Hybrid Edge: data at the right time, at the right place

Edge computing and the benefits of geography in resolving latency

Table of contents

FLEXENTIAL

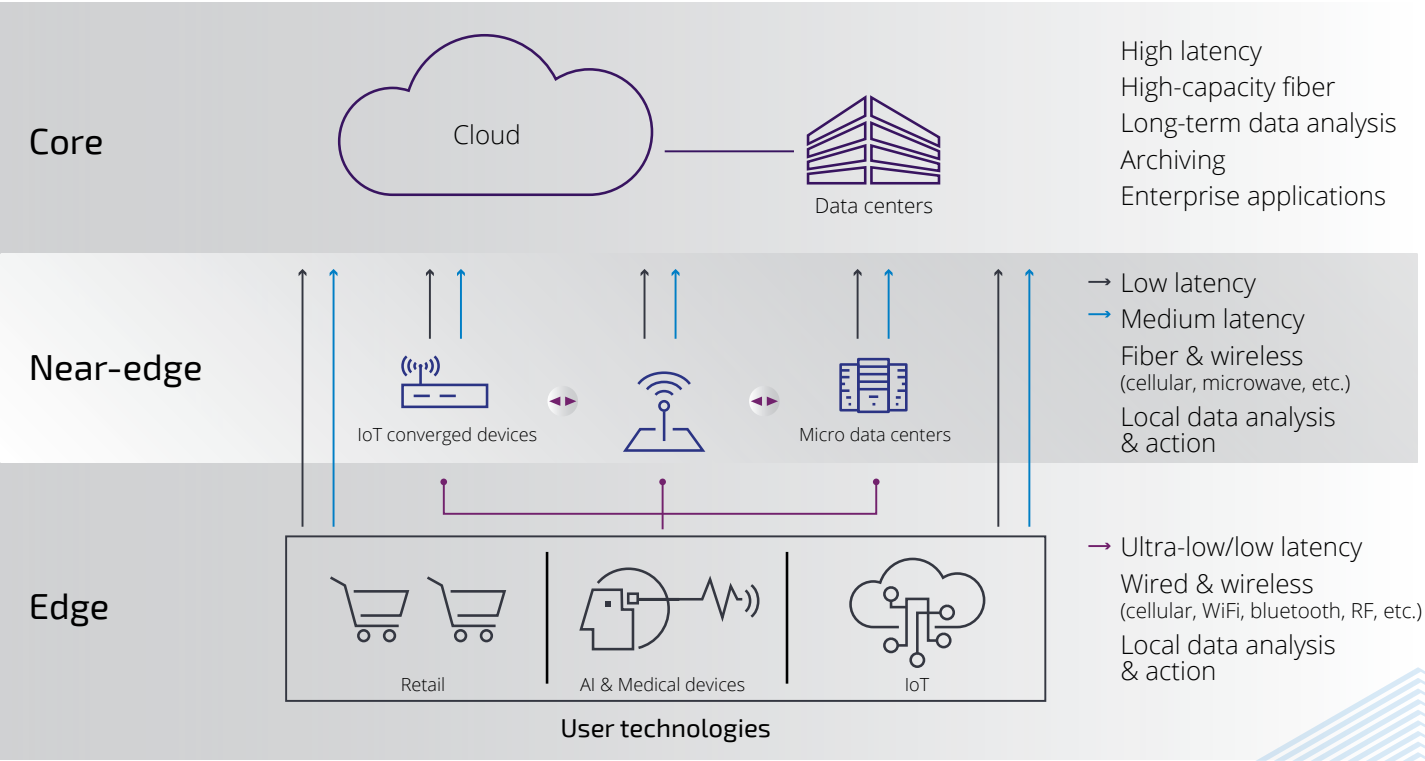
The edge, defined

The "edge" is emerging as a critical concept to reduce latency for network-based services in a world that has become increasingly centralized by public cloud services. Consumption habits of services and the need for analytics are shifting beyond core population centers and becoming local or even hyper-local within a region or city. As the online population continues to grow and new services emerge, the ability to handle data traffic securely and close to the customer or application will become a common pattern for the new service evolution.

