The New "EDGE": Part 2 of our Coverage on EdgeConneX and the Internet

For the last 20 years, the Internet's edge has been located in eight Tier 1 markets. These markets are defined as Tier 1 to the Internet because they facilitate the exchange of data globally. As you may know, the Internet is a collection of networks – not a single network, so the facilitation of interconnecting networks is key to the Internet's ability to provide users with access to content and information. Originally, these networks met to exchange voice traffic, not data. Facilitating the exchange of voice traffic is very different from the exchange of data and different again from the exchange of data which is primarily video.

In the early 90's, the Internet began to take advantage of these pre-existing locations for convenience and network density. With a premium placed on network location, eyeballs (consumers of content) came to the networks over long distances to get to content. Today's exchange points need to be closer to population density, or where the eyeballs are located. This implies that content comes to the eyeballs at a local exchange. But why? Understanding the

“why” allows us to understand the move to what is being labeled the “Edge”. The “why” has to do with content and applications – past, present and future.

In the past, content was primarily zeroes and ones or character-based, and applications were focused on three upper-layer protocols (Telnet, File Transfer Protocol [FTP] and Remote Job Entry [RJE]) serving specific functions and exchanging small amounts of data. In that regard, neither content nor applications were sensitive to latency or bandwidth requirements. You could serve data from just about anywhere; therefore, you served it from where the networks met in a pre-existing architecture for pre-existing reasons. Latency did not matter because content, at the time, was not sensitive to delay.

Character-based content could be disassembled and reassembled at a gingerly pace and still arrive in a usable format, such as a Simple Mail Transfer Protocol (SMTP) email message. Bandwidth did not matter because transporting this data required very little capacity and was therefore affordable to serve at a distance. That solution worked very well until the web really began to take off and content began its dramatic shift to video and other bandwidth-intensive and latency-sensitive formats. In addition, as the web exploded, so did its use by corporate America for confidential information and mission-critical applications, creating the need for enhanced security and network resiliency.

Network architecture originally designed for voice data precedes World War II and is in desperate need of investment and redesign. The move away from character-based content and the overall adoption of the Internet and cloud by enterprise requires a solution for latency, network cost and complexity, resiliency and security. Lastly, it requires a fundamental change in approach to design (i.e., proximity to population or eyeballs). The resulting solution will give the Internet the foundation it needs to support current and future growth, and earn consumer and enterprise trust of online environments.

Video now makes up the vast majority of Internet traffic, and this trend is accelerating. With the growth and change of content as well as the increased need for greater security and reliability, we need to address how many Tier 1 Internet markets are enough. Since content needs to get to the consumer (eyeballs), a Tier 1 Internet market is better defined by density of population versus density of networks – hence the required investment and redesign. Now, you would consider the following metros Tier 1 or potential Tier 1 Internet markets that include Atlanta, Boston, Chicago, Dallas, Denver, Houston, Las Vegas, Los Angeles, Kansas City, Miami, New York City, Philadelphia, Phoenix, Pittsburgh, Portland, San Francisco, Seattle and Washington, D.C. However, only a handful of these markets have established peering and exchange fabrics, leaving many markets underserved by the Internet.

Content no longer functions reliably, securely or cost-effectively from a distance. Redesigning the Internet's Edge and establishing more Tier 1 Internet markets solves these problems.

Latency Matters

But there was a time when it did not. If you think about a blink of an eye taking 300 milliseconds, then who really needs to get content faster than that? Prior to now, the answer was no one. Character-based content and a reliable TCP/IP protocol ensured everything that was sent arrived as it was sent. Whether it was across town or across the globe, it just worked. When it did not work, we threw bandwidth at the problem to solve it. However, not so fast (no pun intended); first, let's discuss why bandwidth can't solve today's content dilemma; it involves physics.

Imagine a congested interstate from New York City to Miami and cargo that needs to be transported between these two markets. You can solve some of the issue of interstate congestion by adding two more lanes, albeit at significant costs. For the sake of this example, let's say we had an unlimited budget and added 20 lanes to eliminate congestion. Let's also assume that now all delivery trucks can safely travel at their max speed of 100MPH without stopping for the entire 1,200 miles.

