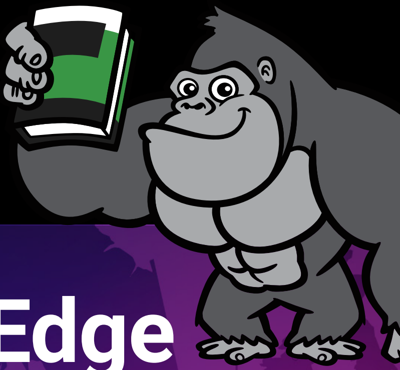
FLEXENTIAL PRESENTS

THE GORILLA GUIDE TO...®
EXPRESS EDITION



Hybrid Edge Computing

Tim Parker and Jason Carolan

INSIDE THE GUIDE:

- Your ticket to edge computing confidence and mastery
- Introduction to the concept of Hybrid Edge Computing
- Practical use cases for understanding how edge computing is being rolled out today
- Top tips for being successful at the edge

TAKE A QUICK WALK THROUGH THE IT JUNGLE!

Compliments of

FLEXENTIAL

THE GORILLA GUIDE TO...

Hybrid Edge Computing

Express Edition

AUTHORS

Tim Parker and Jason Carolan

Printed in the United States of America.

ACTUALTECH MEDIA

Okatie Village Ste 103-157

Bluffton, SC 29909

www.actualtechmedia.com

TABLE OF CONTENTS

CHAPTER 1

Edge Computing: What Is It?

The foundational principle of edge computing is that latency sensitive applications need to be as close to the consumer as possible. Proximity to the user is critical to ensure minimal response times. While the concept of edge computing is simple to describe, it is very difficult to execute. This is especially true of the emerging category of hybrid edge computing.

At its most basic, edge computing can be reasonably viewed as a repackaging of traditional on-premises computing. The concept of processing data as close as possible to where it was collected wasn't considered novel or controversial until cloud computing became a thing; workloads started moving out across the Internet, and started being restricted to a handful of large centralized data centers.

Unfortunately for the enterprise, public clouds aren't the solution to all things IT. One primary problem with public clouds is latency. Try as we might, we haven't yet managed to communicate faster than the speed of light, and the distance between physical premises where data is generated and the data centers of the large public cloud providers can simply be too far apart.

> The concept of processing data as close as possible to where it was collected wasn't considered novel or controversial until cloud computing became a thing.

Each workload has its own latency sensitivity. Latency is the combination of the time it takes data to get from the data generation source to the data processing application, the time it takes the relevant application to process the data, and the time it takes the result to travel to the destination.

As a general rule, latency higher than 100msec, or 1/10th of a second, is concerning. Latencies of 300msec or higher usually mean that an application cannot be considered to serve data in real time. As applications get more complex and require more and more data sources, this problem multiples. Applications within the same data center enjoy a sub-ms or 1-2ms range, where metro traffic between facilities may be slightly higher, but well below 3-5ms.

For many public cloud customers, these limits don't initially seem like a problem. Corporate HQ is normally located in a major city, and connected to the Internet via fiber optic connection. Sitting at one's desk at HQ and pinging the nearest public cloud provider data center will usually return latencies below 100msec. So why does anyone care about edge computing?

Machines Talking to Other Machines

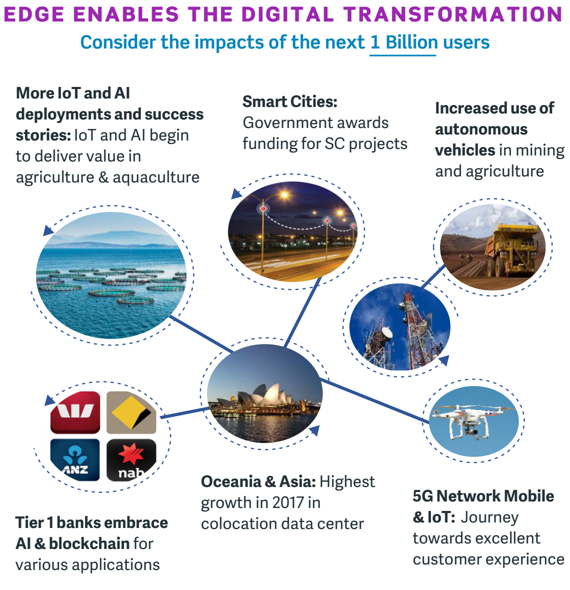
There's more to IT than supporting accounting applications so the CEO can peruse sales reports on his or her desktop. Not only are these types of human-focused applications rarely latency sensitive, but they make up an ever-decreasing share of the IT footprint that today's organizations are charged with maintaining.

