Cloud, Content, Connectivity and the Evolving Internet Edge

A Study from Cedexis and EdgeConneX

Summer, 2017

cedexis.com edgeconnex.com

CLOUD, CONTENT, CONNECTIVITY AND THE EVOLVING INTERNET EDGE

The Internet is always evolving to address changing applications, user habits, and, of course, the insatiable desire for more and more data.

EXECUTIVE SUMMARY

The Internet is always evolving to address changing applications, user habits, and, of course, the insatiable desire for more and more data. Unlike past evolutions of Internet architecture, however, the most recent trends will change not only what is delivered, but from where.

Cloud providers, in particular, find themselves at an evolutionary point and have only now started to realize the need for a broader footprint. Cloud data centers were originally built to serve large regions (like the entire east coast). This means quality of service in any metro area can vary wildly in quality based on latency.

The next step for cloud lies in two key architecture changes:

- First, hyperscale cloud providers and Software Defined Networking (SDN) fabrics supporting on-demand access to the cloud are enabling low-latency direct connectivity from underserved locations. This has taken the form of edge network deployments that send traffic through local, direct, private connections – skipping the public Internet entirely.
  - ***Customers in those underserved markets can experience a 50 to 85 percent reduction in response time by accessing these edge nodes.***
- Next, as demand for localized cloud on-ramp services grow, they will take steps towards deploying incremental cloud infrastructure at "the edge" of the network, closer to the ultimate end user. This will follow the recent deployments of Content providers who have recently pushed their caches further into these underserved locations.
  - ***As a result, CDNs have realized response time improvements ranging from 15 to 47 percent.***

What does this mean for managers of application delivery? CIOs and network architects will be managing static & dynamic content, low-latency applications, and data storage in a rapid deployment environment. The networks handling this will combine centralized & edge facilities, hybrid cloud, DIY caches and 3rd party CDN as critical components.

Operations professionals will need to consider the implications of proximity in planning their network topologies and the potential use of regional deployments and on ramps to guarantee their users' secure access and quality of experience.

Sophisticated global server load balancers (GSLBs) are being deployed to ensure publishers deliver the right quality, at the right price, to their audiences. While traditional load balancing has used round robin, or geography-specific, rules, modern GSLBs base decisions on the actual experience being delivered to the end user. This creates an opportunity and a threat: opportunity to capture business by delivering universally great service, at the threat of losing that business to those who do the job better.

Third parties such as EdgeConneX provide colocation services that bring both content and cloud connectivity closer to the end user. The analysis contained in this whitepaper is based on EdgeConneX deployments, combined with Cedexis Response Time information, which demonstrate the superior performance of both caching and connectivity within the same metro area as households and enterprises.

THE PERFORMANCE RISKS OF CURRENT CLOUD ARCHITECTURE

One key culprit in the challenge of data delivery is, believe it or not, the speed of light – 186 miles per millisecond – so fast that it is essentially the one physical property that defines time.

Yet, the inability of data to travel faster means that there is always a delay between the request for information and its delivery through a computer network. If there is one thing that can be said to consistently impact the time it takes for content to get from one location to another, then, it is proximity. The greater the distance between the sender and receiver, the longer the journey will take.

Not only does light travel slower over fiber (about 2/3rds the speed of light), but fiber is never laid over a direct route. Then, there are the multiple "hops" that are inherent in the Internet: the network is designed to break up our data, send it through multiple routers, and even multiple networks, on the way to its destination.

However, the speed of light is not the whole story of how data is delayed. Not only does light travel slower over fiber (about 2/3rd the speed of light), but fiber is never laid over a direct route. Then, there are the multiple "hops" that are inherent in the Internet: the network is designed to break up our data, send it through multiple routers, and even multiple networks, on the way to its destination. The end result is that latency in ordinary public "ping" tests can be 15 times higher than the speed of light, and the latency of fetching a web site 35 times higher.

