Level 3 COMMUNICATIONS IS NOW CenturyLink®

# WHITE PAPER

## ENHANCING CLOUD CONNECTIVITY FOR OPTIMAL APPLICATION PERFORMANCE

## EXECUTIVE SUMMARY

Increasing dependence on Cloud-based applications hosted in distant data centers makes the quality of backbone networks connecting them increasingly critical. Network quality between Cloud providers and end-users in fact widely varies, and chronically threatens application performance. Differences in carrier network design become more significant as dependence on the Cloud intensifies.

As customers pay increased attention to Cloud application performance, the effects of network throughput issues such as latency, packet loss, routing across multiple carriers, security and weaknesses of typical enterprise network SLAs in regard to Cloud-based performance have not been sufficiently appreciated.

The pressures of the Cloud revolution are exposing meaningful differences among carrier networks. Longstanding SLAs are often no longer adequate. Customers should look for carriers with global reach, deep metro network penetration, minimum inter-city latency and maximum edge router coverage.

Network managers can take measures to bolster network and thus application performance in the Cloud. These include choosing carriers and network solutions that optimize network performance, and actively consulting with their service providers to develop the strongest combination of strategies that mitigate degrading effects of the network on application performance.

**A crucial problem, as Cloud services and applications proliferate, is that the quality of network connectivity among resources and end users varies widely.**

## INTRODUCTION

High-performance websites can deliver a wide range of The Cloud is a focus of today's communications landscape and is becoming central to enterprise computing in the U.S. and around the world. A crucial problem, as Cloud services and applications proliferate, is that the quality of network connectivity among resources and end users varies widely. This performance variability becomes ever more critical as enterprises elevate their dependence on Cloud connectivity for multiple, crucial applications provided to large numbers of end-users, often across great distances.

Cloud deployments take many forms. The most common deployment model is the "Public" Cloud, in which a theoretically unlimited number of customers access third-party resources over the public Internet. Other models that have formed include: "Private" Clouds serving individual organizations; "Hybrid" Clouds — varied combinations of Public and Private; and "Community" Clouds — a solution in which several organizations related by common interests share infrastructure.

## CONNECTIVITY METHODS AND PERFORMANCE VARIABILITY

Most critical from the networking perspective are connectivity options between users and Cloud-based resources. Although the Internet is the normative mode of Public Cloud connectivity, alternate connectivity options for Private Cloud include:

- Layer 3 MPLS virtual networks
- Layer 2 Ethernet virtual networks
- Point-to-point (usually optical) services

There is major network performance variability between and within connectivity options. What the best connectivity choice is depends on application requirements and the criticality of applications to the business.

### PUBLIC CLOUD INTERNET CONNECTIVITY

The Internet offers the easiest and least costly access. However, the Internet also offers the least reliable network performance, commonly referred to as "best effort." Internet performance will inevitably demonstrate major variability, specifically in network latency and packet loss, so those using the Internet for Cloud should expect service delivery problems. IT managers recognize that, in the face of "heavy-duty" application requirements, they need more reliable network options, which might include dedicated data connections from their data centers directly into provider facilities.

### PRIVATE CLOUD CONNECTIVITY OPTIONS

**Virtual Private Network (VPN) Connections:** The next level of performance, more reliable than the Internet, is provided by virtual private network (VPN) connections in which customers buy MPLS-based IP VPNs or Ethernet-based VPLS services. This kind of connection offers assurances that whatever is connected to the VPN will stay on that network. Unlike on the "best effort" Internet, most leading carriers will provide service level agreement (SLA) performance guarantees specifying maximum latency and packet loss levels between network nodes. This bounding of network performance provides much improved control over customer Cloud service experience.

**Optical/Point-to-Point Connections:** Optical point-to-point dedicated circuits provide the lowest latency and highest reliability of the three access choices, along with the potential for complete prevention of packet loss. This service category includes private lines, Ethernet private lines (EPL) and optical wavelengths (DWDM), whether using Ethernet, SONET, Fibre channel or other interfaces.