The goal of the edge is to deliver the heaviest and most latency-sensitive data close to the edge of the internet and customers, often integrating with core or more centralized applications that exist in public cloud centers or corporate data centers. Edge data centers are connectivity-rich and carrier-dense, delivering critical business services over low-latency networks with response times of 5-7 milliseconds or less. Unlike the traditional network approach that delivers outbound content, the edge also supports bidirectional data precipitated by the Internet of Things and 5G cellular networks.

There are two "edge" data center types: near edge and far edge. Near-edge data centers are traditional data centers that are close to users. Far-edge data centers are composed of micro data centers generally thought to be at the base of cell towers.

Edge technologies and resulting "edge native" (following cloud-native applications) will change the common deployment patterns used by CTOs and developers today. Enhanced real-time data and near real-time decision-making should improve scale and availability, which are critical to remaining competitive in today's dynamic business services landscape.



Key technical aspects of the edge

Connectivity-rich. Critical to edge data centers is the ability to service a large number of carriers and service providers. Edge communications and services provide multiple opportunities to deliver data and those services.

This includes:

- Traditional data carriers
- Fiber providers
- Software-defined wide area network services

Proximity. As response time and speed of data transmission become more critical to a world engulfed in IoT, finance and gaming, latency should be 5-7 milliseconds or less. For IoT, 2-3 milliseconds is ideal. The only way to achieve this is to be close to the edge and the customers.

Resilience. Application availability and reliability may improve by shifting from fully centralized applications to a distributed, more intelligent data structure to support these architectures.

Business factors to consider

Digital experience expectations/mobility. Users will continue to consume data through mobile devices at an explosive pace. This will increase even more as 5G deployments boost bandwidth and rich media content. Cisco predicts that wireless and mobile devices will account for 63% of IP traffic.[1] Moving from one place to another and still accessing that content will drive bandwidth and define our edge experience.

Population shifts. In the past, focus has been in Tier 1 cities, but as we become a more mobile society and the population disperses from overly expensive, congested Tier 1 cities, Tier 2 and Tier 3 cities will be pressured to supply the demand of an increased population that is accustomed to Tier 1 services.

Data growth. The explosion of data due to social media, streaming and IoT is going to have an incredible impact on the internet. With volumes expected to surpass 3.3 zettabytes by 2021[2] , there are serious capacity concerns and impacts to transmitted traffic.

Bidirectional traffic. As we progress into the world of greater mobility and 5G capabilities, bidirectional traffic will be just as heavy as the streaming we are experiencing today. Growing traffic from IoT, artificial intelligence and social media will dominate bidirectional bandwidth. Delivering and receiving this wealth of data from a central location is inefficient and costly. Congestion and lagging transmission times will result in poor performance and user experiences, and potentially in the failure of critical services or functions.

Content. As content, such as ultra HD and 3D video, continues to evolve and become more impactful, low-latency, greater-bandwidth services that cannot be delivered effectively from centralized data centers will thrive.

Hyperscale density/locality. Latency is still the gating factor for getting data close to the user. Large hyperscale clouds work well for cost and efficiency, but they can't be everywhere and they don't solve the issues of Tier 2 and Tier 3 markets. Cloud environments need to be close to users and follow the distributed architecture.

Cisco predicts that wireless and mobile devices will account for 63% of IP traffic.[1]

1. Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, June 6, 2017.
2. Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, June 6, 2017.

IoT and AI. As our dependence on IoT and artificial intelligence continues to increase, so does our need to quickly and reliably transmit millions of packets to be locally analyzed and acted upon. This becomes even more critical as IoT devices and robotics manage our day-to-day actions, health and safety. A slow response or lag in communication can add seconds to a response or analysis that needs to be instantaneous. This demand could alter how applications correspond to changing conditions, such as manufacturing, safety systems and controls.

SD-WAN/SD-everything. Software-defined networking is quickly becoming a reality. Fast, automated deployment, coupled with new advances in AI will push network configurations to become more agile. SD-WAN will be a valuable tool in delivering certain types of connectivity. Software-defined services, regardless of the underlying infrastructure, will become the norm and will drive faster network-, server- and security-based deployments.

