The Ultimate Online Experience®

INTERNAP®

Latency: The Achilles Heel of Cloud Computing

Cloud may not be the panacea IT practitioners hope it will be. If end-user customers become disenchanted with application performance at the edge, latency will prove to be the Achilles Heel of Cloud. Find out why cloud computing should be part of a total solution to address IT objectives, while simultaneously improving the customer experience.

An Internap White Paper

©2010 Internap Network Services Corporation. All Rights Reserved. 11/10

Table of Contents

Summary 3

The Problem 4

Every Millisecond Counts to Demanding End-Users 5

Where Cloud Does Impact End-Users 6

Latency and Downtime in the Cloud 6

Measuring Latency and Downtime from the Cloud to End-User System 7

Limitations of the Major Carriers 7

TCP Not Ideally Suited for Internet Performance Today 9

Current SLAs Don’t Hold Up to Scrutiny 10

The Solution 10

Measuring Total System Performance 12

Next-Generation Service level Agreements 12

Summary 13

About Internap 13

Summary

Cloud computing represents a long-standing vision of moving application computing power to the Internet at a much lower cost per cycle. Until recent years, implementation of this vision has been limited for several reasons. Partly, this was due to performance levels from underlying hardware or software being simply not practical or acceptable. Partly, this was due to insufficient network technologies across Internet, WAN and LAN. Finally, necessary technologies such as virtualization, self-provisioning and metering, and security had yet to be fully and reliably developed.

Cloud computing – whether supporting Software as a Service (SaaS), Platform as a Service (PaaS) or Infrastructure as a Service (IaaS) – is a complex, multi-faceted technology solution incorporating numerous hardware, software, middleware, security and monitoring, and other capabilities to ensure application and information processing, reliability, performance and security across the Internet.

Today, cloud computing models are proliferating across the Internet. Performance can be measured and benchmarked across various vendor offerings. To the Information Technology (IT) world, there is strong interest in taking advantage of cloud environments, especially from a flexibility and cost standpoint. Indeed, cost savings probably represents the single most compelling reason to employ cloud services.

"The Cloud" – Defined

"Cloud computing is Internet-based computing, whereby shared resources, software, and information are provided to computers and other devices on demand, like the electricity grid. Cloud computing is a paradigm shift following the shift from mainframe to client-server in the early 1980s. Details are abstracted from the users, who no longer have need for expertise in, or control over, the technology infrastructure "in the cloud" that supports them."

Benefits to IT organizations include:

- Reduces capital expenditures
- On-demand scalability/provisioning and pricing model
- Eliminates IT department infrastructure overhead
- Allows rapid implementation timelines that can align with strategic and tactical initiatives

Source: Wikipedia, citing "Distinguishing Cloud Computing from Utility Computing," Krissi Danielson - March 26, 2008

CIOs and the companies they serve are largely who benefit from deployment of cloud environments in terms of reducing costs of the organizations they oversee. As a result, IT organizations are sure to see substantial change in tasking and resources as cloud's prominence continues to take hold.