### CONNECTIVITY: THE CHOICE

The connectivity choice comes down largely to a cost/quality trade-off. Public IP Internet access for Cloud connectivity can be fine for non-latency-sensitive applications like email, or where security is less of a concern. Virtual networks, whether Ethernet or MPLS-based, can be preferable for applications with a medium degree of latency sensitivity, potentially including networks between customer data centers and Cloud service providers. Virtual networks also offer a higher level of security as services are delivered over the service provider's private MPLS network. However, dedicated optical point-to-point services are best — albeit also typically most costly — for the most mission-critical applications.

The table below illustrates some of the differences between application types in their demands and sensitivities to various dimensions of network quality.

**When on-net connections are deployed, the relationship between the client's network and the carrier's network is greatly simplified.**

Performance Characteristics of Common Business Application Types

| Application | Required BW | Sensitivity to Errors | Latency Sensitivity | Jitter Sensitivity | Burstiness |
|---|---|---|---|---|---|
| Messaging e-mail | Very Low | Low | Low | Low | Medium |
| Voice (TDM) | Low | Low | Low | High | Low |
| Voice over IP (VoIP) | Low | Medium | Low | High | Low |
| Web Browsing (non-critical) | Medium | Medium | Medium | Medium | Medium |
| Web Browsing - SaaS | High | Medium | High | Medium | Medium |
| Video Conferencing | High | Medium | Medium | High | High |
| Telepresence | Very High | High | High | High | High |
| Remote Workers | Medium | Medium | High | Medium | High |
| Streaming Media | High | High | Medium | High | High |
| Storage Area Networks | Very High | High | High | High | High |
| Server Virtualization (WAN) | High | High | High | High | High |
| Unified Communications | Medium | Medium | Medium | Medium | High |

Source: Level 3 Communications

## CLOUD CONNECTIVITY AND APPLICATION PERFORMANCE

Regardless of which Private Cloud connection customers choose, it's important to recognize that there are significant differences between carrier networks and the network's effect on Cloud-based application performance.

Escalating dependence on distant data centers makes the quality of wide-area backbone networks connecting end-users to applications critical. Small differences in network performance can make large differences in application performance. These networks need to handle the considerable tasks of providing huge amounts of bandwidth with low latency and minimal or no packet loss to large numbers of end-users, often across great distances, virtually in real-time.

Managing workload and data between multiple data centers typically means virtually instant transfer of huge amounts of data across distances, with reachability often required between multiple data stores and a much larger number of locations than ever before.

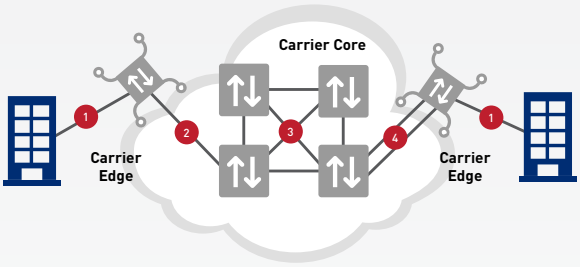
For example, server virtualization means virtual machine mobility becomes a severely complicating factor, itself strongly affected by latency, packet loss and bandwidth throughput. Data replication and business continuity/disaster recovery (BCDR) strategies also become more complex in such environments.

Additionally, the demands for bandwidth can be dynamic, with peaks and valleys, or short-term or long-term requirements. For example, nightly data replication between data centers may require more bandwidth than daily user interactions with the associated applications. Or, a virtual machine migration may require a dramatic, one-time burst in bandwidth to accommodate the redistribution of processing resources. Such varying demands arise prominently between common Cloud-related activities as migration of data storage and virtual machines, data storage updates, and virtual applications traffic between virtual machines.