One aspect of this shift is Machine-to-Machine (M2M) communication, and M2M workloads are far more likely than human-focused workloads to occur outside of a major corporate office. More importantly, M2M workloads can be extremely sensitive to latency. M2M workloads can include everything from the instrumentation of manufacturing equipment to logistics, and even supply chain management.

Anywhere there's a sensor generating data, or data that needs to be exchanged between two organizations (or even between departments within the same organization), you'll find M2M communication. M2M communication occurs when someone buys a widget off of a website, and that website updates the point-of-sales database with a newly-generated invoice, flags production with a new order by notifying the manufacturing platform to build a new widget, and changes inventory status to reflect the purchased items. Increasingly, manufacturers may have additional sensors to adjust and monitor for quality issues, ensuring that the end product meets customer needs.

M2M communication pervades every facet of modern life. It's often invisible, but crucial. A great example to think about: M2M communication occurs when an ambulance approaches an intersection, and the stop-lights need to change to allow it to pass through un-hindered. It's present in modern cars (both with and without drivers) that communicate among each other to prevent collisions. It's likely even present in your local bakery, ensuring that each loaf of bread emerges from the oven perfect every time.

A fraction of a second is a long time to an ambulance hurtling down the road at 60 miles per hour. The safe stopping distance for a vehicle at these speeds is 300



Courtesy of Vertiv

feet, and getting trapped waiting for traffic lights to cycle can cost precious minutes… and even lives. In the real world, even 100msec of latency matters.

Edge Computing Comes in Flavors

It's easy to understand the importance of latency when the use case is an ambulance trying to save lives. It's also just as easy to write that use case off as an (ahem) edge case, and proceed with the assumption that the overwhelming majority of workloads don't need to be processed anywhere near where the data is generated. Very few workloads are quite that latency sensitive, after all; but assuming edge computing is a niche is both reasonable, and wrong.

Even if your workloads don't exhibit real-time requirements, or extreme latency sensitivity, there are a few other reasons why you might want to process data outside a centralized public cloud or service provider cloud. Popular reasons include concerns about privacy, security, data sovereignty, intellectual property, trade secrets and data gravity.

In some cases, edge computing is viewed as a necessity more for throughput reasons than for latency reasons. Even if the latency to a public cloud provider doesn't bother your application, the data to be collected – and/or the data generated – could result in huge monthly Internet access fees.

The diversity of reasons driving edge computing have led to different approaches to edge computing. Beyond the traditional on-premises IT, there are several broad categories of edge computing, each with subtle variations.

- One class of edge computing uses micro data centers which are often located in telco central offices, and at the base of cell towers. These are in high demand, and therefore costly. This class of edge computing is often referred to as metro edge, CO edge, or neighborhood edge, where "neighborhood edge" is usually restricted to base of cell tower deployments.

- The other class of edge computing uses a distributed network of more conventional data centers, but with advanced technologies such as software-defined networking (SDN), software-defined wide area networking (SD-WAN), and software-defined infrastructure (SDI, also known as Infrastructure as Code) to route data and workloads to the nearest data center. This approach ensures that data and workloads live as close to where demand is greatest as possible, rather than in a single centralized location, and is often called "fog computing."

This class of edge computing has given rise to hybrid edge computing.

Hybrid Edge Computing

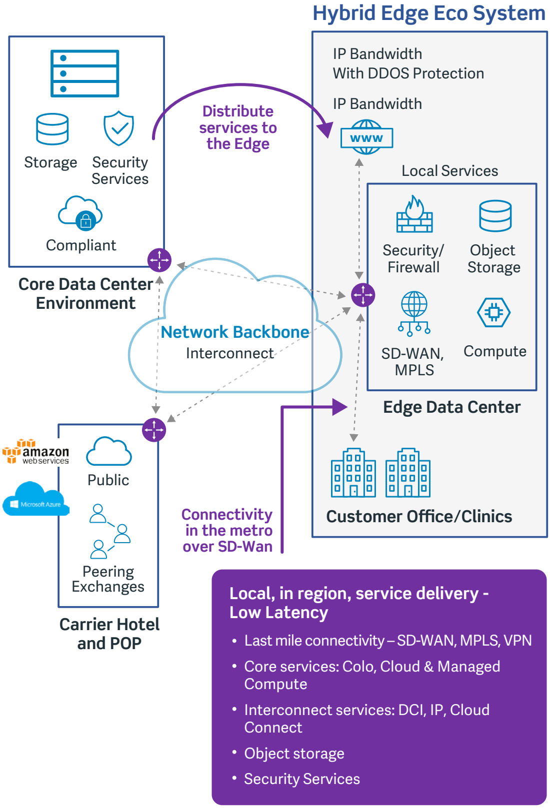
Edge computing differs from traditional on-premise computing because the products used at the edge must have remote management capabilities. Today, many companies consuming edge services may have to physically deploy and manage such resources. In the future, a customer using edge computing resources will only see what looks very much like a public cloud interface or API.