Minimizing latency is achievable. The solution combines data center location and edge computing.

Tips to ensure successful SDN implementations

1. Deploy closer to your customers, where geography can overcome the challenges of latency. Take into consideration population growth areas and the proximity to where data originates.

2. Push a distributed model to the edge. Take lessons from large content providers and apply best practices to your business applications or other services. Identify what services must be moved closer to the edge and deploy them. Consider data architectures that provide latency consistency and share data intelligently using location as a key attribute. Services to consider moving to the edge include:

    - Infrastructure as a service
    - SD-WAN
    - Object storage
    - IoT controls
    - Software as a service
    - Security

As we progress into the automated and highly complicated world of AI, deploying closer to the users of the data, whether human or machine, is critical. While IoT, self-driving cars and robots will use a huge amount of local processing power, there will always be a need for new analytical data feeds from an edge data center.

3. Figure out the last-mile delivery. How do you get that data from the edge data center to the customer or end device without interruption? Whether it's over cellular, the internet or a private connection, the delivery must be reliable with low latency. Define the business tolerances for technologies, such a SD-WAN in a metro region, and reach your remote locations with minimal overhead and low costs.

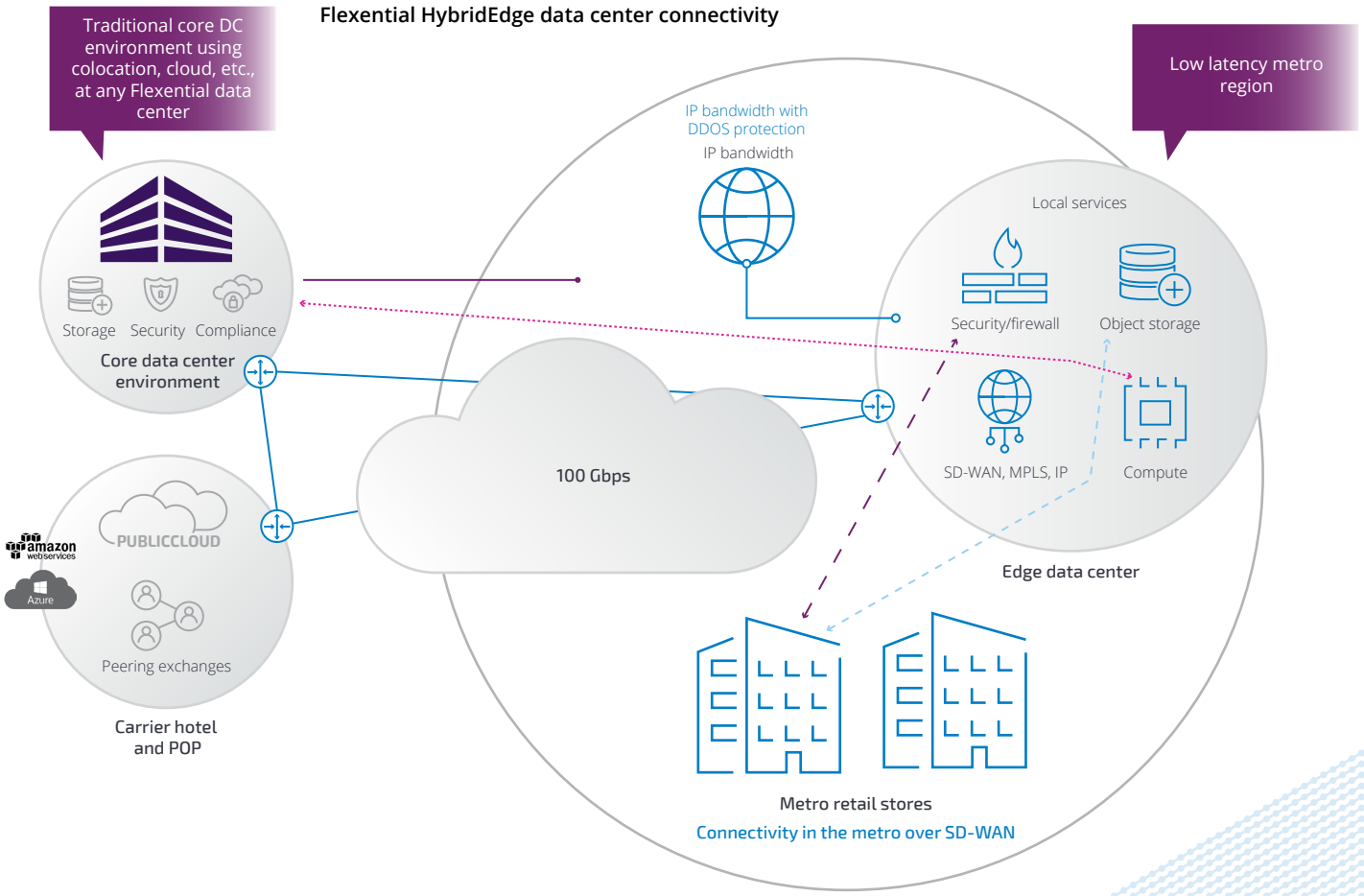
Hybrid edge data centers

As we develop our edge strategy, it is critical to understand the role of the data center in the edge ecosystem. Sometimes referred to as "near edge," the data center has evolved into a connectivity-rich ecosystem that allows a hybrid approach to compute, storage, security and carrier services. Without the ability to provide these services near the data users and transport them to a final destination, we cannot efficiently manage, store and analyze our data.

Local compute and security are relevant in an edge data center, and while both can be done from a large colocation facility or from the base of a cell tower, there are latency and scale limitations that cannot be addressed without impacting quality and efficiency, and greatly increasing costs.

Access to Big Data or public clouds is critical in today's world of data analytics and processing. Companies understand the value of having a multi-cloud solution and the need for hyperconverged data processing. A hybrid approach is required for a data center provider working in the edge.

Allowing this connectivity and tethering retail stores, cell towers or clinics to the hybrid edge data center efficiently brings services to and from the edge without losing the capability to utilize the large colocation environments.



Edge and the cloud

