White Paper

interxion™

TRUTH AND LIES ABOUT LATENCY IN THE CLOUD

An Interxion white paper
by **David Strom** and
Jelle Frank van der Zwet

When it comes to measuring applications performance across our local enterprise network, we think we know what network latency is and how to calculate it. But when we move these applications off premise and onto private or public infrastructures (the cloud) there are a lot of subtleties that can impact latency in ways that we don't immediately realise.

In this paper we will examine more closely what latency means for deploying and or migrating applications to the cloud and how you can both track and measure it. The goal is to both manage and reduce latency to ensure the best performance of your applications in the cloud. As the cloud can mean something different for everyone we refer to cloud as *cloud computing described on Wikipedia*.

For years, latency has bedevilled applications developers who have taken for granted that packets could easily traverse a local network with minimal delays. It didn't take long to realise the folly of this course of action: when it came time to deploy these applications across a widearea network, many applications broke down because of networking delays of tens or hundreds of milliseconds. But these lessons learned decades ago have been forgotten. Today we have a new generation of developers and networking engineers who have to understand a new set of latency delays across the Internet.

Many of the current generation of developers have never experienced anything other than high-speed Internet access and assume that it has always been that way. This tends to make for some sloppy coding decisions, creating unnecessary back-and-forth application communication that introduces increased latency times which impact running their applications. As we will see, as applications are migrating to the cloud, latency becomes even more important than ever before.

TRYING TO DEFINE CLOUD LATENCY IS NOT SIMPLE

In the days before the ubiquitous Internet, understanding latency was relatively simple. You looked at the number of router hops between you and your application, and the delays that the packets took to get from source to destination. For the most part, your corporation owned all of the intervening routers and the network delays remained fairly consistent and predictable.

Those days seem so quaint now, like when we look at one of the original DOS-based IBM dual-floppy drive PC's. With today's cloud applications, the latency calculations are not so simple.

First off, the **endpoints aren't fixed**. The users of our applications can be anywhere in the world, sitting on anything ranging from a high-speed fibre line in a densely served urban area or via a 3G mobile connection. With the flexibility that the cloud offers the applications themselves can also be located pretty much anywhere. That is the beauty and freedom of the cloud, but this flexibility comes at a price. The resulting latencies can be horrific and unpredictable.

We also need to **consider the location of the ultimate end users and the networks that connect them to the destination networks.** Furthermore we need to understand how the cloud infrastructure is configured, and where the particular pieces of network, applications, servers, and storage fabrics are deployed and how they are connected.

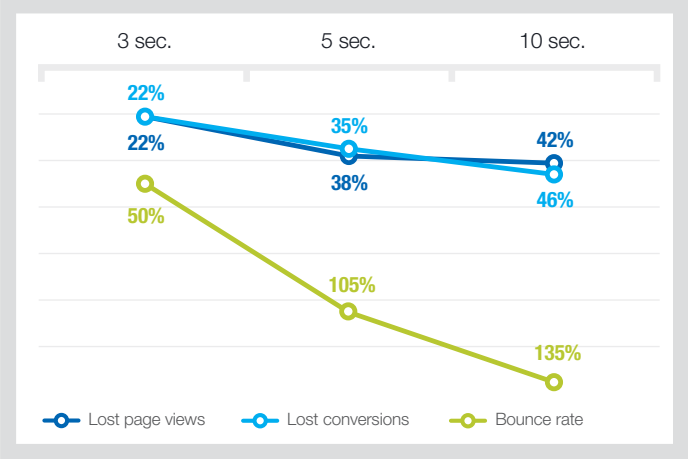
Finally, it depends on who the **ultimate "owners" and "users" of our applications** are. Latency can be important for the end-user experience of an enterprise's applications. If you are a service provider or system integrator, you will want to control the network and deliver the appropriate service levels to your customers, and that means also measuring and controlling the expected latencies as part of these agreements.

THE TRUE COST OF LATENCY CAN ADD UP

Now that we have some understanding of latency, we also need to understand the costs of it and how it impacts our business. There have been some studies that have examined overall website performance with respect to latency. For example, one study shows that reducing latency will have a tremendous effect on page load times, even more so than bandwidth. For example, *every drop of 20ms of network latency will result in a 7-15% decrease in page load times*.

This study isn't just an academic exercise: both Amazon and Google found big drops in sales and traffic when pages took longer to load. A half-second delay will cause a 20% drop in Google's traffic, and a tenth of a second delay can cause a drop in one percent of Amazon's sales. It's not just individual sites that have an interest in speeding up the Web.

Google has been working to try and make the web faster. The reason being that the entire experience needs to be lightning-quick and smooth to keep people happily using its many services. The graph shows that the consequences of delayed loading times can be severe in terms of lost page views or users leaving the site (see image below).