**A virtual machine migration may require a dramatic, one-time burst in bandwidth to accommodate the redistribution of processing resources.**

Source: Level 3

| Inter-Data Center Traffic Type | Flow Duration | Bandwidth per flow | QoS Sensitivity |
|---|---|---|---|
| Data Storage Migration | Medium | Very High | Med–High |
| Virtual Machine Migration | Short | High | Med–High |
| Data Storage Update (Active-Active) | Long | Medium | Very High |
| Distributed vApp Inter-VM traffic | Varies | Low | Varying |




## NETWORK ELEMENTS AND LATENCY

Back-haul mileage: As we've discussed, the physical distance between the end users and applications impacts the amount of time it takes for interaction. This includes the distance from customer premises to carrier edge and the distance the data travels across carrier's (or multiple carriers') network. That distance is not necessarily the shortest path, since carriers may opt for a less-expensive route in lieu of the shortest, an approach known as "hairpinning." Furthermore, the greater the number of carrier edge switches that are available, the smaller the backhaul mileage.

**Inter-switch latency:** The carrier's fiber routing between network switches matters. In order to optimize network performance, the distance between switching centers needs to be as short as possible. Some carrier networks are purposefully architected to have shorter inter-switch paths than others.

**Carrier switching architecture and hierarchy:** The way a network has been designed, which includes decisions about the number and location of edge switches impacts network performance. Intra-network latency falls as the number of edge switches increases.

**Diverse route choices:** Creating multiple paths between users and applications helps ensure continuity. However, diverse routes can have different latency characteristics, which can create challenges when it comes to applications compensating for that variance. It's critical that the provider both has diverse routing options and anticipates latency variation when architecting the user's network.

**Note:** These concepts are discussed in more depth in the following portion of this whitepaper.

**So, while network performance has always been important, the Cloud raises the intensity of these issues to a degree for which network managers may be unready.**

## THE CLOUD'S INFLUENCE: THE "NEW ABNORMAL"

The Cloud adds new challenges and complexities to an already problematic, fast-changing, multi-faceted networking environment. It multiplies the sheer volume of data sent over networks, the distance applications must work across and process complexity.

The more data moved across a network, the more important latency and packet loss becomes. Now, instead of loading software on local hard drives or sharing it over local area networks, software is hosted hundreds or thousands of miles away. Customers seek to access these remotely hosted services on a wide-scale basis in virtually "real time." So, while network performance has always been important, the Cloud raises the intensity of these issues to a degree for which network managers may be unready.

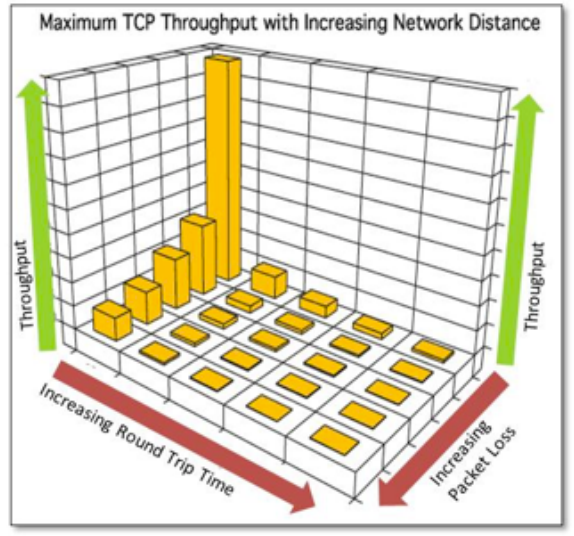
## WHAT USERS MAY OR MAY NOT KNOW

There are profound differences in how competing carrier networks are designed that become increasingly critical as dependence on the Cloud intensifies, and which current standard SLA guarantees may not address. Ongoing connectivity issues affecting backbone and so application performance gain new urgency and complexity. Discussion of these issues follows.

### Latency Building

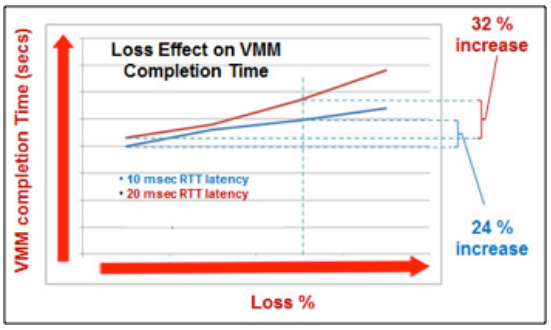
Latency builds with distance and with each piece of equipment on the network. Multiple, seemingly tiny differences — in carrier locations, in how far out networks are pushed, in latency on the backbone that ties customer locations together — add up. When connecting end user locations on a VPN, for example, traffic normally must run through optical multiplexers and connect into a switch fabric that runs its logic to route the packets. The computation time and retransmission process involved introduces multiple, small increments of latency that build and persist across the network.

