Private Cloud Computing Performance Whitepaper

VIRTUAL INSTRUMENTS

Optimizing Private Cloud Infrastructure Performance and Availability

Reducing the risk of implementing key applications in a Private Cloud

It's no secret that cloud computing is enabled by advancements in virtualization, and with all the benefits of virtualization come some challenges. There are many tools to help identify performance problems, but at the end of the day, the whole subject of performance is a sort of a black art, not quite a science. A great deal of performance optimization is still dependent on rules of thumb and best guesses, plus brute-force over-provisioning. But it doesn't have to be.

There's a new class of tools out there designed to work with, and optimize the 'virtual infrastructure', the building block of cloud computing, and with those tools come some new IT best practices.

The Infrastructure Challenge

What is our challenge? Looking around us, we've seen a change happen in the enterprise that has crept up on IT administrators. For a long time, there has been some comfort in the "physicality" of infrastructure components. When problems come up, we've always been able to identify a switch port for examination. Or a server at the end of a cable that might be causing problems, an HBA for inspection, or any number of other physical devices we can turn to for further examination. But in the cloud, that comfort has mostly disappeared. This is primarily due to virtualization technologies. And while this trend is spearheaded by server virtualization, the variations are many. You've got application virtualization, network device virtualization, IO virtualization, file virtualization, storage virtualization, and more. And it's this virtualization which enables private cloud computing to deliver on its promises of enhanced flexibility and cost containment.

At the same time, the actual IT infrastructure has scaled almost out of control. You've got more interdependencies between systems, greater complexity, and usually, more heterogeneity. For a given application, dependencies may cross multiple applications, servers, SAN fabrics, IO adapters, network hops, and even data centers. When the performance limit of a single component is exceeded, experience tells us that these I/O-intensive systems won't see performance gradually degrade, but instead, application

Summary:

- **Infrastructure challenge**
  It's not physical; it's virtual. You have all the old challenges spurred by data growth, plus new ones caused by virtualization
- **Fibre Channel challenge**
  75% of the cloud is supported by Fibre Channel SANs, but Fibre Channel was never designed for the kind of scale seen in cloud infrastructures
- **It's the I/O**
  90% of virtualization performance related problems are associated with the SAN, but 90% of the performance discussion is focused on the server
- **The new reality**
  The cloud adds business agility, but does nothing to reduce trouble tickets or application slowdowns
- **What to look for going forward**
  Virtual Infrastructure Optimization (VIO) solutions are the best way to reduce cloud performance risk
- **VirtualWisdom**
  Without solutions like VirtualWisdom, you are literally flying blind in the cloud
- **Best practices summary**
  Take a fresh look at the requirements for optimizing performance, and avoiding application performance problems
- **Recommendations**
  Perform a self-assessment to see if you're ready for the cloud

latency and performance rocket out of control under a bombardment of delayed, dropped and repeated I/O attempts … attempts that can no longer be queued, cached, backed-off by congestion controls, or otherwise gracefully handled.

In addition, traditional management tools have become overwhelmed. Today's systems operating at immense scale have outpaced their capabilities, leading to application I/O blindness. Faced with this complexity, and potentially catastrophic impacts from any change, you're literally facing … the unknown.

Regardless of whether systems are in the cloud, it's been hard to find any way to evaluate the big picture and then drill down to perform detailed inspection of the physical and virtual infrastructure. For the most part, we just don't have the instrumentation to give us the data for troubleshooting and optimizing performance.

The Fibre Channel Challenge

Most cloud computing environments rely on a very fast, very robust shared storage infrastructure, based on Fibre Channel. When Fibre Channel was invented, no one thought it would ever exceed more than 32 or 64 nodes. Today, we've got customers supporting tens of thousands of nodes. Unfortunately, Fibre Channel has no network protocol for monitoring, like an IP network. On top of that, to make cloud computing even more challenging, we've got increasing momentum behind storage virtualization, which makes it even harder to find and fix performance problems. And it gets worse if you've got two or more vendors for each network component type. The challenge is that there is very little visibility into the SAN. SANs are not like LANs, where there are literally dozens of network monitoring tools that are enabled by protocol stacks. They just don't exist in Fibre Channel. There are really very few mechanisms to look at network traffic, performance, and path management. Yet, over 75% of cloud computing environments run on Fibre Channel SANs.