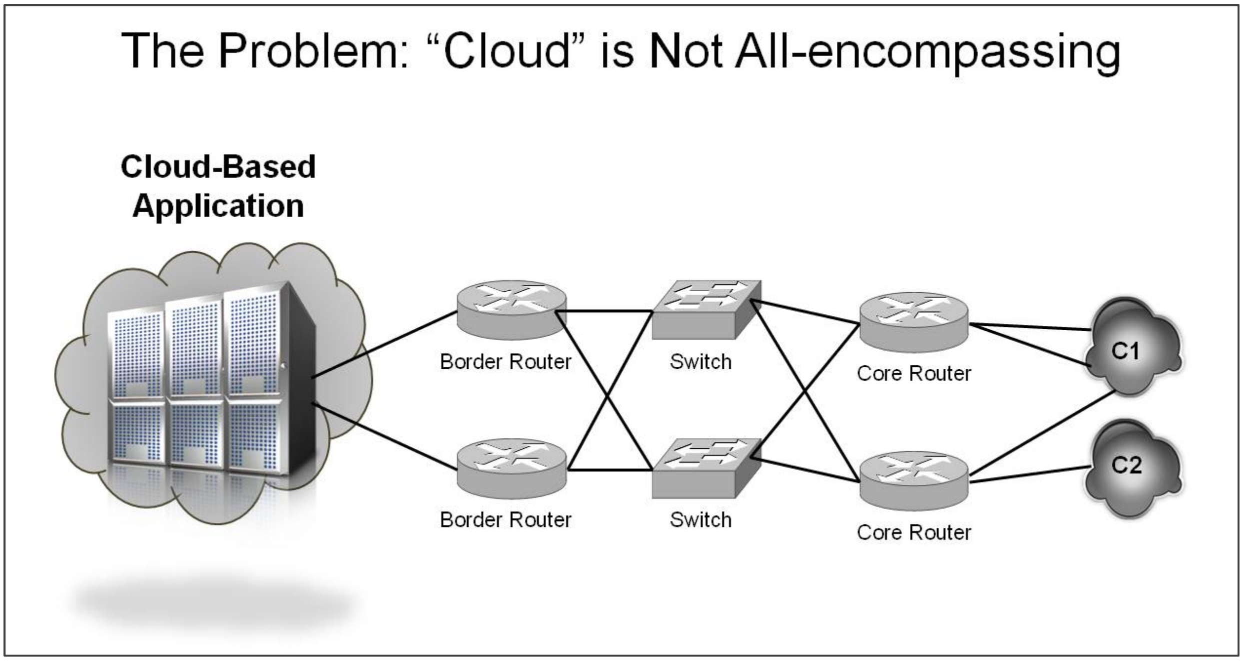
But though cloud computing can provide an IT-oriented solution to reduce costs within the organization, cloud computing doesn't address the broader issue of latency from the cloud edge to the end-user. Considerations of cloud usage often take a myopic view ignoring the end-user, a potentially critical mistake for IT organizations to make.

The Problem

The inherent nature of cloud computing is accompanied by significant risk. Data becomes obscured in the Cloud and may be hosted in multiple remote domains leading to questions about cloud security and compliance risks – risks that stand to delay the adoption of cloud services for anything beyond non-mission-critical applications and infrastructure.

Latency (delay) represents another major issue. However, concerns about intra-cloud performance and reliability often overshadow the performance and reliability of the overall application and content delivery chain from the cloud environment to the end-user. It is this combined latency that can manifest itself as a simple echo annoyance on a VoIP call, or spell disaster for a Massive Multi-player Online Game (MMOG) provider with thousands of users playing performance-sensitive games at any given time.

Cloud computing doesn't effectively deal with the issue of end-user expectations and demands for performance and reliability. This is why the discussion of cloud latency has to shift away from IT-defined acceptable levels of latency to end-user behavioral judgments as to what latency is acceptable. After all, end-users will penalize applications and websites based on the smallest performance delays or downtime.



The Problem: "Cloud" is Not All-encompassing

Cloud computing is a compelling solution that can enable IT organizations to gain on-demand IT infrastructure flexibility and cut costs significantly by using a utility computing model, but cloud computing is not designed to address the issue of latency across the Internet.

Geography and distance still play a key role in estimations of latency. The further the cloud environment is from your internal network systems or the end-user, the greater latency across the network. The end-user is platform-agnostic and simply demanding in their expectations for seamless performance and reliability of your application and websites. Latency within the Cloud, and across the network, could result in a solution that costs your organization money.

Cloud computing must meet all the same stringent demands for availability, performance and reliability that exist for IT systems today. But its consideration must also take into account latency and reliability beyond the cloud edge.

Every Millisecond Counts to Demanding End-Users

When it comes to end-user requirements for application and website performance, every millisecond counts. Numerous studies have shown that end-users are highly demanding. They expect application and website load times to be faster than their inclination to click away.

Application performance has become a critical competitive requirement, which has a direct impact on customer satisfaction, top line revenues and the bottom line.

When evaluating website performance, according to Equation Research[1] sanctioned by Gomez, a poor web experience results in lost revenue opportunity, a poorer customer perception of your company and can boost your competitor's bottom line.

- 78% of site visitors have gone to a competitor's site due to poor performance during peak times.
- 88% are less likely to return to a site after a poor user experience.
- 47% left with a less positive perception of the company.

Aberdeen Group[2] provides a similar snapshot of demanding user requirements for website performance.

- A 1 second delay reduces customer conversions by 7%.
- A 1 second delay decreases customer satisfaction by 16%.
- A 1 second delay decreases page views by 11%.

Though results below are for some of the largest Internet sites, the message remains the same – application downtime or slow performance costs enterprises money.

