Independent market research and competitive analysis of next-generation business and technology solutions for service providers and vendors

HEAVY READING
WHITE PAPER

The Benefits of Edge Cloud On-Ramps for CSPs & End Users

A Heavy Reading white paper produced for EdgeConneX

AUTHOR: ALAN BREZNICK, CONTRIBUTING ANALYST, HEAVY READING

INTRODUCTION: THE DIFFICULTY WITH CLOUD ACCESS

Enterprise applications are moving to the cloud, and for good reason: The cloud offers a host of potential benefits for both small & mid-sized businesses and large enterprises. These benefits include lower costs, scalable and flexible deployment options, lower capital spending, and access to innovative applications. Communications Service Providers (CSPs), meanwhile, see more opportunities to offer more bandwidth to businesses within their footprint.

Despite these benefits, many businesses are hesitant about adopting cloud services, especially when it comes to the public cloud. Continuing concerns about performance (especially with performance-intensive SaaS applications) and security are contributing to this reticence. Many companies and organizations simply do not consider the Internet to be reliable enough on its own to deliver latency-sensitive, cloud-hosted solutions such as SaaS applications, interactive databases or virtual desktops. This concern will only grow as businesses look to implement services relying on next-generation capabilities such as the Internet of Things (IoT), augmented/virtual reality (AR/VR) and AI intensive applications.

To cite one recent study, research firm Ovum reported that latency is "rising in importance as a buying criterion in the high-bandwidth portion of the market." In general, Ovum found, enterprise and wholesale clients wish to better understand latency performance on both primary and failover routes before they make a major commitment to cloud usage. They conclude that "applications need to perform as if local" to raise adoption. This is buttressed by data from 451 Research indicating performance and availability are key factors hindering public cloud adoption *and* causing workloads to be shifted from the public cloud.

As long as these performance issues linger, enterprise IT departments will remain rightfully concerned about putting their critical applications at risk for their organizations. Thus, they will remain either unable or unwilling to leverage all that the cloud has to offer.

Fortunately, though, the outlook for the cloud may not be so… well, *cloudy* after all. There are well-defined steps that service providers, enterprises and others can take to realize key benefits of direct cloud access, while mitigating or avoiding some of the major risks. Simply put: there's a better way to tap into the cloud using local edge on-ramps.

One such ecosystem is already developing in Portland, Oregon, where EdgeConneX, a provider of edge colocation facilities, has deployed both a native cloud on-ramp provided by Amazon Web Services (AWS) and software-defined on-ramps from Megaport and PacketFabric. These partnerships also help local CSPs that have infrastructure inside the Edge Data Center and can thus offer connectivity from the business customer to the cloud on-ramp.

This ecosystem brings the cloud to "the edge," enabling the local-like performance and security that is in such high demand. Upon becoming part of such an edge ecosystem, CSPs can further penetrate the business market. Their customers, meanwhile, can overcome key latency, cost and security obstacles and enjoy the full benefits of the cloud. *As this white paper will demonstrate, these companies can see an average performance improvement of 45 percent to 85 percent, depending on their cloud and connectivity architecture.*

This paper will look at the public cloud's performance, security and cost issues and examine what can be done to resolve them, focusing especially on the impact of latency and the benefits for the growing SaaS segment. It will also explain the cost, performance, security and other benefits of leveraging edge cloud on-ramps to access the hyperscale public cloud providers, and examine how CSPs can use this local infra-structure to promote their own connectivity infrastructure.

THE SOLUTION: CLOUD CONNECTION FABRICS AT THE EDGE

As noted in the introduction, there is a way out of the performance and security dilemma for companies that still want to benefit from the cloud. The solution recommended in this paper calls for using a private connection, otherwise known as a dedicated cloud on-ramp, to reach the cloud, thereby skipping the public Internet and avoiding the performance, security and other problems that the Internet frequently poses.

The idea of such a cloud on-ramp is actually nothing new. In fact, cloud providers have offered these dedicated private connections to customers in a handful of core Internet markets for years. In these markets, businesses at the network edge can use costly long-haul fiber connections to reach legacy core markets.

What is new here is that both large and small companies can now take advantage of similar types of cloud on-ramps in Tier 2 and 3 local markets. Commercial customers can now establish local connectivity using either a dedicated metro-area fiber link or a carrier transport connection (e.g., a metro-Ethernet connection directly from the enterprise office to a data center with a cloud on-ramp). In other words, businesses no longer require a costly long-haul transport connection to gain private access to the nearest cloud on-ramp, which may reside in a distant Tier 1 market. Instead, they can now turn to either:

- A local native on-ramp node (e.g., an AWS deployment sitting inside a local Edge Data Center, connected to the CSP's network, or
- An SDN-based solution that offers virtual cloud connections locally, such as Megaport and Packet Fabric, available within their local area.