Cloud computing is a mainstay in today's high-tech world and will be here for quite some time. The practical uses of a virtual cloud environment have been demonstrated and highly monetized. But how does this dynamic, agile deployment integrate with today's edge trend?

Simply put, cloud environments — like applications — need to be close to the users and follow a distributed architecture. Large hyperscale clouds provide cost and efficiency benefits, but they can't accommodate the low-latency needs of businesses in more remote markets.

To be relevant in the shift to the edge, cloud environments need to adopt the same architectural principles as distributed computing. Small cloud environments will be required near the end user and must offer dynamic storage and the ability to communicate back to large-scale cloud deployments for fast, efficient local compute.

Micro clouds, similar to the micro data centers at the far edge, will be required to process and analyze IoT and AI data in a mobile world. Deploying these at the base of a cell tower or very close proximity will be mandatory.

Practical use cases

While a lot of attention is focused on IoT and AI, as it should be, providing IT services for today's applications is still in high demand. Service delivery and needs continue to increase at tremendous rates in the wake of the automated age, and businesses need to get in front of the curve before their efforts are consumed by accommodating these technologies.

Latency kills regardless of the intent. Holding onto the concept of a single compute infrastructure at a single colocation facility because moving to the edge is complex or not necessary may impact your competitiveness. Today's service providers need to have an edge strategy and embrace the distributed service delivery. Developers must understand how these new architecture paradigms share data and develop additional intelligence attributes that address location and latency as applications continue to scale to meet new demands. Additional data acquisition technologies related to IoT will continue to change these needs and drive new concerns that reveal new business models to stay competitive.

Retailer with multiple locations

For retailers with multiple, geographically dispersed locations, distributing customer-impacting workloads quickly and efficiently can be a challenge. Routing network communications from the centralized data center thousands of miles away is expensive and slow, and provides a subpar customer experience. This is particularly relevant as retailers shift to a more automated presence within their stores and increasingly utilize IoT devices.

Restaurants use in-store kiosks and tablets to allow patrons to order food. Diners upload the menu — complete with pictures and descriptions — select their items and place their orders. To ensure a positive customer experience, menu data must be accessed quickly, without any lag in upload speed.