- Shopzilla: A 5 second speedup in website performance resulted in 25% more page views, a 12% increase in revenue, and a 50% reduction in hardware.
- Amazon.com: Every 0.1 seconds in latency reduces sales by 1%.
- Google: Every 0.5 seconds in latency reduces traffic by 20%.

Latency also impedes business models. For example, cloud providers today are unable to offer premium Service Level Agreements (SLAs) for mission-critical applications and infrastructure.

[1] "When more Website visitors hurt your business: *Are you ready for peak traffic?,"* Equation Research - 2010

[2] "Performance of Web Applications – Customers are Won or Lost in One Second," Bojan Simic, Aberdeen Group - November 2008

Enterprises would gladly pay for such agreements, but there literally is no way cloud providers can guarantee the appropriate service levels. In fact, the "big 3" cloud providers today – Amazon.com, Google, and Microsoft – can offer no better than 99.95% as an SLA. This represents 21.6 minutes of unscheduled downtime per month, which is insufficient for enterprise-class applications.

As a result, latency not only prevents enterprises from embracing cloud providers for critical business infrastructure and applications, it prevents cloud providers from being able to generate revenue from providing such infrastructure.

Where Cloud Does Impact End-Users

IaaS and PaaS cloud solutions can allow organizations to more rapidly deploy new applications and capabilities that intersect and advance changing customer requirements. This is a key benefit of Cloud – to provide enterprises the ability to be more agile in application testing and deployment. Indeed, in a 2010 KPMG survey[3] of current cloud adopters, the highest rated realized benefit from Cloud cited was "more flexibility." The result is that organizations have the potential to shorten development and production timelines by using IaaS, PaaS, and SaaS cloud environments.

Cloud's lasting impact on end-users may be this: more applications get delivered sooner as software product life cycles and go-to-market life cycles are accelerated. And if the application is monetized, an accelerated launch schedule can result in additional revenues to the company that otherwise would have been delayed or lost through the previous deployment model.

End-users experience new capabilities sooner.

End-user interest and stickiness is promoted.

Latency and Downtime in the Cloud

Performance and availability within the Cloud are key concerns expressed across a number of surveys.

- "Performance" was ranked a 3.9 (on a scale of 5, "5" being "very concerned) in an InformationWeek Analytics[4] survey.
- "Performance" was cited as a top 5 indicated concern by SearchCIO.com[5]

KPMG's report surveyed current cloud adopters to identify realized benefits of Cloud and what these early adopters would like to see improved. Surprisingly, "Performance" and "Availability" were not cited as realized benefits from their cloud usage. Both were identified as areas that should be improved.

[3] "From Hype to Future: KPMG's 2010 Cloud Computing Survey," - KPMG the Netherlands - 2010

[4] "InformationWeek Analytics/Bank Systems and Technology Cloud Computing Survey of 186 banking technology professionals," InformationWeek - January 2010

[5] "Five Top Concerns About Cloud Service Providers" SearchCIO.com - May 6, 2010

Interestingly, in Aberdeen's research study on cloud infrastructure performance, cloud adopters were surveyed on any realized performance improvement from their cloud deployments.

- Respondents overall reported an average 5% performance improvement.
- 30% of respondents indicated greater than a 10% performance improvement.
- 29% of respondents indicated a 1-9% performance improvement.
- 32% did indicate no change to performance levels as a result of deploying Cloud.

Given the low average improvement of 5%, arguably it should be no surprise that performance remains a concern for cloud computing for many IT organizations.

Measuring Latency and Downtime from the Cloud to End-User System

Determination of cloud and solution providers must take into consideration end-user behavior, as it relates to application or website poor performance and downtime.

For purposes of this white paper, Internap applies the following simplified formula to highlight the crucial perspective that IT organizations must employ when utilizing cloud services.

$$C_L + N_L = TS_L$$

Where C_L equals intra-cloud latency, N_L equals network or Internet latency, and TS_L equals Total System latency. TS_L is the statistic whereby the application or website should truly be measured against in terms of end-user requirements.

What's your application or website's Total System latency or "TSL" and how does it match up to pressures to perform?

Downtime can be indicated by traditional percentage measures in terms of the percentage of availability over a given 12-month period. As cloud providers will usually indicate their level of "guaranteed" uptime in their SLAs, it is still critical to measure uptime using independent tools. Understanding SLA definitions of downtime can also identify gaps in what the IT organization's definition may be versus that of the cloud provider.