It's the I/O, the 3rd leg of the stool

"90% of the performance problems we see are I/O related, mostly due to mis-configured storage networks."
Performance Specialists VMware

When you look at where people historically focus on system management, it's around capacity and utilization. And there are good tools which do a pretty good job here. But what about performance? When you look at a virtualized environment, optimizing CPU and memory only gets you so far. For I/O-intensive applications such as those using databases, looking only at server-related metrics is simply inadequate. I/O optimization is the key as it is the third leg of the "systems optimization stool".

So we find it a little ironic that 90% of performance discussions today are all about what happens on the server. In this paper, we're going to address the effect of the storage network on application performance. We are going to discuss the other 10%, which really makes up 90% of the performance problems.

How the 'new reality' affects IT

Without true infrastructure visbility, the virtualization and cloud computing trend is hitting your business in increased capital and operating expenses. We like to say that you can't optimize what you can't measure, and today, you can't measure what's going on inside the virtualized cloud infrastructure, even with the most popular management tools found in most IT shops.

Probably the most common thing we hear is that the number of trouble tickets is going up and the time to resolution is doing likewise. One of our larger customers told us that it used to take 4 – 8 hours to find a problem, and 5 minutes to fix it. This is a huge loss in productivity and increased expenses, not to mention the impact on the end-user. Because you can't measure latency or response time in the network, the fallback is to

allocate more expensive disk arrays, and acquire more network ports. Or maybe reduce the number of VMs per physical server. The point is, there's an inability to "right-size" the infrastructure. And this means a much greater CAPEX growth than is needed. Also, we tend to see very low levels of network utilization … the SAN is heavily over-provisioned. Traditionally, the answer to application performance problems has been to deploy more hardware.

On the server side, we're seeing project delays for mission-critical applications and lower consolidation ratios than promised, resulting in decreased ROI and higher than planned CAPEX. And there are hidden challenges that people don't often talk about, the unmeasured risk of disaster, the lack of auditability, and the challenges in measuring SLA compliance. These all result in hidden OPEX costs and potentially lost revenues. As an example, we are currently engaged with a large online application provider that depends on software downloads. Their cost of downtime is quoted as over $100,000 per hour, per geography. What is the cost for an hour of downtime in your environment?

"The impact of server virtualization and the related virtual ecosystem requires updates to traditional performance monitoring and instrumentation approaches."

"It's key to remember that, in some cases, virtual infrastructures are deployed on top of a physical fabric that may not have been designed to support the increase in virtual servers, etc. Care should be taken to have visibility into potential critical choke points, especially the back-end SAN infrastructure, which is one of the most common areas where performance problems may originate."

Gartner
"Reviewing the Dimensions of Virtual Server Infrastructure Performance Monitoring"

Today's common practices

How are IT administrators dealing with this today? Well, from our conversations with even the largest enterprise data center managers, we have found that the majority of businesses today rely on their initial testing of known good configurations or arbitrary rules of thumb. They lack meaningful data when they manage and make decisions about their emerging cloud infrastructures. This makes planning an exercise in waste that consistently over-provisions resources. There are no guarantees that SLAs can be met. Troubleshooting is a bear, and flexibility is just a theory. As configurations change over time, IT administrators have no guarantees about what their current infrastructure capabilities are. We find that the majority of VMware customers have still not deployed VMotion, and of the users that have, they often have a hard time determining the performance impact when they make changes to their infrastructure. The REAL challenge is getting visible, meaningful data, about the physical and virtual infrastructure on a continuous real-time basis.

With the trend towards virtualizing the server, the abstraction layers only make things harder to see. This makes it risky to move mission-critical applications into production. Issues come up around visibility into I/O loads and characteristics. So today, the majority of virtual server deployments are in development or "light" production use … perhaps with tier 2 or tier 3 applications. Server consolidation ratios often fall well short of their original plans. Keep in mind that VMware loses sight of transactions once the I/O request is made. It has no physical layer awareness and no ability to see inside the SAN. An additional challenge in most IT organizations is that you have distinct storage and server groups. This makes resolving private cloud problems even more challenging and often leads to finger-pointing.