The edge is all about being connectivity-rich, carrier-dense, and providing service offerings close to customers within respective regions.

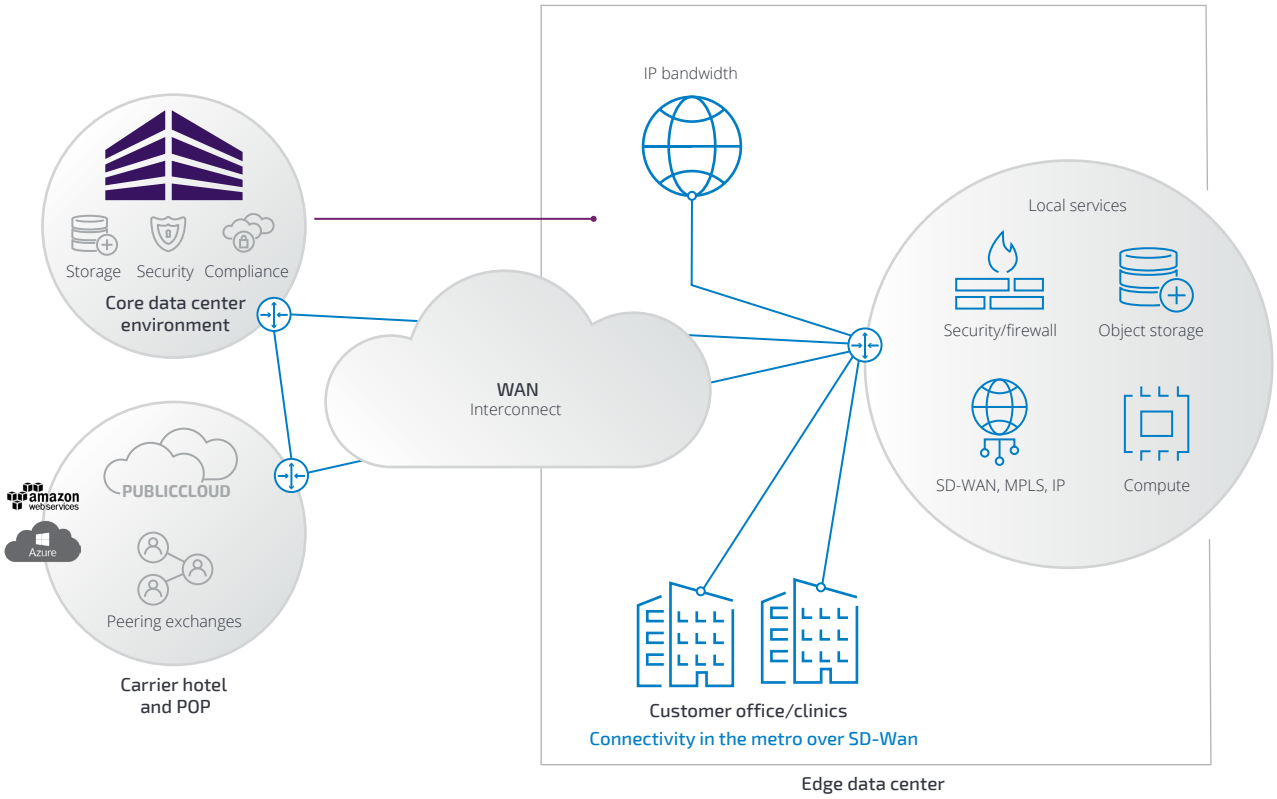
Retail stores have also adopted kiosk models allowing shoppers to research products, check availability and order from within the store. Delays in retrieving information can deter a customer from completing the order, impacting sales, the bottom line and the retailer's reputation.

To reduce latency and optimize the customer experience, retailers need to move content closer to the end-user. One way to do this is to install servers at every retail location. For a large retailer with thousands of stores across the country, this means deploying and locally managing, servicing and updating these servers at each location — a time consuming, resource-draining proposition.

The edge is a more efficient option. With an edge deployment, a retailer uses a small fleet of edge data centers in select metro areas to service a large subset of stores near the edge facility. This enables the retailer to manage compute and connectivity closer to the end-user instead of transporting it across the country to a core data center. This local deployment offers high availability and places point-of-sale information and other content-rich data near the user to minimize latency and enhance the customer experience. Information is securely and rapidly delivered from the edge to individual locations.