Applying these formulas allows the enterprise to determine the cost of latency and downtime across the total system from an end-user perspective. Admittedly, this would be distinctly separate but complementary to the ROI for utilizing a cloud-based environment versus other in-sourced or out-sourced IT options.

Limitations of the Major Carriers

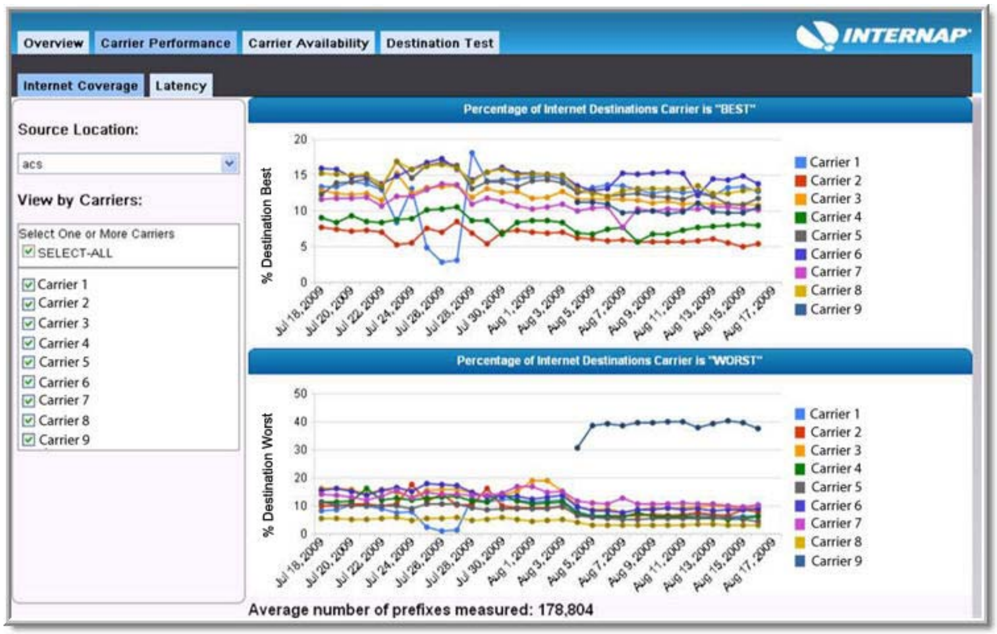
At the foundation of the Internet's framework is the intelligence called Border Gateway Protocol (BGP) – which directs how large amounts of data are transported between one network node and another. BGP was implemented to allow for a decentralized Internet routing scheme replacing Exterior Gateway Protocol, which routed based on a more tree-like network topology.

BGP's sole purpose is to determine the route or path that data is to take over the network configuration (i.e. the Internet) – a purpose BGP has facilitated generally very well since its inception. However, a limitation in its methodology is that because of the load shouldered by the network, the shortest path is not always the best route, and alternate, more optimal paths are not prioritized. As a result, this less-than-optimum routing method over the Internet causes latency, which is represented by slow reacting applications and websites, and frustration on the part of the end-user.

It has long been a standard practice for organizations seeking a reliable and consistent connection to the Internet to use multiple carriers for that connection. This "redundancy" served as insurance and ensured that as one carrier's network failed, the other carrier would take on the organization's traffic and keep applications up and running. The use of multiple carriers to maintain a connection to the Internet speaks to the importance of the applications that are now placed there.

The problem with this approach is that no one or combination of two (or three or four) carriers ensures an optimal connection to the Internet. Redundancy may keep the connections up, but they are not addressing the core need or expectation – which is optimal routing of traffic to reduce latency, jitter and packet loss.

Traffic patterns across the major Internet backbone carriers[6] show that individually, they perform similarly to the structure of the "Bell Curve" or to a normal distribution.



Source Location:

acs

View by Carriers:

Select One or More Carriers

☑ SELECT-ALL

☑ Carrier 1
☑ Carrier 2
☑ Carrier 3
☑ Carrier 4
☑ Carrier 5
☑ Carrier 6
☑ Carrier 7
☑ Carrier 8
☑ Carrier 9

Average number of prefixes measured: 178,804

The chart above points to some significant realities:

- A given carrier possesses the most efficient traffic route approximately 12% of the time.
- A given carrier possesses the least efficient traffic route approximately 12% of the time.