These solutions are helping to close the gap with the major markets by providing private cloud on-ramp connections at the network edge.

Native cloud on-ramps at the network edge show the greatest performance improvement potential, as they connect directly into the nearest cloud region. SDN services, meanwhile, can help companies facilitate a multi-cloud strategy by offering private connectivity to a variety of hyperscale cloud providers, all through the same local connection to the SDN fabric. Like native links, the virtual connections deliver reduced latency and greater security.

Taking this concept further, business customers can now use local CSP partners to make the private connections to the cloud. For example, local cable operator sales teams can use this approach to sell a metro Ethernet line that connects to an Edge Data Center. Depending on the bandwidth and the cloud providers required, that business could either use a native or virtual on-ramp. Either way, it will have an efficient, secure, private cloud connection.

Such an approach promises to deliver multiple benefits for both service providers and their commercial customers. To sum them up, these expected benefits include:

- Significantly lower and more consistent latency levels
- Greater security for the information transmitted
- Lower overall costs and flexibility
    - Lower costs: lower data transfer fees without long-haul transport costs
    - Ability to scale bandwidth
    - Multi-cloud access (via SDNs)

THE NEED FOR LOW-LATENCY CLOUD CONNECTIONS

Latency is a growing performance issue for service providers and their commercial customers. It has become a larger concern as the data and applications have become more interactive, more video-intensive and far more complex in nature.

For corporate end users accessing applications entirely in the cloud, the conventional wisdom had been that latency levels just needed to be below 100 ms (eg, traditional Internet in the same region) to be acceptable. This was true for file transfers, and the occasional, non-critical use of cloud-based email or office applications. Meanwhile, ultra-low latency demands (<10 ms) are largely driven by hybrid cloud providers integrating with public cloud.

But that no longer applies, as more cloud workloads are driving tighter latency requirements. This is true for any interactive or data-intensive application, but is most notable in the SaaS segment, which includes such key enterprise-oriented tools as office productivity/ collaboration tools and hosted desktops. For example, business intelligence tools such as Power BI and Tableau Online have different performance requirements than file transfers.

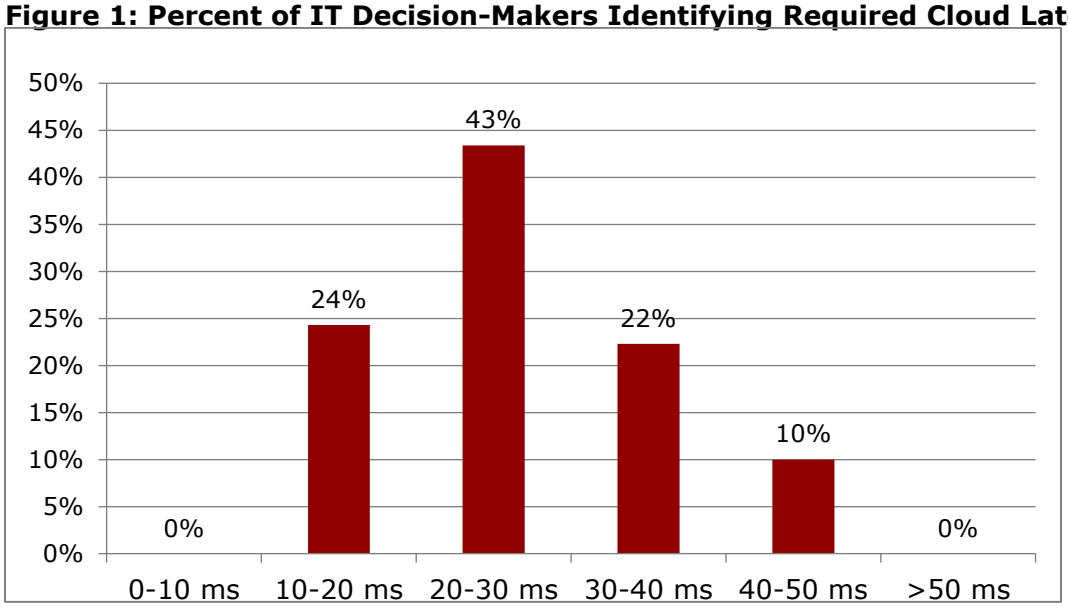
Here are examples of how sensitive applications are to latency, even when used in fixed laboratory environments that are not subject to the traffic spikes of the public Internet:

- **Online Collaboration Tools:** In testing conducted by Gartner, users reported "noticeable performance degradation" for the application once round-trip latencies exceeded 25 ms. And when latency levels rose significantly higher than 30 ms, the application became "virtually unusable." Even the 25 ms latency level saw relatively slow page loads of about 8.5 seconds (Gartner Research, Kyle Davis).
- **Productivity Suite:** Nokia Bell Labs has conducted experiments on the Mean Opinion Score (MOS) of office applications. This works similar to MOS measurements in voice, whereby users rank their experience on a 1 to 5 scale. Nokia Bell Labs found that the applications work well at 20 ms, but by the time latency reaches 40 ms, the experience has "degraded noticeably" ("Tips for Enhancing Virtual Desktop Quality of Experience," RJ Vale, Nokia Bell Labs, and "Virtual Desktop Performance and Quality of Experience," Alcatel-Lucent/Nokia Bell Labs).
- **Virtual Desktop:** Similar to Nokia Bell Labs, the Telecom Research Center in Vienna conducted experiments with virtual desktop applications. Researchers found that the MOS and "ease of use" began to drop at latencies above 20 ms. At 25 ms, the time it took to conduct drag-and-drop and menu exercises increased by 10 percent and 20 percent, respectively ("Quality of Experience in Remote Virtual Desktop Services," Pedro Casas, et al., Telecommunications Research Center, Vienna).

What does all this mean for IT decision-makers? Due to such rapidly developing trends, commercial customers are increasingly seeking *much lower and more consistent* latency levels than ever before from their trusted service providers. In a recent survey, for instance, the equity research firm Cowen & Company found that two thirds (67 percent) of organizations want latency levels of less than 30 ms for their various applications. Furthermore, the study found that:

- 67 percent of organizations say that the quality of traditional connectivity is "very important" in the colocation buying decision
- 41 percent of organizations need help in "optimizing performance" of public cloud deployments

Figure 1: Percent of IT Decision-Makers Identifying Required Cloud Latency (in ms)



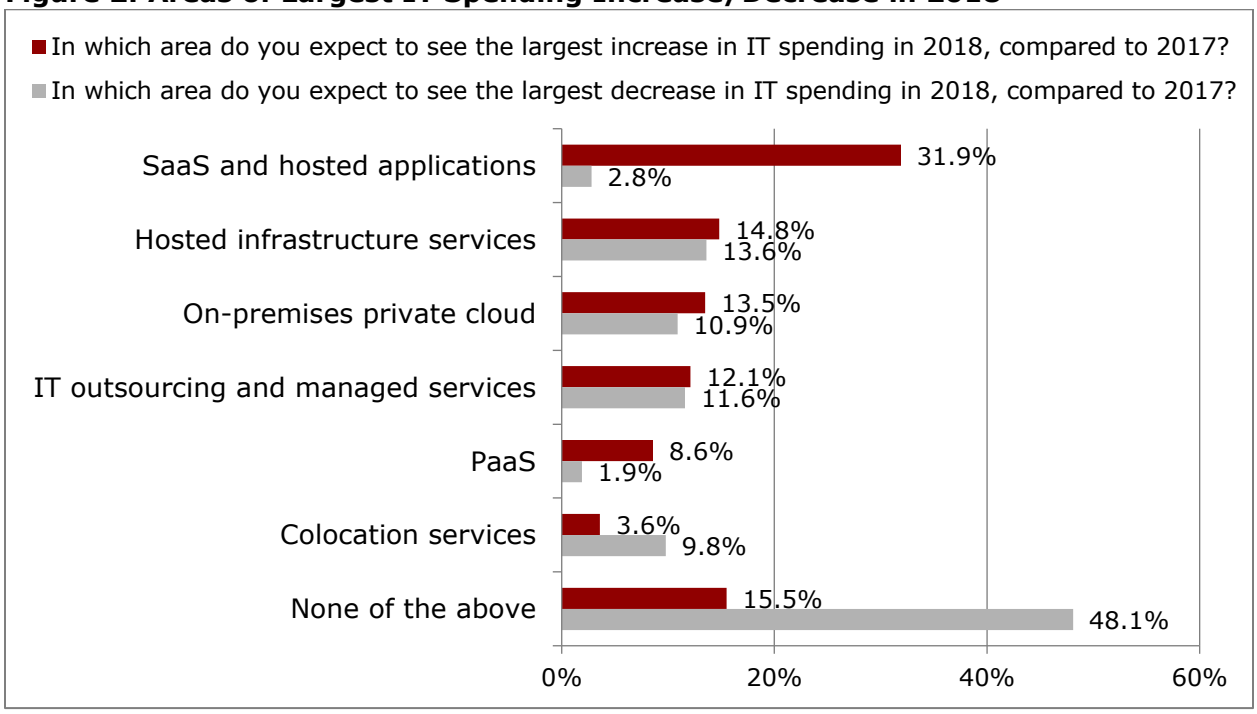
Source: Cowen & Company Equity Research/Altman Vilandrie & Co. Cloud Survey, May 2017

This impacts public cloud adoption. 451's 2018 Voice of the Enterprise survey found that:

- "Performance (latency & availability)" ranks highest as the driver for enterprises moving workloads off the public cloud, with 26 percent citing it is the primary reason.
- 23 percent of enterprises also said it was the problem that most contributed to their view that certain workloads are unsuitable for the public cloud.