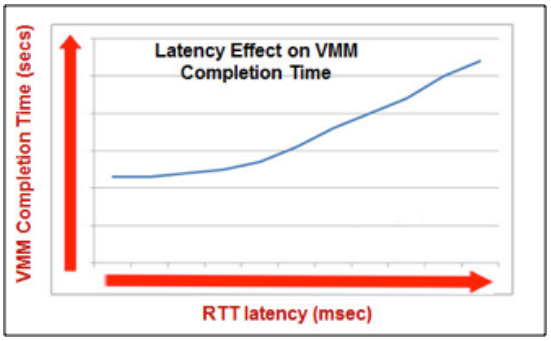
Figure 1.




Source: Level 3 Communications

Figure 2.




Source: Level 3 Communications

Problems can arise with an application when users buy "compute" Cloud services and need to move huge chunks of data.

### LATENCY + PACKET LOSS REDUCE THROUGHPUT

Latency and packet loss can radically reduce effective data throughput between customer data centers and Cloud service providers. Sometimes this will not matter, as when data flow is small. But it can be of utmost importance when users have major bandwidth requirements. Problems can arise with an application when users buy "compute" Cloud services and need to move huge chunks of data. Clearly, significantly more data can pass when round-trip latency is 10 milliseconds rather than 30, and switching systems don't shed data packets in an effort to maximize throughput.

### JITTER AND THROUGHPUT

The TCP protocol is designed to adjust to latency variations, also known as jitter, by staying at "lower common denominator" throughput levels. While this may provide a stable throughput level for the TCP protocol, that level may not be sufficient for mission critical applications.

For example, if latency on a connection starts at 15 milliseconds at one moment and changes to 30 milliseconds the next, a logical assumption may be that throughput will adjust back to its earlier, better-performing level when latency again drops. The way TCP actually operates, however, is that the network adjusts to jitter (latency variance) by performing consistently at the level dictated by the higher (worse) latency, the base level it can "count on," delivering less bandwidth than anticipated. This exacerbates inherent problems with using the Internet for Cloud connectivity (see Figure 1 for a conceptual illustration of this concept).

### SMALL LATENCY DIFFERENCES = MAJOR NETWORK DEGRADATION

Network effects are disproportionate. A round-trip delay increase between two nodes of just over 10 milliseconds, for example, can cause network throughput degradation of 25 to 50 percent. Latency differences among carrier networks, especially on long-haul routes, often actually are dramatically larger. This issue is illustrated below by the effect of latency on virtual machine migration (VMM), a basic Cloud activity that is highly latency-sensitive.

### Latency & the Carrier Edge

Proximity of a carrier network's VPN "edge" to end user facilities fosters improved latency performance. The more edge switches a carrier has, the more dispersed the switches can be around a metropolitan area, meaning that when enterprise customers connect to the edge of that provider's metro network, less back-haul distance must be traversed between the customer's VPN and the carrier's switching infrastructure.

### Routing Traffic Between Providers

Passing customer traffic between carriers brings an expected reduction in network performance over time, in the face of ongoing risks of peering congestion. There is particularly high-risk when involving providers with lower-quality connectivity that may "run hot" to save money. Often without recognizing this issue, enterprise customers sometimes buy lower-cost Internet service from resellers that traverse three, four or five separate Internet switching fabrics. Of additional concern, when multiple service providers are involved, is the difficulty of rapidly locating, diagnosing and solving network problems.

In contrast, there are important advantages working with a carrier that can provide the most comprehensive geographic coverage, and so has less need to "hop" traffic between networks or "point fingers" toward other carriers responsible for disparate network portions.