[6] Source: Internap's IP-Scope monitoring solution. More than 200,000 prefixes are constantly monitored to determine the optimal carrier path / individual carrier's performance.

- There resides a more optimal routing path among other carriers 85% of the time. Inversely, any one carrier is providing “less than optimal” routing 88% of the time.

Selecting a combination of carriers that can somehow comprise a more-efficient routing solution for web-based applications is a “hit and miss” proposition since finding the optimal path across carriers is a fluctuating objective, if not incomprehensible to determine and manage.

TCP Not Ideally Suited for Internet Performance Today

Transmission Control Protocol (TCP) is focused more on ensuring packet delivery versus achieving performance, and is an overly conservative protocol given the characteristics of the Internet today.

From the graph below, one can see throughput performance of native TCP (blue) versus TCP acceleration technology (red).

<u>Test Parameters</u>
Receiver: Windows XP
File: 100 MB Compressed
Link: 100 Mbps

Native TCP has two primary objectives that it tries to satisfy in every communication – govern the flow of data by monitoring congestion and ensure that all packets arrive at the destination.

The problem is that native TCP uses packet loss as the metric to evaluate whether it is achieving both objectives. The error is that packet loss is not always due to congestion (especially in high performing networks), so invoking a start / re-start in this instance is too inefficient.

As a result, TCP’s check mechanisms often result in added latency for data transmission and reduce overall bandwidth utilization rates significantly.

Current SLAs Don’t Hold Up to Scrutiny

Promises for performance, especially when they relate to such a complex system involving the two major components of cloud infrastructure and the network, must be backed up by service level agreements that mean something. Today, Service Level Agreements across cloud providers fail to align with market realities that dictate a highly responsive application or website experience. Often, the cloud provider will guarantee levels of uptime but won’t guarantee any latency threshold. Some cloud vendors will go so far as to guarantee a threshold for latency, but this is calculated only for operations within the Cloud and completely ignores the network aspect of content communication across the Internet.

All of the major vendors offer credit for what they define as excessive downtime. Some of these processes require detailed documentation from the customer on when the outages occurred and how long the network was affected. The customer has to initiate the request for credit. One could argue that the cloud provider would be in the best possible position to know when an outage and downtime has occurred and spare their customers the burden of proving their case for consideration.

WARNING: Understanding your ROI is critical.

A “credit” for downtime may not replace lost customers and revenue and typically does not address the impact of a slow network.

The Solution

End-user requirements or demands must take precedence in assessments of cloud computing environments and their usage.

For any application deployment that is crucial to the organization’s revenue model, IT’s objective should be to optimize application processing and content delivery paths into a cohesive, structurally optimized end-user experience. To achieve this, the organization must address latency requirements across the total system with the end-user in mind.

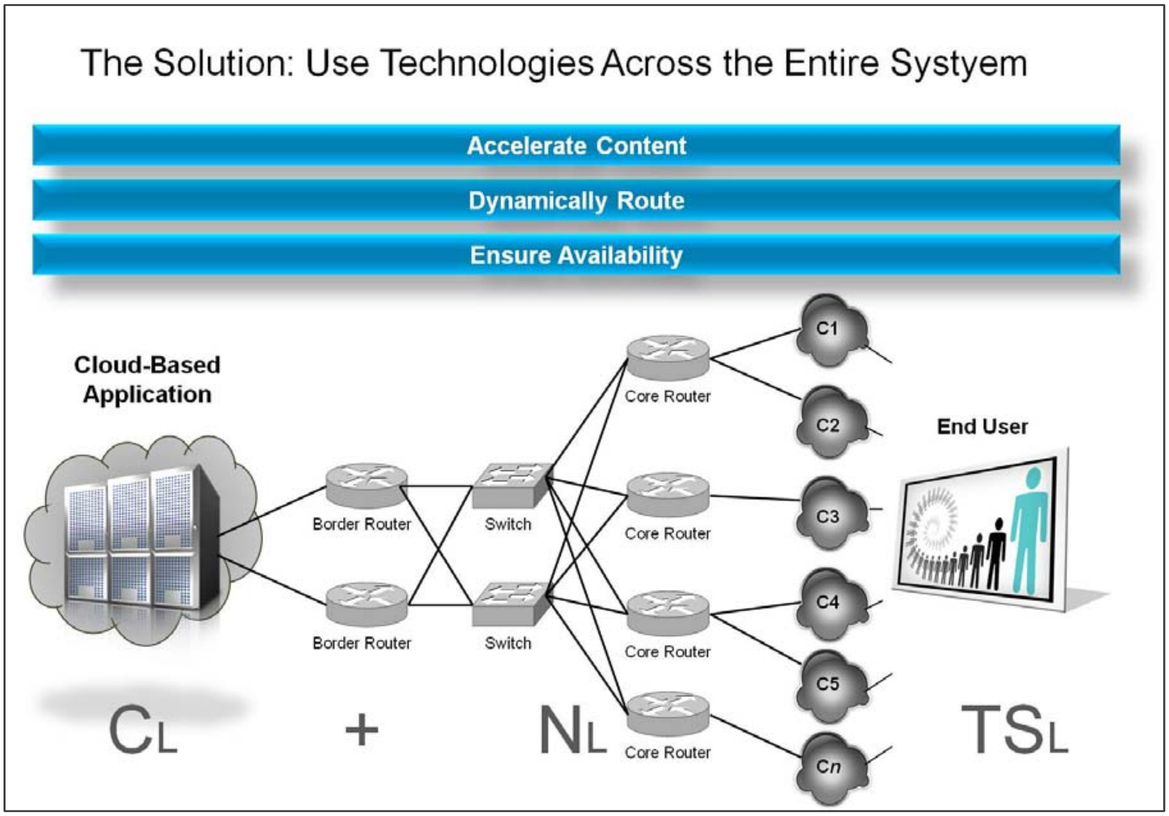
To summarize in the formula below:

$$C_L + N_L = TS_L$$

whereby,

$$TS_L \leq \text{End-User Requirements}$$

Total System latency should not exceed the estimated acceptable performance requirement from the end-user’s perspective. Anything above acceptable performance levels raises the risk and likelihood that end-users will penalize an application or website by abandoning it.



The Solution: Use Technologies Across the Entire Systyem

Accelerate Content

Dynamically Route

Ensure Availability

Solution Components

- Reliable data center architecture with multiple layers of redundant infrastructure to remove single points of failure.
- Optimization of cloud infrastructure (i.e. use of high-performance physical and virtual cloud architectures, and related software technologies).
- Optimization of IP traffic through dynamic, intelligent traffic routing mechanism to reduce latency and improve reliability.
- Optimization of application code for a cloud-based deployment model.
- Optimization of website coding (HTML5, CSS3, Flash, etc.), and optimized small and large object content.
- Optimization of IP traffic delivery through TCP acceleration technologies.
- Optimization of static and dynamic content delivery from the network edge to end-user (i.e. a content delivery network).
- Measurement mechanism for performance across the entire “cloud to end-user” system.

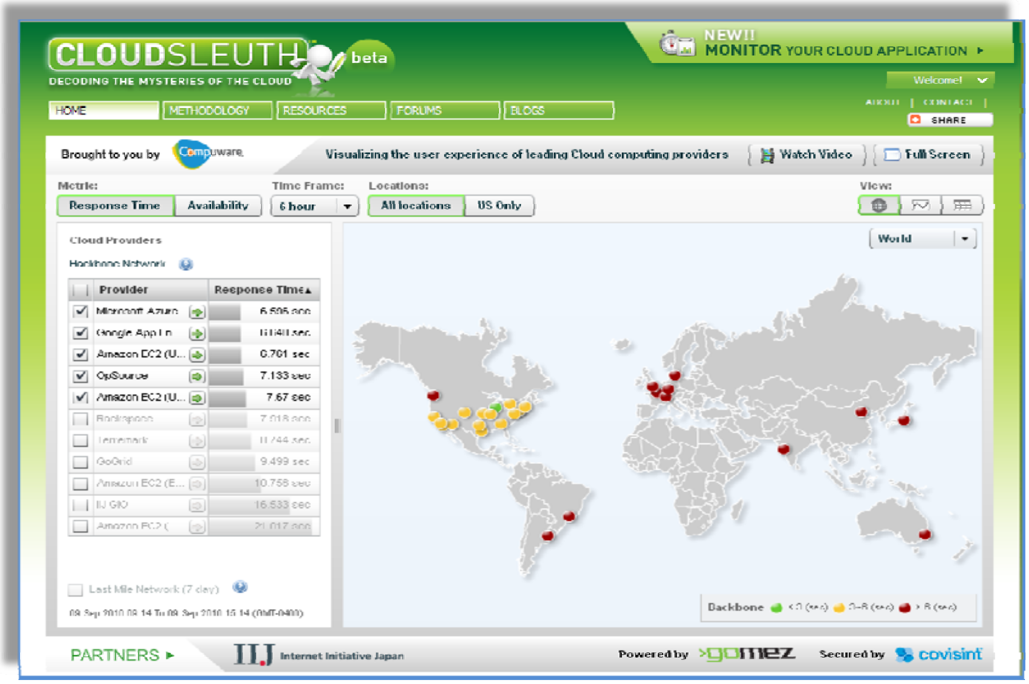
- Robust Service Level Agreement with performance and availability guarantees.