So, after proximity, the network must be designed to minimize the hops, traffic jams, and indirect routes. The further the data must travel, and the more time it spends in routers and servers along the way, the longer it must take.

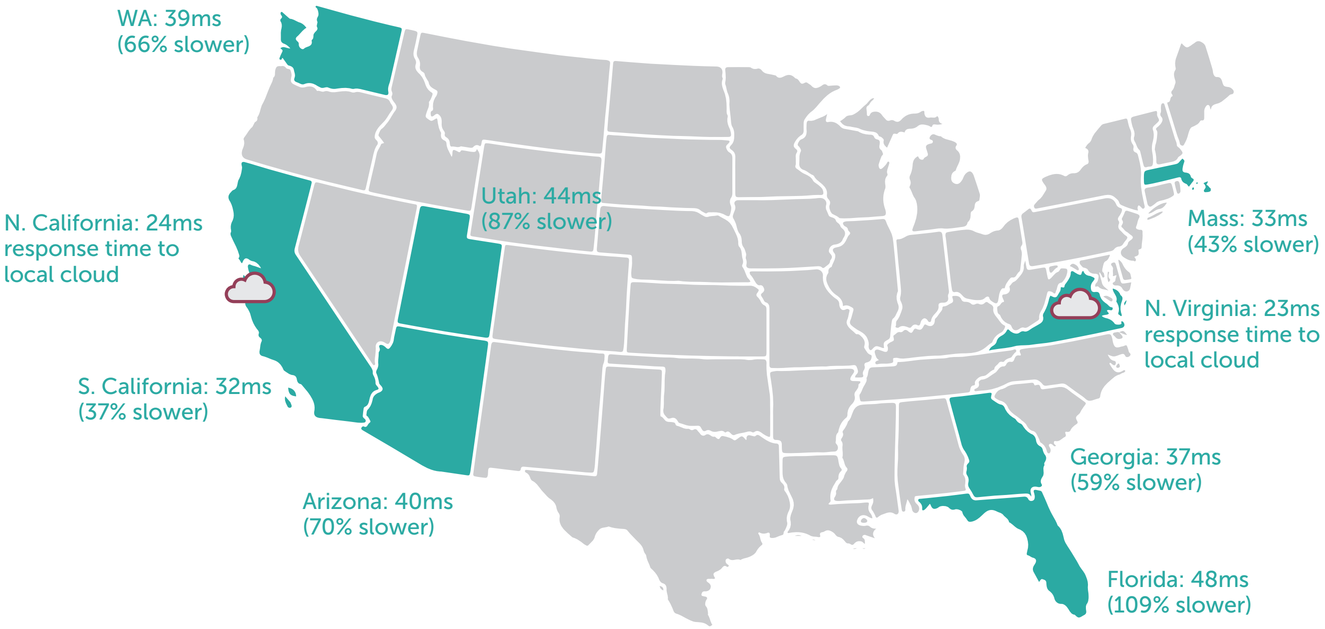
THE RISK WITH CLOUD DISTANCE

As cloud use continues to increase, we will see that improving performance will soon become a matter of need for each of the cloud providers. For instance, the following illustration shows the response time for several major cloud providers' Virginia and Silicon Valley infrastructure to points further in the eastern and western US.

Cloud providers' regional architecture means slower performance away from local nodes.

Cedexis Radar Data of Real User Measurements: How does cloud performance change with distance?

Cloud providers' regional architecture means slower performance away from local nodes. Response time to two major hyperscale cloud providers in Ashburn, VA and Silicon Valley grows with distance, causing performance degradation.



Compared performance of two hyperscale cloud providers in various regions. East coast measurements compared response time to Ashburn, VA location; west coast measurements to Silicon Valley.

As we might have expected, users nearby are receiving a much better experience than those who are several hundred miles away.

This does not yet mean that users in those further-off locations are necessarily getting a bad experience – but it does mean that they are at greater risk over time as Internet traffic grows. The additional milliseconds of delay, combined with the growth in traffic volume, mean it is highly likely that they will see more and more performance issues.