**For customers increasingly using networks to support mission-critical Cloud applications, current standards are not universally meeting expectations.Latency & the Carrier Edge**

Proximity of a carrier network's VPN "edge" to end user facilities fosters improved latency performance. The more edge switches a carrier has, the more dispersed the switches can be around a metropolitan area, meaning that when enterprise customers connect to the edge of that provider's metro network, less back-haul distance must be traversed between the customer's VPN and the carrier's switching infrastructure.

### ROUTING TRAFFIC BETWEEN PROVIDERS

Passing customer traffic between carriers brings an expected reduction in network performance over time, in the face of ongoing risks of peering congestion. There is particularly high-risk when involving providers with lower-quality connectivity that may "run hot" to save money. Often without recognizing this issue, enterprise customers sometimes buy lower-cost Internet service from resellers that traverse three, four or five separate Internet switching fabrics. Of additional concern, when multiple service providers are involved, is the difficulty of rapidly locating, diagnosing and solving network problems.

In contrast, there are important advantages working with a carrier that can provide the most comprehensive geographic coverage, and so has less need to "hop" traffic between networks or "point fingers" toward other carriers responsible for disparate network portions.

### SLAS & NETWORK PERFORMANCE

Service providers have arrived over time at de facto market standards for SLAs, which typically are quantified in maximum allowable packet loss and latency. A problem is that once safely within the SLA's standards, providers may optimize for cost rather than performance. So a standard IP-VPN guarantee of 40 milliseconds round-trip latency domestically edge-to-edge within a carrier's U.S. network may gravitate from minimum standard to de facto norm — at a level of greater latency and thus lower performance — than is optimal for certain applications. A 10-millisecond latency advantage can be very important in Cloud application performance, so this is a real problem. For customers increasingly using networks to support mission-critical Cloud applications, current standards are not universally meeting expectations.

## WHAT IS TO BE DONE TO IMPROVE NETWORK/APPLICATION PERFORMANCE?

What steps should network managers be taking to ensure the enterprise networks they are responsible for are appropriately tuned for Cloud applications?

First, consider Cloud backbone options in relation to applications requirements as well as overall cost. Assess Public, Private and Hybrid Cloud solutions, for which options are most appropriate to the demands of your applications. Consider the ideal types of connectivity for these. Evaluate how much packet loss, latency and jitter are tolerable for particular applications. Also, take into account related issues of dedicated vs. shared infrastructure in terms of economics, security and other considerations.

Once these choices have been made, some prominent paths to improving network and application performance include:

Front-end optimization: Including reducing numbers and sizes of files (video, graphics, texts, etc.) sent, as well as using TCP optimization and devices to prevent frequent window re-set.

TCP pooling: Better enabling multiple communications streams to share common paths rather than many individual ones.

TCP connection persistence: By maintaining connections after communication sessions have temporarily stopped, in the generally correct expectation two end-points will soon resume communication, repetitive short-term shrinkage of TCP windows can be avoided or reduced.

**Customers need to sit down with their chosen provider to think through how, together, they will further mitigate network factors that degrade application performance.**

Altering TCP's "slow start" mechanism: Better accommodating host (receiving machine) buffers, improving consistency between TCP and the latter. Manually adjusting TCP window size is also an option; older operating systems have default 64 kilobyte window sizes, too small for long-haul high-capacity transmission.

Writing or re-writing applications: Revising software to optimize performance.

Application Performance Management: Service and tools are readily available that monitor customer traffic information and upload performance information into a centralized server; customers can apply analytics to better understand their networks and applications.

Implement WAN acceleration techniques: These include data reduction and compression to QoS, as well as application-specific acceleration to prioritize applications. Here again, there are tools and services readily available in the market.

Carefully choosing carrier network services.

After choosing carrier backbone networks for optimal service performance, customers need to sit down with their chosen provider to think through how, together, they will further mitigate network factors that degrade application performance. It will be increasingly incumbent on network managers to secure expertise to help design and optimize networks that enable productivity and meet business requirements.