451 has also found that SaaS applications -- likely more latency-intensive than traditional services -- will generate the largest spending increases for the cloud. IT decision-makers are seeing the challenge of latency for these increasingly popular services.

Figure 2: Areas of Largest IT Spending Increase/Decrease in 2018



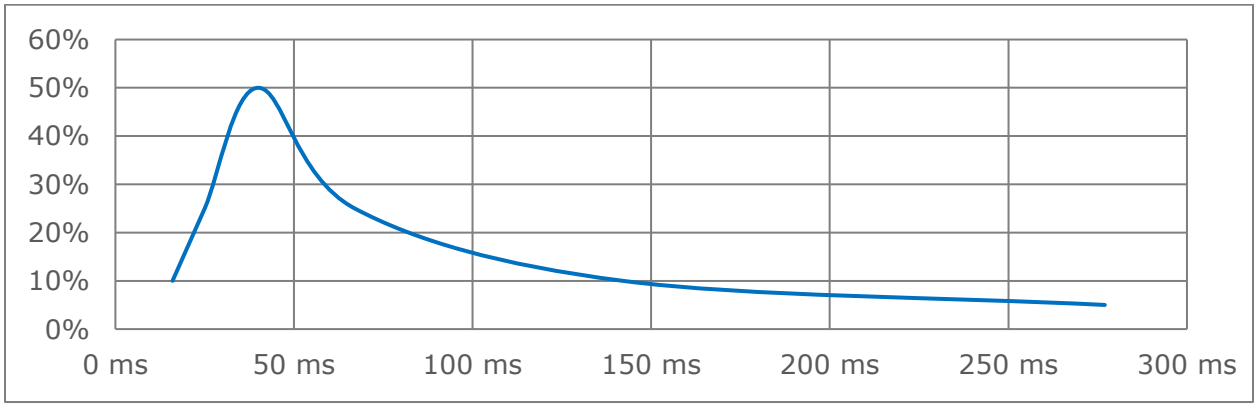
Source: 451 Research; n=561

In addition to the hybrid cloud configurations and SaaS applications already mentioned, 451 has identified that ultra-low latency, 1 ms to 10 ms, will be critical for upcoming "Industry 4.0" applications, such as manufacturing equipment that makes split-second decisions based on cameras, sensor readings, machine learning and machine vision. Researchers also identify IoT and Engineering Augmented Reality (e.g., used to support product R&D) as requiring such sub-10 ms latencies ("Which Workloads are Attracted to the Edge?," Daniel Bizo, 451 Research).

PERFORMANCE RESULTS: LOCAL CLOUD ON-RAMP SOLUTIONS VS. THE PUBLIC INTERNET

Can the public cloud deliver such desired low latency levels? Not for most Internet users right now, it can't. For example, computer scientists at the University of Waterloo, Ontario, tested multiple locations throughout North America. What they found is that only one fifth (20 percent) of the U.S. population can regularly experience cloud latency levels of 30 ms or lower today ("The Brewing Storm in Cloud Gaming: A Measurement Study on Cloud to End-User Latency," Choy, et al.).

Recent data from Cedexis further demonstrates the magnitude of the problem. **Figure 3** shows the "long tail" of the response times to an Ashburn-based cloud provider from East Coast locations. While the median is 42 ms, 10 percent of latencies observed are above 142 ms, and 5 percent of latencies are above 240 ms.



Source: Cedexis Radar

These high measurements may sound like an infrequent "hiccup," but they have significant ramifications. The packets that are caught in congestion at the tail are usually holding up the entire exercise – a page load, a click on an office tool, a menu in a SaaS application. That is why the "tail latency" is the focus of hyperscalers, who are working to reduce it within their data centers. For example, Kathryn McKinley, principal researcher at Microsoft, has written that "Data centers that service interactive user requests must be carefully engineered to opti-mize the tail response times or *users abandon the service*" (emphasis added).