All this translates into wasted resources on underutilized servers, switches, and storage, an inability to track or maintain SLAs, an unnecessary IT staff resources deployed on maintaining performance. Most importantly, you've got unhappy users, or even lost revenues. Plus, more and more, datacenters are looking hard at how to save power and floor space.

Alternative solutions – what to look for

Today, many sophisticated technologies try to guide the IT administrator through the cloud. The technology analysis firm, Taneja Group, labels these solutions under the Virtual Infrastructure Optimization (VIO) banner. VIO assesses the entire virtual infrastructure, not just pieces. VIO provides IT with the data necessary to make intelligent

decisions about performance for every layer of the infrastructure, including the network, server, storage, and applications. What administrators need with VIO, is a way to see into multiple dimensions of the infrastructure, in real or at least near real-time. They need solutions that deliver the integrated monitoring and analytics required to optimize or troubleshoot virtual infrastructure performance in real time, and across every involved system – from the application to the spindle. These solutions provide the granular data needed for good decision making. Without intelligent assessment of performance, capacity and utilization can only be planned on top of assumptions that may or may not apply to any single system, or that may change at any given point in time. In our opinion, real-time performance-based analytics, inherent in the **Virtual Instruments' VirtualWisdom** VIO solution, is the required foundation for building and managing a virtual private cloud infrastructure.

Besides monitoring in real time, today's best monitoring solutions measure performance from the point of view of the application. And though every IT vendor says that they do root cause analysis, at best, they point to a slowdown in the I/O path, and it's up to the IT administrator to do some serious detective work to actually find the root cause. It's the difference between saying "hey, my son is late for his meeting" … and knowing that he has a flat tire on the interstate 30 kilometers away from home.

Even with the most modern monitoring products from the larger vendors, there's a wide gap in capabilities. Software monitoring tends to be agent based, with no interface to the physical link, and by virtue of the agent, it affects, and is affected by host system performance. Typically, if your cloud component is having a problem, it's probably overloaded. The agent might be contributing to that load problem, or to avoid compounding the problem, the monitoring turns off or takes a lower priority in the queue, when you need it the most.

Agentless software monitoring products usually use APIs that change. These products rely on the cloud component to report its own health, and they poll the device to reduce the effect on device performance. Because there's no physical layer interface and no view into the lower levels of the stack, they can't detect degraded behavior. The truth is that behavior observed high in the stack can often be predicted by observing changes at the very lowest layers. It's like getting a CAT scan and seeing blocked arteries, vs. waiting for the irregular heartbeat that indicates a heart attack.

Why do many performance tools fail to solve performance problems? One big reason is that they generate so much data, that they get ignored. Most IT infrastructures can be measured in well over 150 ways. Storage systems often report over 40 metrics, and likewise for switches. VMware's vCenter reports over 100 metrics. What do you do with well over 150 metrics? The best solutions, like VirtualWisdom, can help you weed out the noise, and point you to only the things you actually need to do something about.

"As more companies consider deploying mission-critical applications in a virtualized environment or look into deploying private clouds, it will be imperative for them to have complete real-time visibility into the response times of their virtual and physical IT infrastructure. Without the visibility enabled by solutions such as VirtualWisdom, they will be flying blind, exposing their businesses to needless risk and wasted IT resources."
Bernd Harzog
The Virtualization Practice

The VirtualWisdom solution

VirtualWisdom is a combined software and hardware monitoring solution, designed to monitor and optimize enterprise-class datacenters entering the world of cloud computing. See figure 1.