While performing some compute at the edge is appropriate, not all core business functions should reside at the edge. Some applications, services and data will need to be routed back to the main data center or public cloud. For example, historical data needs to be communicated back to the core data center to allow businesses to compile and analyze data from all locations to better understand customer preferences and maintain their competitive edge.

Accounting and payment processing data will also likely need to be routed back to the core compliant environment for managed IP with DDoS protections and PCI compliance. A select number of edge environments offer the availability, connectivity and scalability to manage



and process this compute on the edge and redistribute it to core environments when appropriate. It's important to seek data center providers that can quickly deliver this bidirectional traffic for a peak customer experience.

Healthcare organizations

The healthcare segment has been using IoT devices longer than most other industries, and is inundated with IoT and AI communications, data and images. Monitors are used to assess a patient's condition and report its readings back to a clinician; tablets are used to conduct office visits; implanted medical devices regulate body functions; and video-call appointments allow patients to remotely connect with their healthcare providers to address concerns. Even our personal devices — smart phones, watches and fitness monitors — can track heart rate and other health factors and report the compiled metrics to providers.

The ability to efficiently attain and distribute this abundance of health information not only impacts the patient experience, but may impact the patient's well-being. A tablet slow to upload a patient's information can result in a longer visit. A delayed video call can complicate communication. Lagging delivery of results from a monitor or device can impact care. When this mass of content is transported from the primary, centralized data center to unique healthcare providers or facilities, the network becomes congested, slowing response time and increasing the time needed to service patients. This data is also competing for bandwidth with the necessary security protocols and controls that are built into the network to protect patient confidentiality.

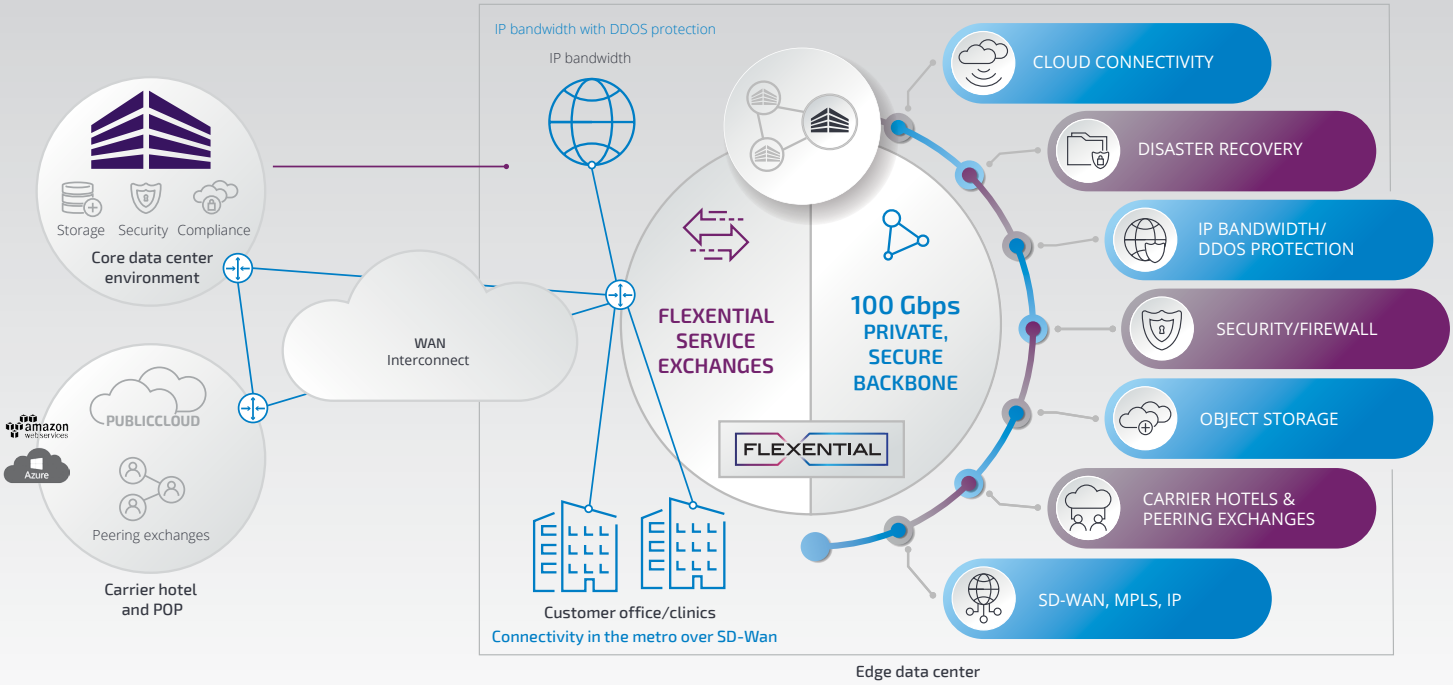
Again, the solution to reducing latency is to get the data and compute closer to the user. Rather than deploying servers to every healthcare facility or provider office, the edge offers a secure, high-availability, local deployment that is easier to manage and more cost effective.