Though the cloud computing environment and Internet connectivity environments largely fall out of direct influence for IT organizations, organizations do have choices as to what cloud and IP solutions they employ to satisfy their performance and reliability requirements.

Measuring and reducing latency while improving site reliability across the content path from Cloud to end-user is key to maintaining future competitive standing.

Measuring Total System Performance

The good news is that there are a number of reporting capabilities, also known as Application Performance Management (APM) applications that can provide a holistic view of total system performance from Cloud to end-user. In addition, these tools can measure performance across more than one cloud environment, a clear benefit to IT organizations looking for a central console to measure activity and performance for numerous cloud deployments.



Visualizing the user experience of leading Cloud computing providers

Metric: Response Time | Availability

Time Frame: 6 hour

Locations: All locations | US Only

View:

Cloud Providers

Backbone Network

	Provider	Response Time▲
☑	Microsoft Azure	6.595 sec
☑	Google App En	6.640 sec
☑	Amazon EC2 (U...	6.701 sec
☑	OpSource	7.133 sec
☑	Amazon EC2 (U...	7.67 sec
□	Rackspace	7.918 sec
□	Terremark	8.744 sec
□	GoGrid	9.499 sec
□	Amazon EC2 (E...	10.758 sec
□	IIJ GIO	16.533 sec
□	Amazon EC2 (	21.017 sec

□ Last Mile Network (7 day)

08 Sep 2010 09:14 To 08 Sep 2010 15:14 (GMT-0400)

Backbone 🟢 < 3 (sec) 🟡 3–6 (sec) 🔴 > 6 (sec)

CloudSleuth™ powered by Compuware Gomez Application Performance Management tool. Measures various cloud computing environments' performance and availability.

So, total system performance can now be measured from Cloud to end-user which allows IT organizations to look closely at how they can address latency in both major areas, as opposed to within the cloud environment only.

Next-Generation Service level Agreements

A point of differentiation across cloud providers may stem from the SLAs themselves. Given the level of "trust" that customers must have to place applications and information in the cloud environment, cloud vendors must have higher-level SLA and support models to address cloud performance/reliability and customer-related inquiries.

Some key aspects of next generation SLAs may encompass:

- Guarantees for performance and reliability across the total system from Cloud to end-user
- Transparent cumulative capture of how much downtime has occurred across the system
- Transparent capture of actual performance in terms of latency to SLA guarantees
- Proactive, automated crediting mechanism when downtime SLA guarantees are exceeded – in other words, automated “reverse metering”
- Generous credit scheme
- Robust customer support mechanisms that scale with the level of customer services commitment

Summary

Cloud computing represents a compelling and proven solution for IT departments to increase flexibility and cut costs typically associated with the deployment of new platforms. However, the consideration of cloud computing should not be limited to this view and should take into consideration the entire system solution – from Cloud to end-user.

Selection of a cloud solution in this regard can yield benefits for IT while addressing requirements for application performance by the end-user.

Use of redundant cloud architecture, redundant data center infrastructure and an innovative route optimization technology can provide the required performance and availability of applications. Additionally, use of technologies beyond the cloud edge such as TCP acceleration and content delivery networks can further improve content delivery to end-users.

About Internap

Internap is a leading Internet products and services company providing *The Ultimate Online Experience®* by managing, delivering and distributing applications and content with 100 percent performance and reliability. With a global platform of data centers, managed Internet services and a content delivery network (CDN), Internap frees its customers to innovate, improve service levels, and lower the cost of IT operations. Thousands of companies across the globe trust Internap to help them achieve their Internet business goals.

To learn more about Internap, visit us online at www.internap.com or call 877.843.7627 to speak to a representative.