What was once a 22-hour drive due to the congested infrastructure (or two lanes of bandwidth in a network) and limited velocity (60MPH speed) is now a 12-hour drive. Because of physics, if distance is absolute and velocity is at its maximum, than the equation for the fastest delivery time bottoms out at 12 hours. There is no way around the physics to improve this time. (See figure 2) How could you improve this then from what appears to be the absolute optimal equation? Change the only variable possible – distance.

By shortening the physical distance, we are able to change the time it takes to deliver the product to the consumer. By shipping the cargo once to Miami, storing it (in a digital world, think 'caching it'), and delivering it five miles across town at an average speed of 30MPH, we have completely changed the equation – 12 hours now becomes 10 minutes. For data and content, the only variable you can adjust is distance, as velocity is fixed at the speed of light for content. Therefore, adding bandwidth can only solve latency to a point. There comes a stage where increased bandwidth (more highway lanes) brings no incremental latency reduction. However, the reduction of distance never stops reducing latency.

Next, let us discuss applications and content. We'll use the same highway example and go back

to when the current infrastructure was designed. Back then, content and applications were analogous to that era's Model T. Due to the vehicle's relatively small size, we were able to fit many Model T's on the highway at the same time. Today, those Model T's have transformed into tractor-trailers because we have gone from small data to big data. Today, content and applications take up much more space on the highway and need to get to their destinations quickly or they don't work. So, not only is the cargo larger now, it is also perishable.

Today's content is highly sensitive to delays. The slightest network latency can have a profoundly negative impact on the performance of applications. It can cause abandonment and drops, stop voice recognition applications that require fast, bi-directional transport of packets like Siri and Cortana from working altogether, and can even cease the evolution of specific geographic markets due to an inability to serve High-Definition (HD) versus Standard Definition (SD). Proximity lowers latency each step of the way as distance is reduced. Bandwidth hits a limiting factor of helping latency and then raw physics take over.

Decreasing Network Cost

Revisiting once again our highway analogy, imagine the astronomical cost of adding those additional 20 traffic lanes to ease congestion from New York City to Miami. Now, imagine if you moved the goods only once and could then deliver them from Miami; you would not have to invest in more lanes and could use the current highway infrastructure for one-time traffic transfer. For example, if 1 million users in Miami want to view video content being stored in New York City, they would each have to go one-by-one to the city in order to retrieve it. However, if that video were moved once to Miami and then stored locally for all other users, the backbone needed to distribute the content amongst viewers would be reduced exponentially.

A key finding by ACG Research revealed that localizing traffic to serve a subscriber base of 1 million households produced backbone transport reductions of $110 million over five years. Applied nationwide to a subscriber base of 50 million households, this represents a $5.5 billion reduction in network and associated costs.

In addition, the possibility of an accident, road construction, weather, or some other disrupting factor on the highway between New York City and Miami can increase delivery time even further (you may know this as the spinning wheel of death on your screen). Therefore, by moving the content closer to the user, you eliminate all of the routers and networks between the cities as well as all of the potential points of failure. Between New York City and Miami, there is equipment at least every 60 miles and a minimum of three networks involved. This complexity creates a high likelihood that network failure is introduced to the equation. By moving data once, you eliminate all of the complexity and decrease the hardware, networks and hops involved to get your data where it needs to be.

Ensuring Security at the “EDGE”

There is a massive paradigm shift occurring in the banking world. In the past, you went to a branch office and dealt directly with a bank employee in a secure environment to deposit checks, withdraw money, and transfer funds. You also paid your bills using sealed envelopes and dropping them into a secure mailbox or directly into the hands of the postmaster at the post office. Over the last decade, this process has shifted entirely online for many consumers. Now more than ever, sensitive financial information is online and vulnerable to cyber attack, and banks have transformed to content companies with a financial focus.

One of the most secure ways to prevent such attacks and protect your information is via a

distributed architecture of data centers (i.e., the Edge). Traditional network architecture in eight peering points is no longer sufficient to support the security demands of the financial industry, which is one of the fastest growing adopters of online technology to service its customer base. The notion that every company will be a content company in the near future means that the Edge must grow significantly to offer a secure and trusted environment for this transformation.