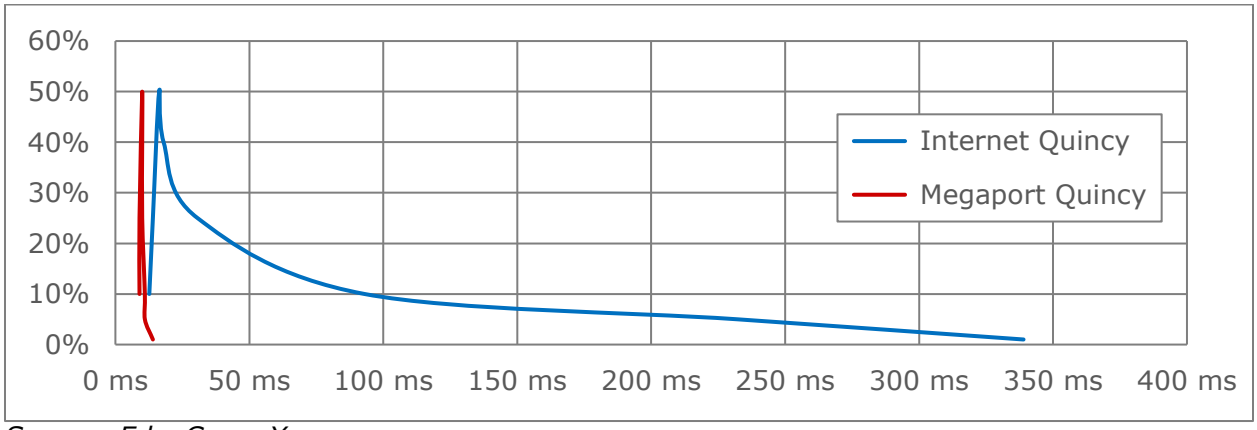
So how do local cloud on-ramp solutions compare with the public Internet in terms of both overall latency levels and tail latencies? EdgeConneX, along with key experts in the Internet measurement and information technology space, have examined this question and found significant improvements by using direct cloud connectivity solutions.

EdgeConneX and Cedexis published results of a native cloud on-ramp latency in 2017 ("Cloud, Content, Connectivity and the Evolving Internet Edge"), with significant findings:

- The native on-ramp to AWS US West-02 (200 miles away) had latency of 4 ms -- a dramatically fast response, superior to an Internet connection to a local data center.
- Testing from the same location showed 24 ms over the public Internet, consistent with Cedexis' real user measurements, taken throughout Portland, of 24-31 ms, looking at the best case scenario (10th-50th percentile). However, latencies routinely reach 45-60 ms in the 90th percentile.
- Therefore, the native on-ramp showed latency improvement of **85% or higher.**

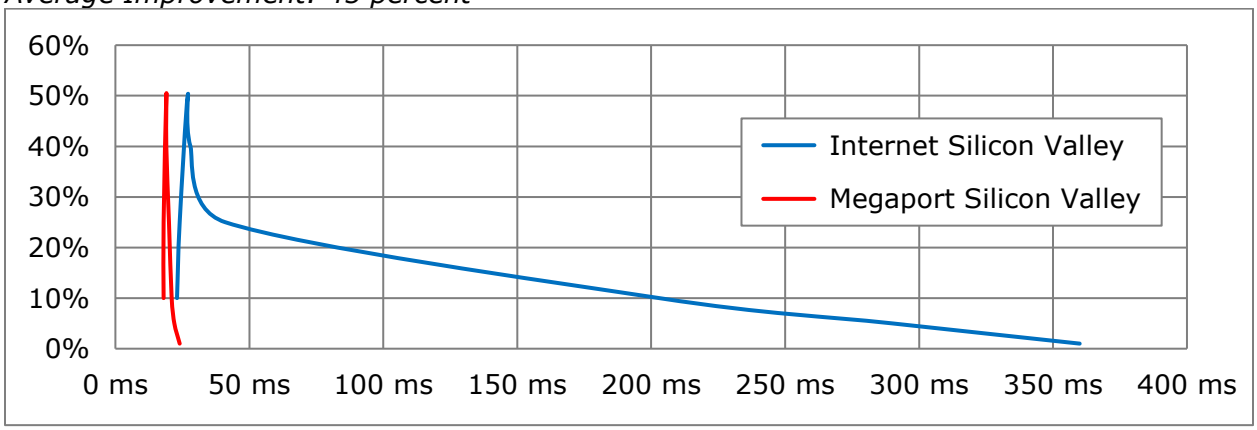
Now EdgeConneX has conducted new measurements. Along with its infrastructure provider Curvature LLC, the firm analyzed 4 days of latency data from its Edge Data Center in the Portland, Oregon, area to Microsoft Azure locations in Quincy, Wash., and Silicon Valley.

Figure 4: Latencies From Hillsboro, Ore., to Azure Cloud (Quincy, Wash.)
Statistical distribution of 4-day period/12 hours per day
SDN (Megaport) Average: 9 ms; Internet Average: 28 ms
Average Improvement: 67 percent



Source: EdgeConneX

Figure 5: Latencies From Hillsboro, Ore., to Azure Cloud (Silicon Valley)
Statistical distribution of 4-day period/12 hours per day
SDN (Megaport) Average: 18 ms; Internet Average: 33 ms
Average Improvement: 45 percent



Source: EdgeConneX

Figures 4 and **5** show the distribution of latency levels for data traffic flowing from Portland, Ore., to Quincy, Wash., and Silicon Valley during the busiest 12 hours of each day over the four-day period. The blue lines denote the latency levels for Internet traffic; the red lines denote the latency levels for Megaport traffic. They show the consistency of the cloud on-ramp, specifically using the software-defined fabric (in this case, from Megaport).