Granted, these studies just look at website pages rather than the complete applications server spectrum, but they give us a good starting point in terms of understanding how critical latency is when it comes to the cost of cloud computing.

ONE SOLUTION: TRIAGE YOUR APPLICATIONS

While reducing latency is desirable, not every app requires the lowest latencies. Certainly, we have gotten more demanding of our Internet performance as we distribute our applications throughout various cloud-based providers. Many businesses are extremely demanding and will continue to require the lowest latencies possible from their Internet connections. Applications such as algorithmic/high frequency trading, video streaming, more complex web/database services and 3-D engineering modelling are in this category. But applications such as email, analytics and some kinds of document management aren't as demanding. As a way down the path towards understanding latency, perhaps we need to start with some kind of triage and separating those applications that will really benefit from the lowest latencies.

We suggest preparing a chart such as the one shown opposite that classifies your applications according to their various properties of computing intensity, network bandwidth and latency requirements.

UNDERSTANDING THE TRUE EFFECT OF LATENCY

In the past, latency has had three different measures: roundtrip time (RTT), jitter and endpoint computational speed. Adding traceroutes as a tool, each of these is important to understanding the true effect of latency, and only after understanding each of these metrics can you get the full picture.

RTT measures the time it takes one packet to transit a network from source to destination and back to the source, or the time it takes for an initial server connection. This is useful in interactive applications, and also in examining app-to-app situations, such as measuring the way a Web server and database server interact and exchange data.

Jitter -Jitter is a variation in packet transit delay caused by queuing; contention and serialisation effects on the path through the network. This can have a large impact on interactive applications such as video or voice.

	Cloud Infrastructure Requirements		
Applications	**Compute Intensity**	**Network Bandwidth**	**Network Latency***
Testing & Development	High	Low	High
Web Browsing	Low	Low	High
Backup & Recovery	Low	High	High
Email & Calendar	Low	Low	High
HRM	Medium	Low	High
Doument Management	Low	High	Medium
CRM	Medium	Low	Medium
Finance & Accounting	High	Medium	Medium
ERP	High	Medium	Medium
Payment & Transactions	Medium	High	Medium
Virtual Desktops	Medium	Medium	Low
Network Storage	High	Medium	Low
Unified Communication	High	Medium	Low
Online Gaming	High	Medium	Low
HD Video Streaming	High	High	Low
Black Box (M2M) Trading	High	High	Proximity

**RTD (Round Trip Delay)*

The speed of the computers at the core of the application: their configuration will determine how quickly they can process the data. While this seems simple, it can be difficult to calculate once we start using cloud-based computer servers.

Finally traceroute is the name of a popular command that examines the individual hops or network routers that a packet takes to go from one place to another. Each hop can also introduce more or less latency. The path with the fewest and quickest hops may or may not correspond to what we would commonly think of as geographically the shortest link. For example, the lowest latency and fastest path between a computer in Singapore and one in Sydney Australia might go through San Francisco.

Let's keep each of these in mind as we look at how the cloud plays into each calculation. As we will see, they aren't the only ones that we need to consider when it comes to the cloud.

FIRST COMPLICATING FACTOR: DISTRIBUTED COMPUTING

As we said earlier, in the days when everything was contained inside an enterprise data centre, it was easier to locate bottlenecks because the enterprise owned the entire infrastructure from source to destination. But with the rise of Big Data applications built using tools such as *Hadoop* and *R* (the major open source statistics language used for data analytics), the nature of applications has changed and is a lot more distributed. These applications employ tens or even thousands of compute servers that may be located all over the world, and have varying degrees of latency with each of their internet connections. And depending on when these applications are running, the latencies can be better or worse as other Internet traffic waxes or wanes to compete for the same infrastructure and bandwidth.

VIRTUALISATION ADDS ANOTHER LAYER OF COMPLEXITY, TOO

Today's modern data centre isn't just a bunch of rackmounted servers but a complex web of hypervisors running dozens of virtual machines. This introduces yet another layer of complexity, since the virtualised network infrastructure can introduce its own series of packet delays before any data even leaves the rack itself!

But we also have to add to this the delays in running virtualised desktops, if that is relevant to our situation. Many corporations have begun deploying these situations and that introduces yet another source of latency. If not designed properly *(see this reference here)*, you can have tremendous delays with just logging into the network, let alone running your applications on these desktops.

IT ISN'T JUST ABOUT PING, EITHER

One of the major problems with measuring latency and using tools such as traceroute is that applications don't usually use the same protocols as ICMP (Internet Control Message Protocol). In fact, most don't have anything to do with ICMP, the protocol behind ping and traceroutes. *But most modern applications and networks don't give much priority to ping packets*, and so the best way to measure and understand latency is to measure the performance of the same protocols that your applications use: HTTP, FTP and so forth. This brings us to our next layer of complexity: QoS (Quality of Service).