## WHAT CUSTOMERS SHOULD LOOK FOR IN CARRIER NETWORKS

The Cloud service revolution is exposing meaningful differences between carrier networks. Latency is more critical than ever, with variance tolerances now measured in single digits of milliseconds. There is near zero-tolerance for packet loss, making the Internet potentially unusable for "industrial-strength" Public Cloud services. Important elements customers should focus on when comparing carriers' network capabilities include:

Global reach: The global distribution of computing and content will become increasingly critical as customers expect to be serviced globally.

Inter-city latency: Customers should seek carriers with the lowest average latency between key city-pair combinations.

Deep metro penetration: The ability to provide direct fiber connectivity to crucial data centers within metros will be essential for optimal customer experience and performance transparency.

Edge router coverage: The maximum number of edge switches (as well as their dispersion across metro markets and proximity to customer facilities) provides the most efficient routing, resulting in the best network performance.

Transparency in the relationship: Although Cloud connectivity networking can be complex, the provider should be willing to candidly discuss the solution, including showing the infrastructure architecture.

## CONCLUSION

As enterprises shift more and more business processes to the Cloud, the quality of wide-area network connectivity with distant data centers — and its resulting effects on application performance — becomes increasingly important. This quality currently widely varies.

**Without appropriate support and tools to understand and manage traffic flow, it is very difficult to make the adjustments needed to improve application performance.**

The accumulation and interaction of network performance problems have an intensifying effect on application performance, due to problems such as latency build-up over networks, the effects of distance-driven latency, packet loss, bandwidth variability and multiple-provider network data hand-off.

The Cloud adds a stark new level of complexity to the networking scene. The network manager's job inevitably becomes more challenging as new Cloud applications emerge. Adding to the complexity is need for "real-time" application performance for huge numbers of end-users, in multiple locations, across vast distances. Without appropriate support and tools to understand and manage traffic flow, it is very difficult to make the adjustments needed to improve application performance.

Network managers need to take a wide variety of measures and strongly consider the qualities of the carriers and carrier networks they use. They need to work closely with their carrier, leveraging the provider's capabilities and expertise to optimize network performance. Customers should look for carriers with global reach, deep metro network penetration, minimum inter-city latency, maximum edge router coverage and comprehensive managed and professional services.

## THE LEVEL 3 POSITION

The Cloud revolution exposes clear, meaningful differences between carrier networks. Level 3 offers important differentiation in connectivity options. Some of Level 3's network strengths are inherent and some result from a thoughtful, purpose-built portfolio of professional service and management solutions. A commitment to delivering a quality customer experience is also part of Level 3's solution. As such, transparency in network design, operation and optimization is foremost in the relationship.

Level 3 has built into its network ecosystem a powerful capability to advise customers (before they buy) as to the latency they might expect on specific routes and to provide appropriate SLAs. Latency is measured at circuit turn-up with this information stored in Level 3's provisioning system and the circuit flagged so operations staff knows its latency is paramount and the circuit cannot be moved, groomed or reconfigured without customer consultation. This system was initially built for low-latency financial industry trading applications. With advent of the Cloud, Level 3 extended it to the rapidly-growing and increasingly universal set of enterprise customers with intensifying latency concerns.

Level 3 has 116 U.S. metro markets and its own long-haul optical backbone connecting them with 100,000 U.S. fiber-miles, giving it the ability to keep traffic on its own network and control performance end-to-end. Along with Level 3's number of edge switches, the company offers competitive latency between many U.S. city pair combinations.

## ABOUT LEVEL 3

We build, operate and take end-to-end responsibility for the network solutions that connect you to the world. We put customers first and take ownership of reliability and security across our broad portfolio.

**1.877.2LEVEL3**
**info@level3.com**
**level3.com**

DATA NETWORKS | SECURITY | CONTENT DISTRIBUTION | DATA CENTERS | APPLICATION PERFORMANCE | VOICE | UCC

© 2014 Level 3 Communications, LLC. All Rights Reserved. Level 3, Level 3 Communications, and the Level 3 Logo are either registered service marks or service marks of Level 3 Communications, LLC and/or one of its Affiliates in the United States and/or other countries. Level 3 services are provided by wholly owned subsidiaries of Level 3 Communications, Inc. Any other service names, product names, company names or logos included herein are the trademarks or service marks of their respective owners.