Private Cloud Computing Performance Whitepaper

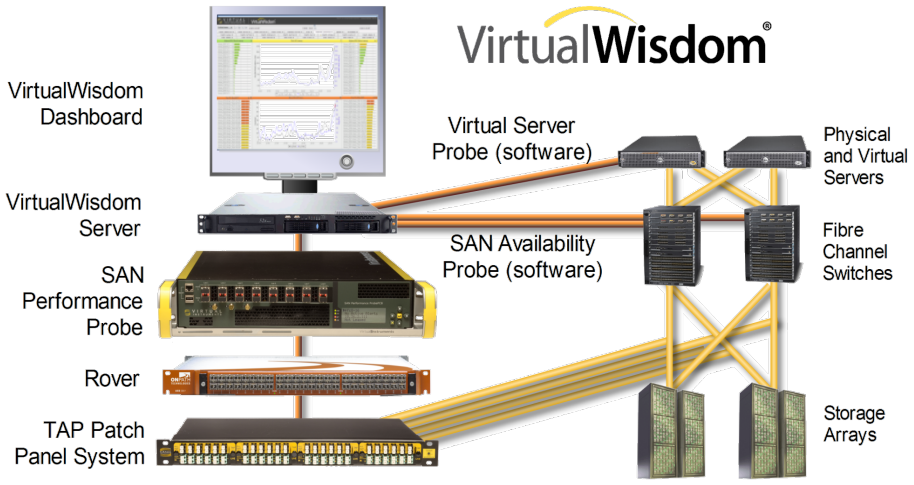

Figure 1: VirtualWisdom topology

Above, you have a representation of a SAN on the left, connected via software and hardware to the VirtualWisdom instrumentation, monitoring, measurement and analysis components on the right. VirtualWisdom monitors and analyzes Fibre Channel frames "out of band" in real-time, via optical splitters we call TAPs. With a real time physical layer connection, there is no "polling" or "averaging"; every transaction is tracked. To correlate what is captured from the Fibre Channel frames, VirtualWisdom monitors the switches themselves via SNMP. Lastly, to complete the VM to LUN correlation, VirtualWisdom monitors the VMware environment via vCenter.

VirtualWisdom's hardware monitoring is non-intrusive to the link and to the cloud component. It sees evidence of degraded behavior because it sees everything, not just the upper layers of the stack, and not just what a cloud component tells it. It has a virtually unlimited ability to record and play back transactions. Hardware monitoring is like the CAT scan or MRI, only continuous and in real-time.

When you look at where people historically focus on system management, it's around capacity and utilization. Server and storage virtualization now make capacity planning relatively easy, shifting the key criteria for success in Cloud Computing to performance. When you look at a virtualized or private cloud environment, optimizing CPU and memory is incomplete. For I/O-intensive applications such as those running OLTP databases, looking only at server-related metrics is simply inadequate. I/O optimization is the key as it is the third leg of the "systems optimization stool".

Virtual Instruments' VirtualWisdom solution assesses the entire virtual infrastructure, not just pieces. It provides IT with the data necessary to make intelligent decisions about performance for every layer of the infrastructure, from the server to the storage. What administrators need is a way to see into multiple dimensions of the infrastructure, in real or at least near real-time. They need solutions that deliver the integrated monitoring and analytics required to optimize or troubleshoot private cloud infrastructure performance in real time, and across every involved system. As industry analyst Bernd Herzog recently noted, real-time performance-based analytics, inherent in the Virtual Instruments' VirtualWisdom solution, is the required foundation for building and managing a virtual private cloud infrastructure.

As we mentioned earlier, no one is looking for reports on 150+ metrics, they're looking for intelligence to point them where action is needed. A great example of that would be reports and alerts on Exchange Completion Time. ECT will tell you the effect of the

infrastructure on application response time, as latency. Performance experts at VMware point to the storage infrastructure as the culprit in 90% of the performance problems related to production use of virtualization. Real-time transaction latency data will immediately tell you if the SAN is the cause of the problem. No guessing or hunting is required.

In figure 2 below, the graph on the left shows latency as reported by a typical vCenter tool. According to the graph, the highest write latency for this SAN is 22 milliseconds, not too bad. On the right, the same SAN is monitored by a VirtualWisdom hardware probe. You can quickly see, at a glance, what your longest latencies are for each Initiator-Target-LUN (ITL) path. In this example, the customer set a threshold at 40 milliseconds, and you can see that one ITL exceeds that threshold. Without granular, real-time monitoring, you would miss the problem, though your application user unfortunately wouldn't. One of the great things about infrastructure latency is that it will immediately show what effect any SAN has on the cloud-based application. If you consolidate a couple of switch ports, you'll know immediately what difference it makes.

