility and analysis to help enterprises fix problems, optimize utilization and improve performance for both virtual and physical servers and storage. IT organizations use BalancePoint to reduce IT infrastructure total cost of ownership and ultimately enable a service oriented infrastructure. Follow Akorri at http://twitter.com/Akorri. For more information, visit www.akorri.com.

Akorri • 305 Foster Street •Littleton, MA 01460 •+1.978.431.1200 •Fax +1.978.428.0091 • www.akorri.com