The same is true for major retailers; for example, let's use Nike or Adidas. The traditional behavior of purchasing shoes at a store inside the mall is diminishing. Increasingly, consumers go online to buy their shoes, customize their look, and sync their key vitals from wearable fitness technology. Nike and Adidas are becoming content companies because the consumer interaction with them is moving to an online experience. This can only happen if we as the consumers trust that experience and can ensure that our data is safe.

Distributed Denial of Service (DDoS) attacks are often used by hackers to cripple sites on the Internet. By creating a distributed data center footprint, DDoS attacks are mitigated and sites are able to stay online and available to the consumer by shifting compute resources automatically to other sites where the attack is not happening.

Rethinking Network Design and Accessibility

Existing data centers and Network Access Points (NAPs) commissioned in the mid-90's were built in large Tier 1 markets far away from most of the people they serve and contain insufficient technology to handle today's massive traffic growth and low latency content delivery demands. Since distance to users was not a factor in their location, these data centers and NAPs focused instead on factors such as proximity to cheap power and land.

Today, the latency caused by moving packets across the country and back can significantly hinder the performance of real-time applications. For this reason, current data center site selection processes are hyper-focused on proximity to the user population and getting content closer to the highest concentrations of eyeballs. Also critical is close proximity to network provider aggregation points such as submarine cable head ends as well as rich fiber density.

The focus has changed from co-locating in NFL cities such as New York, Chicago and Los Angeles to targeting smaller Tier 1 and Tier 2 markets like Nashville, Orlando and Pittsburgh, which are now comparable to markets like Washington, D.C. in terms of Internet demand. But what these markets boast in growing Internet demand, they lack in public peering infrastructure. Today’s massive traffic growth requires the establishment of data centers and NAPs in markets that were not considered in the past in order to ensure the fastest and most reliable physical delivery of content to local-market consumers with a more secure design.

Defining the Real “EDGE”

The rapid growth of mission-critical or real-time traffic such as video, cloud-based applications and gaming has created the need to move the Internet's Edge closer to where the users are located in order to avoid significantly degraded performance and increased security risks. In response, many data center and colocation providers have commenced initiatives to expand their Edge footprint.

However, simply deploying a group of data centers in new Tier 1 and Tier 2 markets does not expand the Internet's Edge or make it more secure. Expanding a provider's Edge requires moving content closer to the end-user, a task not solely accomplished by merely placing a data center in an underserved market. So, what defines a new Edge market? Perhaps Zeus Kerravala of Network World said it best when he described infrastructure at the Edge and what it had to have in order to live up to the name:

• The services provided by the data center or exchange point must reach at least 50 percent of the broadband users, or eyeballs, located in its service area.

• At least 75 percent of the local Internet usage, including content, cloud services, or gaming sites, should be served by the provider.

• The provider must enable the shift of Internet peering traffic from the core to its new local peering point in the smaller Tier 1 or Tier 2 metro in order to bypass large markets and reduce latency.

• The provider should offer customers optimized performance at a lower cost.

• Because content has moved closer to the user, the provider should have the ability to deliver a richer, higher-quality media experience.

• Due to closer proximity to peering points and content, providers expanding their Edge should offer improved security gained from the ability to isolate their networks from DDoS attacks.

• As part of the Internet Edge, providers' data centers should be designed to N+1 standards and guarantee 99.99% uptime as well as redundant power and links.

Conclusion

Content delivery has changed the Internet as we know it. Growing traffic demands for mission-critical, bandwidth-intensive content and applications require the abandonment of yesteryear's network infrastructure and the creation of a new Edge of the Internet for traditional networks.

By deploying data centers and colocation facilities at the new Edge, providers can ensure their content is local, more secure, faster and produces a premium consumer experience. Many predict that infrastructure is returning to its glory days, whereby it held a significant key to the future of technology. We predict something very similar: that infrastructure will be as important

and game-changing as the technology that rides over it. It should be fun to watch.