UNDERSTAND QUALITY OF SERVICE AND WHAT TRAFFIC IS PRIORITISED

In the pre-cloud days, Service Level Agreements (SLA's) and QoS were created to prioritise traffic and to make sure that latency-sensitive applications would have the network resources to run properly. These agreements were also put in place to ensure minimal downtime by penalising the ISP's and other vendors who supplied the bandwidth and the computing resources.

But with the rise of more cloud and virtualised services, it isn't so cut and dried. For one thing, the older SLA's typically didn't differentiate between an outage in a server, a network card, a piece of the storage infrastructure, or a security exploit. But these different pieces are part and parcel to the smooth and continuous operation of any cloud infrastructure.

Just as different applications have different tolerances for network latency, the same can be said when it comes to downtime. "Applications with varying uptime requirements also commonly reside at different companies or within a company. Some applications may be less critical to the business and can tolerate lower uptime in exchange for lower cost. Other applications cannot tolerate any amount of downtime without resulting in a significant impact on the business," said Reed Smith, a Savvis director *writing in the ComputerWorld UK CloudVision blog*.

An example of this is a back office application that produces daily summary charts about a particular business process. If one of the many components of this app is down briefly, no one would probably notice nor would they really care, as long as the reports are produced eventually. We've put together the chart opposite that summarises our thoughts on how critical particular applications are and under what circumstances they should be prioritised for particular SLA's. This means that your SLA's need to handle a variety of situations. You don't want to enforce (nor pay for) the same service levels on your test & development cloud that you would on a production cloud.

CLOUD AND APPLICATIONS MAPPED ON PERFORMANCE AND AVAILABILITY

MANY CLOUD PROVIDERS ARE CAGEY WHEN IT COMES TO DESCRIBING THEIR INFRASTRUCTURE

A final complicating factor is in trying to track down the actual infrastructure that is running your cloud applications. Some cloud providers, such as Amazon, don't even want to tell you exactly where their cloud data centres are located (we know that one of their data centres is in Dublin Ireland, *but not much more than that for example*). Let alone what kind of gear is in place in their data centre and what kinds of connection paths are used to hook them up. But to really understand latency, you should know answers to questions such as:

- Are your VM's stored on different SAN's or different hypervisors, for example?
- Do you have any say in decisions that will impact your own latency?
- How many router hops are in your cloud provider's internal network and what bandwidth is used in their own infrastructure?

REDUCING LATENCY HAS SEVERAL DIMENSIONS

So now that we have a better understanding of some of the complicating factors, the next step is to start to examine how you can reduce latencies in particular segments of your computing infrastructure. *In a paper for Arista Networks*, they mention four broad areas of focus:

- Reduce latency of each network node
- Reduce number of network nodes needed to traverse from one stage to another
- Eliminate network congestion
- Reduce transport protocol latency

Of course, they sell some of the gear that can help you reduce network switch transit times or cut network congestion, but still it is worth examining these more mundane pieces of your cloud provider's network infrastructure (if you can) to see where you can start to apply some of these savings.

CAN A CONTENT DELIVERY NETWORK (CDN) HELP?

Not much. CDN's are designed mostly for delivering static content to a broad collection of distributed end users. One of the largest CDN's is *Akamai,* which is based on 95,000 servers in 71 countries within nearly 1,900 networks around the world. But many cloud applications have a different type of treatment, and in many cases won't get much of a latency improvement from a CDN because they aren't using static pieces of content. Nevertheless, CDN's are expanding their capabilities and trying to help reduce latencies by caching more than just static HTML pages. Certainly, it is worth investigating whether a CDN partner can improve your particular situation.

ONE IMPORTANT TOOL: CLOUDSLEUTH

To help cloud-based application developers understand some of these issues, Compuware created its *Cloud Sleuth benchmark*. They have attempted to mirror real-world app conditions and set up a series of servers running simultaneous transactions from 32 locations around the world. They make use of standard app servers from Apache, Tomcat, Microsoft and Google to build an ecommerce app. They then measure the response time of the application hosted at 17 cloud providers. A sample analysis that is produced from them looks like the image above.

This is a good first cut at seeing what is going on across these various Infrastructures as Service (IaaS) providers at a particular moment in time. And while the tool can be used to provide some trends, it still is just an estimate and your actual performance will still need careful and closer examination.