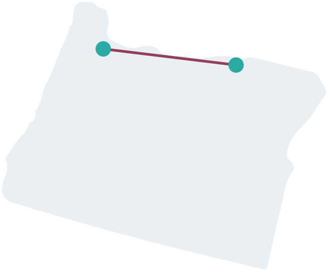
THE FIRST EVOLUTION: CLOUD CONNECTIVITY AT THE NETWORK EDGE

Until they determine how to distribute information effectively between the core and the edge, the constraint of investing in physical presence everywhere would be a challenge for cloud providers. Instead, businesses at the network edge can bridge the performance gap by using direct connections to the cloud (and increase in security by bypassing the public Internet).

Cloud on-ramps have traditionally been offered in the core Internet markets, and are now being expanded to edge markets. Customers can establish connectivity using a dedicated metro-area fiber or carrier transport connection (e.g., a metro-e connection directly from the enterprise office to a data center with a cloud connection capability).

This provides not just a secure connection, but a low-latency performance improvement that occurs when skipping the public Internet. Even over a distance of 200 miles, this can make a significant difference.

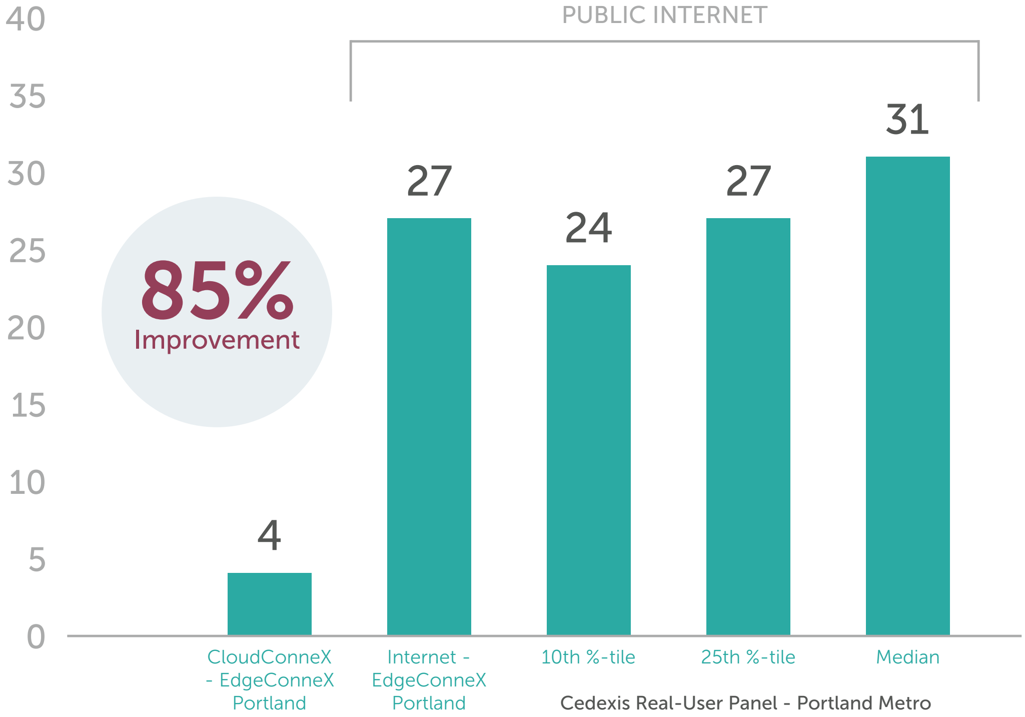
For example, analysis of EdgeConneX AWS Direct Connect service out of their Portland area facility shows the following vs. the public Internet to AWS US West locations.



approximately
~200 Miles

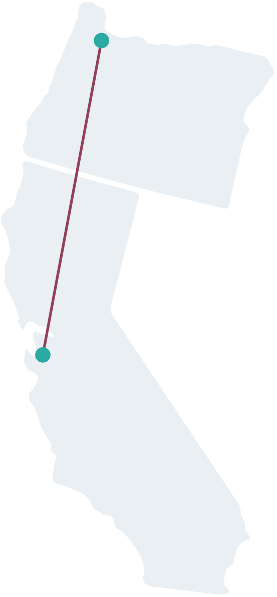
This means that a business using direct connectivity to connect to the cloud 200 miles away can cut the response times by about 85 percent.