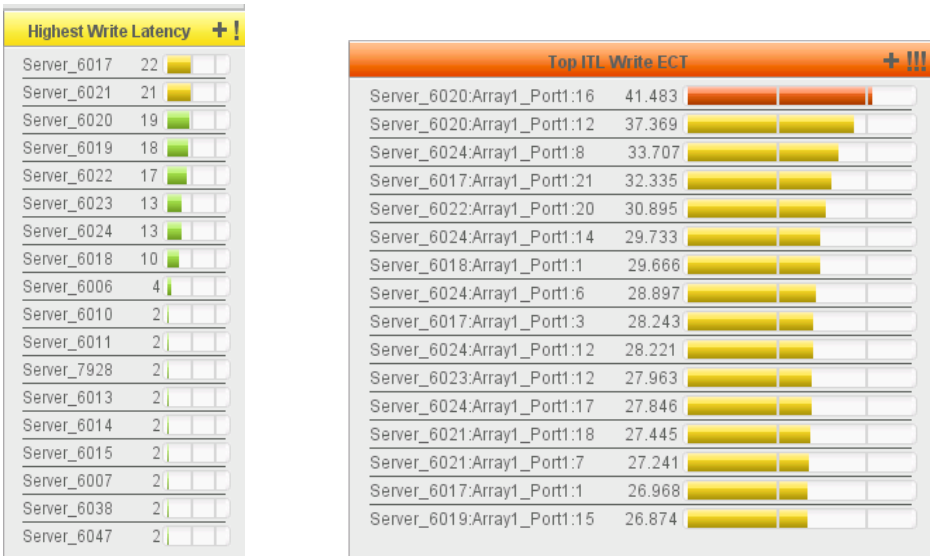

Figure 2: Exchange Completion Times

If you are preparing reports for your application team, you can use VirtualWisdom's User Defined Context capability to build reports that are meaningful to them. For instance, in figure 3 below, we show the "before" and "after" effects of configuration changes on application read response times. Clearly, the changes made were very effective.

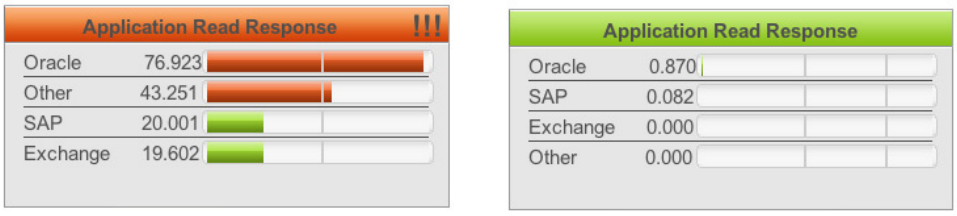

Figure 3: application performance reporting

Private Cloud Computing Performance Whitepaper

Best practices summary

Some things to look for in a cloud-computing monitoring environment include:

- Vendor-independence, as it avoids finger pointing and, at the same time, provides your vendors with helpful metrics
- Availability of predictive metrics, to avoid problems and reduce trouble tickets
- A real-time solution with fine and course-grained metrics to the deepest levels
- Ability to monitor across multiple domains, not just single components
- Whether it's called in-band or out of band, it should have no impact on the production applications
- 100% heterogeneous, works with a servers, switches, and storage arrays to avoid lock-in and reduce the number of separate tools
- Presents actionable data, not just metrics
- Proven scalability to tens of petabytes, and tens of thousands of servers and switch ports.

Lastly, though performance is hardly the only thing that creates risk for a cloud environment; there are security issues and data availability risks; a properly instrumented cloud environment can go a long way towards eliminating the performance and availability risk.

Recommendations

We suggest you complete a self-assessment and ask yourself these questions.

- Are your cloud infrastructures being planned with the essential perfofmance and utilization information, or are they just happening?
- Are you delaying implementations because you're afraid that the cloud will only make current problems worse?
- How are you monitoring both cloud infrastructure and application performance?
- Do you know if your response times are degrading due to infrastructure issues?
- Are your queue depth settings correct? Are you measuring them?
- Are your I/O paths sufficiently provisioned for throughput and redundancy?
- What would it do to your business if response times substantially lengthened due to errors or congestion?
- Does your monitoring alert you to conditions that are irrelevant, while not informing you of conditions that are likely to hurt your business?
- Finally, ask yourself, when it comes to the health and performance of your cloud-based application, are you flying blind?

VIRTUAL INSTRUMENTS

Virtual Instruments
25 Metro Drive Suite 400
San Jose, CA 95110
Phone: 408-579-4000
Sales Information
sales@virtualinstruments.com
Phone: 408-579-4080
Customer Support
support@virtualinstruments.com
www.virtualinstruments.com