OTHER SOLUTIONS FOR IMPROVING CLOUD LATENCIES: PROVIDING MORE CONSISTENT NETWORK CONNECTIONS

As you can see, it isn't just poor latency in the cloud but the unpredictable nature of the various network connections between your on-premises applications and your cloud provider that can cause problems. What is needed is some way to reduce these daily or even minuteby- minute variations so you can have a better handle on what to expect.

Clearly, the best available option is to directly connect to a public cloud platform. One way that Amazon has managed to provide a more consistent network experience than Internet-based connections, is through their product called Direct Connect. This is where a dedicated network connection can be established between your network and one of the Amazon Direct Connect locations, rather than having to send all traffic over the public Internet. In a nutshell, the key benefit is that traffic is no longer subject to the unpredictability of the general Internet, and so basic metrics such as bandwidth and latency will be far more predictable.

For the connection from the customer's premises to the Direct Connect locations these metrics will be subject to strict quality of service guarantees, i.e. a bandwidth of X with a defined maximum latency and an SLA for the connection. For the connection from the Direct Connect locations to Amazon Web Services cluster in your region, you can expect improved network characteristics but there is not an SLA that defines guaranteed bandwidth.

Direct connect options like Amazon, but also Windows Azure, offer to build a hybrid solution that uses both on-premise and cloud-based resources. Application code and data can be stored in an appropriate on- premise location according to regulations, privacy concerns, and a measurement of acceptable risk, while solution components requiring the features and pricing model of cloud computing can be migrated to the cloud.

CONCLUSION

Relying on the internet for application connectivity in the cloud introduces a degree of variability and uncertainty around bandwidth, speed and latency. This can be unacceptable to many large and medium sized enterprises, which are increasingly putting the emphasis on end-to-end quality of service management. Using dedicated connectivity to cloud providers overcomes this and hooking up via carrier neutral data centres and connectivity providers can also have benefits in terms of costs, monitoring, troubleshooting and support.

As you can see, cloud latency isn't just about doing traceroutes and reducing router hops. It has several dimensions and complicating factors. Latency in itself does not have to be an issue; it's the unpredictability of latency that really causes the problems.

Hopefully, we have given you some food for thought and provided some direction so that you can explore some of the specific issues with measuring and reducing latencies for your own cloud applications, along with some ideas on how you can better architect your own applications and networks.

For more information about Interxion Cloud Hubs, visit www.interxion.com/cloud

AUTHOR BIOGRAPHIES

David Strom is a leading expert on networking and Internet communications. He has written two books and thousands of articles on the topic over the past 25 years for most of the major IT publications and appeared on numerous American TV and radio programs as well. He was the founding editor-in-chief of Network Computing magazine and has managed dozens of Web-based editorial sites for computing enthusiasts, IT managers, VARs, electronics engineers, and network applications developers.

Jelle Frank (JF) van der Zwet has more than 14 years of experience as a B2B marketer, working on ICT product and business development. As Global Marketing Manager for Interxion he is primarily responsible for the company's continually growing cloud community, and is a frequent speaker at technology events on cloud computing. He has published articles on TechCrunch and Wired and written white papers on latency, hybrid clouds and application migration. Prior to joining Interxion he held senior marketing and product management roles with Imtech, UPC, KPN and Schiphol Airport.

ABOUT INTERXION

Interxion (NYSE: INXN) is a leading provider of carrier and cloud-neutral colocation data centre services in Europe, serving a wide range of customers through over 35 data centres in 11 European countries. Interxion's uniformly designed, energy efficient data centres offer customers extensive security and uptime for their mission-critical applications. With over 500 connectivity providers, 20 European Internet exchanges, and most leading cloud and digital media platforms across its footprint, Interxion has created connectivity, cloud, content and finance hubs that foster growing customer communities of interest. **For more information, visit** www.interxion.com

INDUSTRY ASSOCIATIONS

Cofounder: Uptime Institute EMEA chapter

Founding member: European Data Centre Association

Patron: European Internet Exchange Association

Member: The Green Grid, with role on Advisory Council and Technical Committee

Contributor: EC Joint Research Centre on Sustainability

Member: EuroCloud

ACCREDITATIONS

ISO 22301 Business Continuity Management

BCMS 560099

ISO/IEC 27001 Information Security Management

IS 537141

ITILv3-certified Service Centre Members and Facilities Managers

interxion™

www.interxion.com

INTERNATIONAL HEADQUARTERS
Main: + 44 207 375 7070
Fax: + 44 207 375 7059
E-mail: hq.info@interxion.com

EUROPEAN CUSTOMER SERVICE CENTRE (ECSC)
Toll free from Europe: + 800 00 999 222
Toll free from the US: 185 55 999 222
E-mail: customer.services@interxion.com

WP-CL-HQ-TRUTH-HQ-eng-7/15