Edge data centers allow content-dense information, including images using object storage services, to be processed and managed close to the healthcare provider for a faster upload and processing speed. While at the edge, transactions are secured with firewalls and IDPS to allow the healthcare provider to securely generate and process data, run reports and conduct payment transactions. Patient information is also encrypted at the edge to render it unidentifiable until it is communicated back to the main cloud environment.

An edge deployment also provides healthcare organizations access to hundreds of carriers, hyperscale cloud providers and their healthcare networks through interconnect products to enable efficient communication between diverse environments. Enabling this bidirectional traffic, in addition to powering compute at the local level, creates a dynamic environment capable of handling the intense processing, reporting and storage needs of the healthcare sector.

> The edge delivers the heaviest and most latency-sensitive data close to the edge of the internet and customers.

FLEX ANYWHERE



Flexential FlexAnywhere

Developing a full fabric of edge data centers to distribute or receive data from multiple locations is an important part of a successful edge strategy. Managing services at the edge, while connecting them back to a centralized compute or cloud environment delivers the low-latency applications and services without sacrificing the strength of large compute or hyperscale cloud environments for analytics, reporting, business services and archiving.

With a broad portfolio of interconnected edge data centers across the country, the Flexential fabric delivers high-capacity, low-latency and secure connectivity from the data center and cloud, to the edge. The Flexential North American 100 Gbps network backbone offers full-service, single-hop connections to get anywhere — eliminating extra network delays. High-speed, predictable throughput yields faster-performing applications and reduces costs by eliminating the need for multiple, complex network deployments. Offering access to a comprehensive range of domestic and international carriers; numerous cloud nodes located across the country; multiple peering exchanges; a host of network, data center, managed service and cloud providers, including Amazon Web Services, Microsoft Azure and Google Cloud; and direct network connections to Asia and the South Pacific via the New Cross Pacific and Hawaiki subsea cables, the Flexential FlexAnywhere software-defined infrastructure solution provides the flexibility and tools to elastically adjust capacity, easily integrate new technologies and move data and compute to the edge of the internet.

Flexential fabric delivers high-capacity, low-latency and secure connectivity from the data center and cloud, to the edge.

3. Cisco Visual Networking Index: Forecast and Methodology, 2016-2021, June 6, 2017.

FLEXENTIAL

Not science fiction

Edge computing is being driven by the production of massive amounts of data and the need for low-latency response and the transmittal of that data. As IoT and AI continue to gain steam as the next big waves of technology, so does the demand for low-latency, high-compute infrastructure close to end users. With enormous amounts of analytical data being pushed back into compute environments and the advent of 5G mobile technologies, the need for a hybrid edge solution becomes critical.

The need for data centers, both near and far, to accommodate the demands, growth and latency of this data-intense landscape can only be met at the edge. With data amounts expected to increase exponentially in the next several years, we need to build these data centers now, or risk losing our ability to progress and grow.

Annual global IP traffic is expected reach 3.3 zettabytes by 2021.[3]