Response Time from Hillsboro, OR to AWS US-West-2 (Central Oregon, aprox. ~200 Miles)



Not all businesses have a native hyperscale cloud connection, like AWS Direct Connect, in their area. Fortunately, SDN solutions offering virtual cloud connections, such as Megaport, are helping to close the gap with virtual connections at the network edge. Moreover, they can help facilitate a multi-cloud strategy by offering private connectivity to a variety of hyperscale cloud providers, all through the same local connection into the SDN fabric.

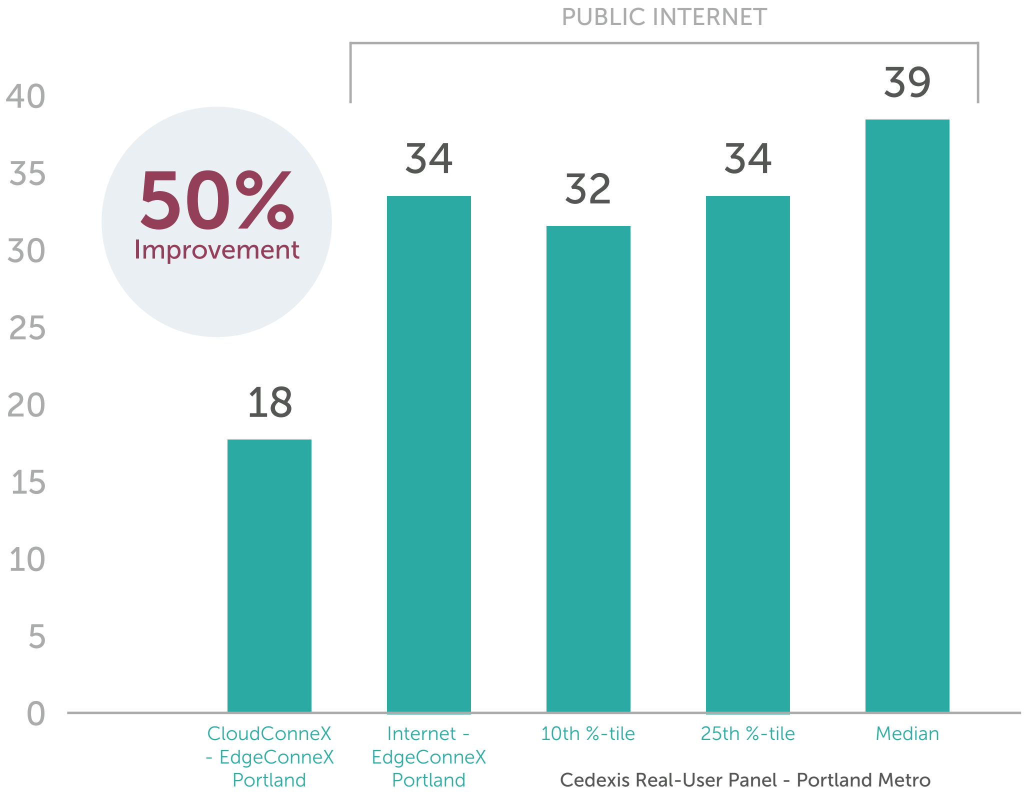
Like a physical link, these virtual links also provide response time benefits.



approximately
~600 Miles

In this case, using a virtual connection to transmit data to the cloud location 600 miles away reduces the response times by ~50% vs. the public Internet.