For example, the SDN on-ramp has a lower average latency:

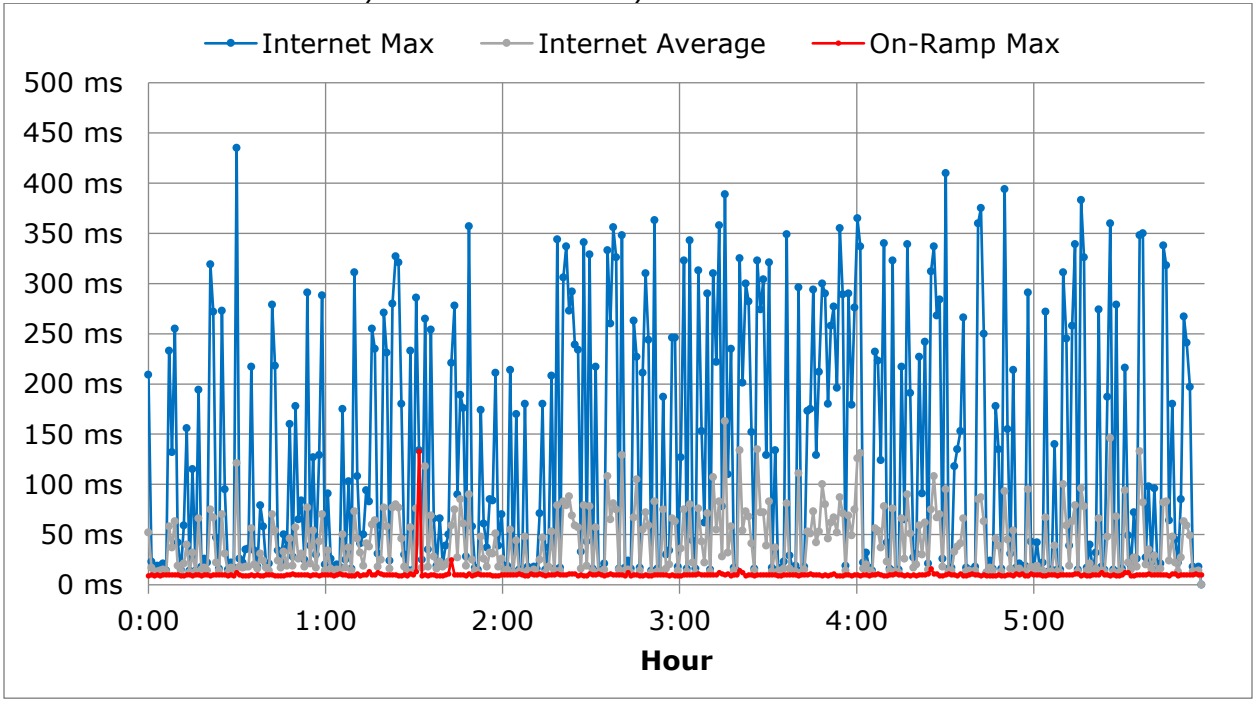
- 9 ms vs. 28 ms for Quincy (67 percent improvement for SDN Cloud On-Ramp)
- 18 ms vs. 33 ms for Silicon Valley (45 percent improvement for SDN Cloud On-Ramp)

The SDN on-ramp also shows remarkable consistency:

- A mere .005 percent of observed latencies to Quincy were above 12 ms. For the public Internet, 10 percent were measured at or above 93 ms and 5 percent above 232 ms.
- Less than 5 percent of observed latencies to Silicon Valley were above 22 ms. In contrast, 25 percent of latencies for the public Internet were above 41 ms, and 10 percent were above 200 ms.