Response Time from Hillsboro, Oregon to AWS US-West-1 (Silicon Valley, aprox. ~600 Miles)



These results indicate that an edge connection to the cloud can provide local performance from a regional data center.

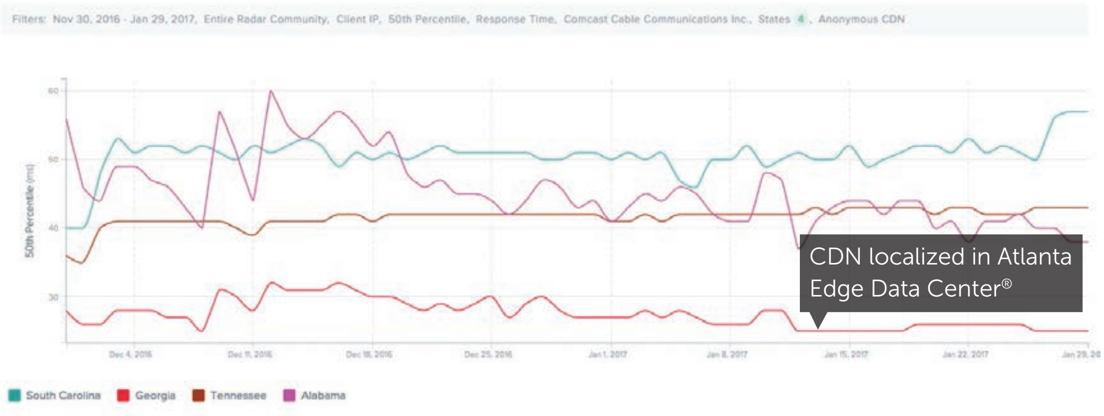
ENABLING PROXIMITY: CLOUDS FOLLOWING CDNS

As cloud hyperscalers grow, they will imitate the model of major Content Delivery Networks. That is, build out a growing number of nodes to cache more frequently accessed information close to the end user.

While geographic distribution of networks nodes has been part of their business model since their inception, CDNs and content providers are driving the new evolution of the Internet to the edge.

Cedexis Radar real-user measurements provide data on EdgeConneX-housed content peering. These consist of caches serving a metro area, directly connected into the local ISP networks ("one hop" from the customer).

The example below shows how this works with the same CDN on the same ISP over three different states. The CDN is cached locally in an EdgeConneX facility, directly connected into the Atlanta-area ISP. This is where the response time is fastest. In the neighboring states, where there is no local cache, the response time gets worse.



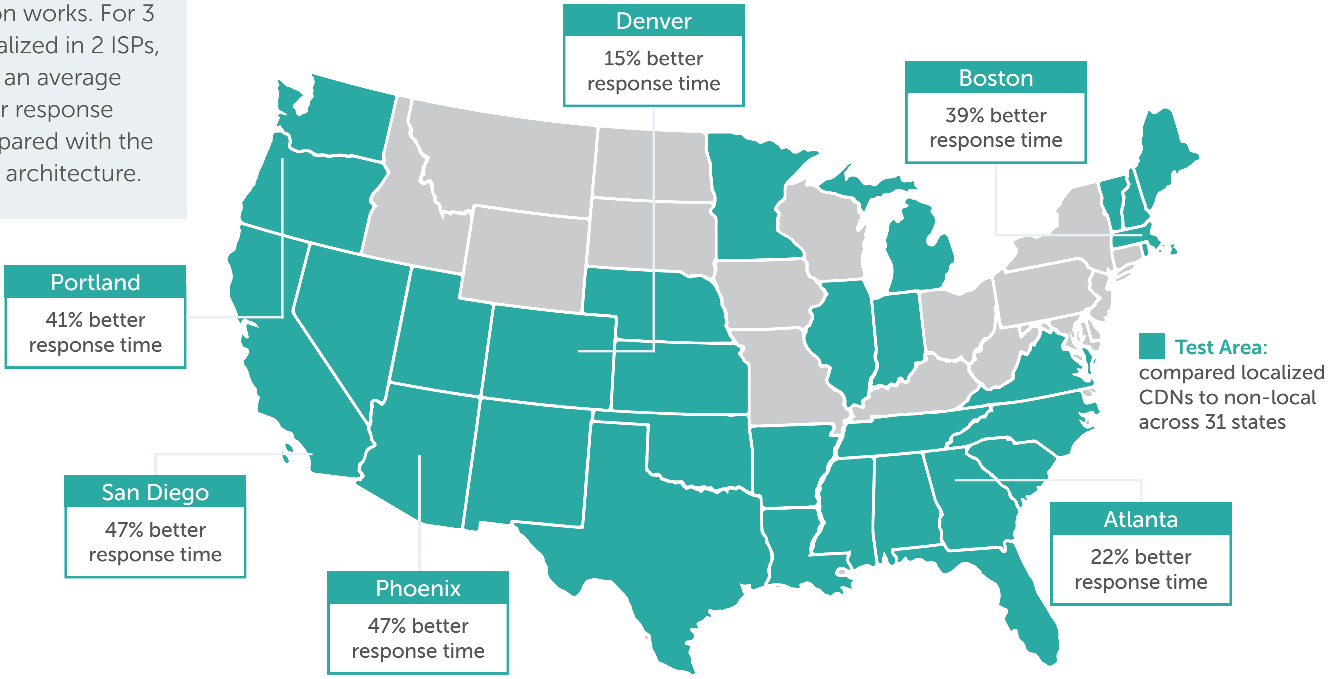
Filters: Nov 30, 2016 - Jan 29, 2017, Entire Radar Community, Client IP, 50th Percentile, Response Time, Comcast Cable Communications Inc., States 4, Anonymous CDN

CDN localized in Atlanta Edge Data Center®

This pattern is repeated at the national level. The response time of the CDN is 15-47% better vs. the same CDN/ISP combination in neighboring states. In other words, as users get closer to the cache, their response time improves dramatically.

Content Delivery Networks have found localization works. For 3 CDNs localized in 2 ISPs, there was an average 39% better response time compared with the non-local architecture.

Cedexis Radar Data of Real User Measurements: How much does localization improve response time vs a non-local architecture?



Denver
15% better response time

Boston
39% better response time

Portland
41% better response time

San Diego
47% better response time

Phoenix
47% better response time

Atlanta
22% better response time

🟩 Test Area: compared localized CDNs to non-local across 31 states

Compared performance of CDN in EdgeConneX territory (with a localized cache) to the same CDN on the same ISP in territories without a local cache. Used Cedexis Radar tool median response time measurement.

This is having implications across the industry. First, hitting both the CDNs and "do it yourself" content distributors. Next the cloud providers will want to find a way to move a portion of their data –mostly likely the most highly accessed application functions and databases – to a more localized architecture.

MANAGING RESOURCES: MULTIPLE PROVIDERS, MULTIPLE LOCATIONS

Both of these localization techniques will improve delivery, while at the same time adding to the complexity of managing public, private, and hybrid cloud solutions, cloud connections, and CDN services for application management and delivery.

As the edge evolves to include more cloud and content deployments, the complexity will increase. **Companies must learn to manage:**

- **Multiple CDNs**
- **Private CDN**
- **Private Cloud**
- **Public Cloud over Internet**
- **Public Clouds over direct and virtual connections**
- **Hybrid Cloud**