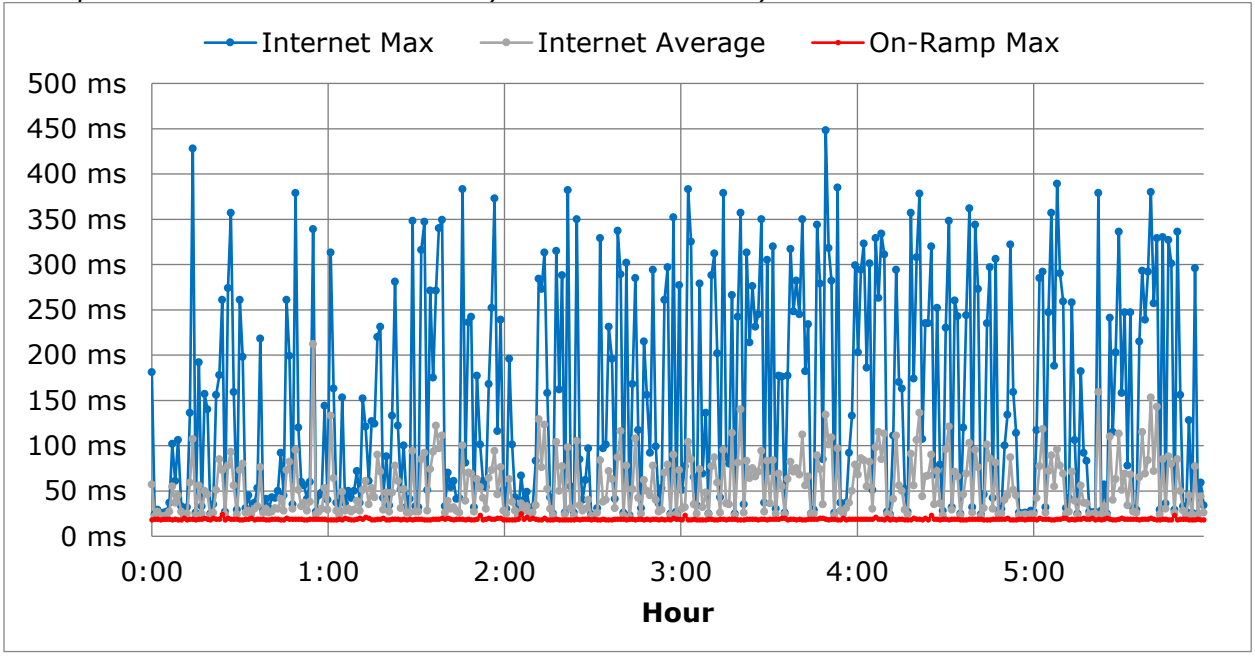
This is also visible when looking at the individual measurements, as shown below in a sample of one-minute measurements taken over a six-hour period. The "max latency" is measured by taking several latency measurements every minute, and looking at the maximum value. As discussed above, this is a key metric in determining application performance, due to the "long tail" impact important to hyperscalers. For the cloud on-ramp, this remains at a steady red line throughout the entire busy-hour period.

Figure 6: Latencies From Hillsboro, Ore., to Azure Cloud (Quincy, Wash.) *Example of latencies measured every minute over 6 busy hours*



Source: EdgeConneX

Figure 7: Latencies (in ms) From Hillsboro, Ore., to Azure Cloud (Silicon Valley)
Example of latencies measured every minute over 6 busy hours



Source: EdgeConneX

COST SAVINGS OF LOCAL CLOUD ON-RAMPS

In addition to the performance benefits, the cost savings of using on-ramps, either native or virtual, to reach the cloud can be quite significant. This is because businesses can avoid the high outbound data transfer rates that the cloud hyperscalers charge when sending out data over the public Internet.

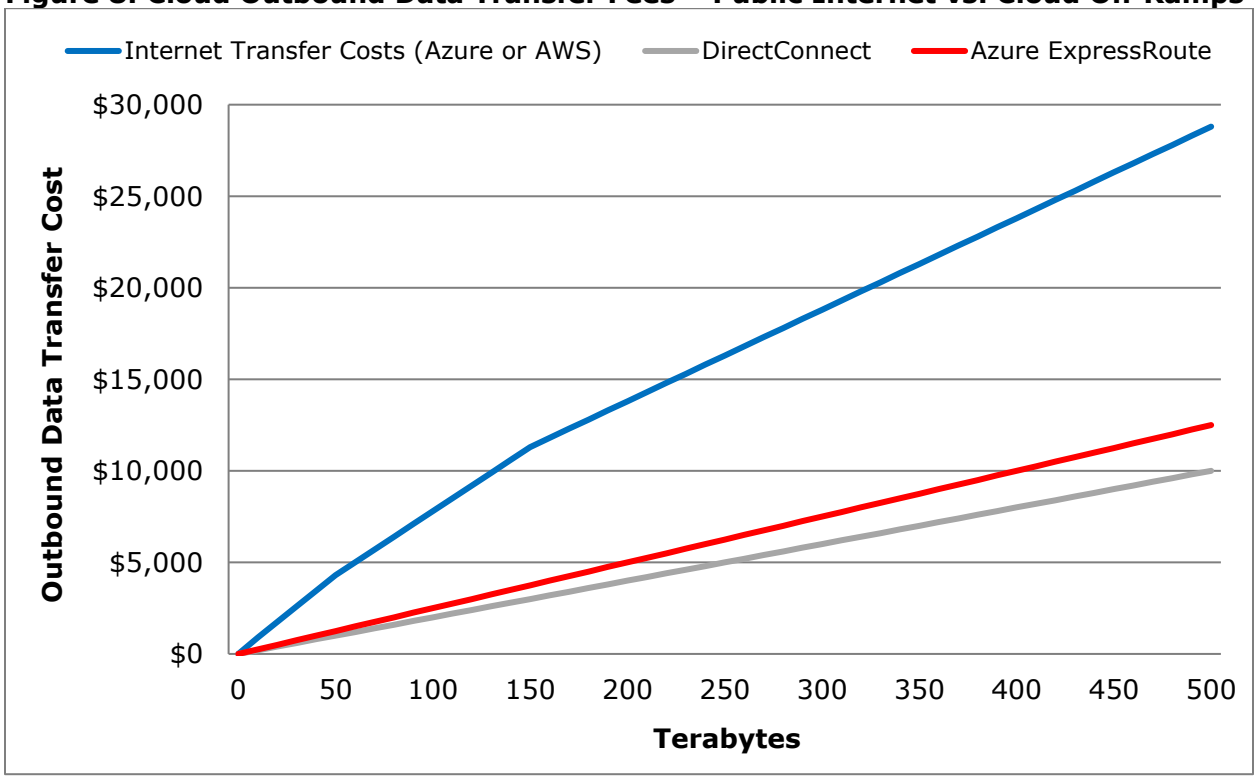
Figure 8 highlights the magnitude of the savings, using the pricing information available on AWS and Azure websites. Customers can save 56 percent to 77 percent, depending on how much data and which cloud provider they use. For example:

- A customer transmitting 50 TB of outbound traffic saves about $3,000, or 70 percent, on Azure, and $3,300, or 77 percent, on AWS vs. the public Internet.
- For larger usage of 200 TB, the customer saves $11,300 on AWS and $9,940 on Azure, for savings of 69 percent and 61 percent, respectively.

Edge on-ramps generate additional savings for businesses by avoiding the long-haul costs of connecting with a core on-ramp for the same type of connectivity. For example, in the traditional "core" model, businesses had a choice: they could either pay for a network connection over hundreds of miles or endure the higher costs and lower performance of cloud connectivity through the Internet. With edge on-ramps, businesses can avoid both these problems.

Added to this is the flexibility of the cloud connections. Not only can businesses choose to flexibly increase or decrease their bandwidth depending on demand, but in the case of SDN fabrics, they can access multiple public cloud providers through the Edge Data Center.

Figure 8: Cloud Outbound Data Transfer Fees – Public Internet vs. Cloud On-Ramps



Source: Created Based on Data Transfer Cost Published on AWS/Azure Web Sites

SECURITY BENEFITS OF LOCAL CLOUD ON-RAMPS

Besides the cost savings covered above, cloud on-ramps can deliver substantial security benefits as well. These benefits include:

- Leveraging a private connection, skipping the public Internet
- Using a dedicated Layer 2 connection
- Taking advantage of dedicated capacity

Not too surprisingly, "private" and "dedicated" are the key words here. With the security of the data they are transmitting clearly a strong and growing concern for service providers and their commercial customers alike in an era when DDoS attacks, network hacking and other security threats have become all too common, both parties want to make sure that the data is protected from prying eyes as much as possible. The best way to do this is to isolate the data from the highly vulnerable public Internet by relying on private, dedicated connections and capacity to reach the cloud untouched and unscathed.

Cloud on-ramps offer such security capabilities because their links to the cloud are always private and dedicated, enabling service providers and businesses to safeguard critical information from wrongdoers. In addition, these improved capabilities can help companies to comply better and more easily with the increasingly stringent regulatory requirements of government agencies, which are also greatly concerned about data and network security.

KEY TAKEAWAYS/CONCLUSION

Enterprises are increasingly finding the need for low, fixed latency for the next generation of public cloud applications. While latency in the cloud began as hybrid cloud providers needed fast connections between their computing resources and the public cloud, the need has continued with the large-scale adoption of hosted office applications, hosted desktop and SaaS. IT decision makers are justifiably hesitant to deploy new applications on the public cloud if latency is a concern.

Ideally, latencies supporting these applications should be as low as possible because performance levels typically degrade in the congested, "best efforts," environment of the public Internet. Even in a fixed latency environment, performance begins to degrade between 20 and 30 milliseconds. These requirements will become even more stringent for the next generation of cloud applications, such as data-intensive AR, VR and machine-vision applications requiring as low as 1 ms to 10 ms.

Can the cloud support such applications? With edge cloud on-ramps, enterprises can experience the cloud as if it were local, with latencies 45 percent to 85 percent lower than the public Internet. Moreover, enterprises will enjoy greater overall consistency, ensuring that applications don't stall during times of high usage.

Meanwhile, having more applications in the cloud mean more services that must be secured. This includes the network transport that is transporting data to the public cloud. Companies will increasingly look for solutions that meet stringent security requirements. Again, edge on-ramps can provide this security using private, dedicated connectivity.

Finally, while the cost and flexibility of cloud applications are advantageous for enterprises, they will also want to avoid the high cost of either Internet transfer or long-haul private cloud connectivity whenever possible.

While edge cloud on-ramps offer a solution that meet performance, security and cost criteria, these solutions can be even more successful with the promotion and partnership of CSPs. By becoming part of the connectivity, infrastructure and sales ecosystem, CSPs can help businesses reach their full potential in the cloud.