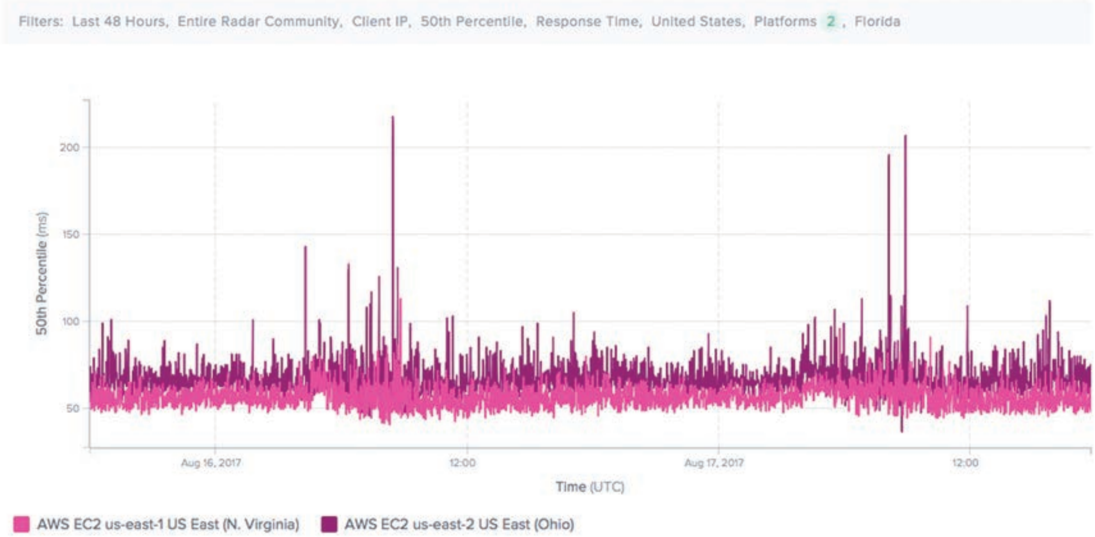
Localizing a network to bring information closer to end users means ensuring information comes from the right source. This source selection can be based on proximity, but also must factor in network hops, network congestion, server congestion, bandwidth and similar performance factors of all available resources.

Source selection can be based on proximity, but also must factor in network hops, network congestion, server congestion, bandwidth and similar performance factors of all available resources.

For example, a cluster of server instances in the nearest POP may be overloaded; peering between the network and an individual POP may be congested; or there may be simply degraded QoE measurements detecting a problem with one of the resources.

The chart on the next page demonstrates how this works. You can see AWS traffic to Florida from Virginia and Ohio. Most of the time, Virginia has the lower (better) response rate. But you can also see instances where the median response time is better from Ohio – 15 percent of observations, in fact. In order to maintain consistent service, it will from time to time be necessary to switch between the two locations.

This plays out across multiple geographies, CDNs, Clouds, ISPs, and every other element in the Internet value chain. Proximity creates the opportunity to deliver an excellent experience at the lowest cost as volumes increase. Detecting congestion and re-directing traffic to avoid it is the balancing activity that delivers on that promise.


Filters: Last 48 Hours, Entire Radar Community, Client IP, 50th Percentile, Response Time, United States, Platforms 2 , Florida

An adequate GSLB, then, will:

- Monitor, track, analyze, and make actionable real user measurements, which ensure that the outcome (high QoE) is the focus.
- Ingest, integrate, and use data from synthetic monitoring, as well as other 3rd party sources (APM, contractual agreements, etc,), which ensure that the global network is aware of activity at every node.
- Automatically route traffic to the optimal system node, based on easily-adjustable algorithms that guarantee QoE at the point of consumption that meets or exceeds user expectations.

CONCLUSION: A NEW DIRECTION FOR INTERNET ARCHITECTURE

As Cedexis Real User measurements show, geography and connectivity are key factors in user performance and will continue to drive the future architecture of the Internet. Some key trends include:

- The evolution of “fog computing” as cloud providers determine and how to distribute data between edge, region, and core.
- New applications – such as IoT, smart cities or intelligent vehicles – will drive further need for low-latency network service.
- Data will move from customer-premise equipment as hybrid cloud and content continues to take off.

As the exponential growth in traffic congests peering exchanges and long-haul networks, look for organizations to become more sophisticated in how they manage data across this new environment.