This is deliberate: much of what's offered as edge computing is being offered by public cloud providers who want to solve the latency problem without losing customers. Having customers subscribe to IT services is a much more attractive revenue model than selling them servers, so providing fully-managed services geographically close to demand is simply good business.

The edge computing clusters themselves, however, are designed to live in someone else's network. They are designed to exist in shared environments, and often contain carefully-guarded intellectual property that the cloud providers don't want getting out. Amazon, for example, probably doesn't want anyone peeking under the hood and cloning their proprietary services, like Rekognition.

This approach to managing edge computing equipment separates edge computing from traditional on-premise IT.

Like edge computing itself, hybrid edge computing will be cynically dismissed by many as merely a



buzzword, repackaging that which existed before. At first blush, hybrid edge computing sounds a lot like classic colocation and service provider hosting, which it is. Like the more general edge computing, however, hybrid edge computing is an old concept with a new twist.

In addition to software-defined everything, hybrid edge computing relies heavily on service exchanges and carrier interconnection to deliver IT services with the lowest possible latency. With traditional service provider hosting, organizations treated the workloads placed in the hosting provider's data center as islands unto themselves. If those workloads needed to communicate with other workloads, they would typically interact with other workloads located within the service provider's network, or with workloads located on one of the major public clouds: in other words, data center-to-data center communication with enormous fibre optic connections available between source and destination.

Hybrid edge computing needs to be far more aware of how modern, distributed applications work. Hybrid edge data centers need to have interconnection with every carrier possible, because the data sources (and possibly destinations) may be coming from mobile networks, or even consumer broadband networks. Sensors can live anywhere, and they'll use whatever Internet connectivity is convenient.

Hybrid edge providers aim to be “one network hop from anywhere.” This focus on connectivity, when combined with dynamic, software-defined connectivity, infrastructure, and service provision make hybrid edge computing as different from traditional service provider hosting as public cloud computing is from traditional x86 virtualization.

CHAPTER 2

Who Needs The Edge? Practical Use Cases

As organizations develop their edge computing strategies, it's critical to understand the role of the service provider data center in the edge computing ecosystem. Sometimes referred to as "near edge," the service provider data center has evolved into a connectivity-rich ecosystem with a dynamic, software-defined approach to infrastructure.

Infrastructure elements being automated and managed with infrastructure orchestration applications include compute, storage, security, hosted service provisioning, trans-infrastructure layer 2 networking and carrier services.

While there are multiple approaches to edge computing – and thus multiple places where edge workloads can live – there are latency and scale limitations that make hybrid edge computing an attractive approach. Telco micro data centers have severe restrictions on scale.

Attempting to resolve the scale issue by bursting some parts of a workload to the public cloud during periods of high demand dramatically increases latency. Latency has a dramatic impact on the quality and efficiency of edge computing projects. Any edge computing project using telco micro data centers or the network extrusion model runs the risk of inaccurately assessing demand, missing the mark on scale requirements, and ultimately greatly increasing costs.

> Sometimes referred to as "near edge," the service provider data center has evolved into a connectivity-rich ecosystem with a dynamic, software-defined approach to infrastructure.

A Bridge to the Cloud

For today's organizations, access to big data or public clouds is non-negotiable. Public cloud providers provide the more effective, easiest-to-use versions of critical IT services ranging from facial recognition to big data analytics. Unfortunately, the real world tends to get in the way.

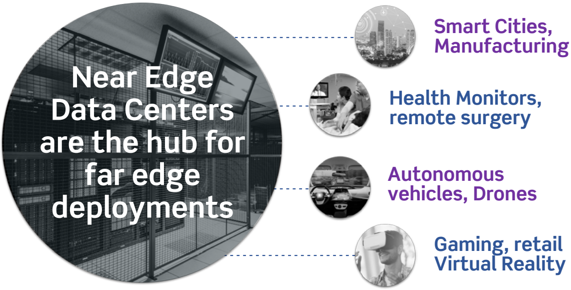
Adequate Internet connectivity may be too expensive or even unavailable in many areas data needs to be connected. Another consideration is that many last-mile telecommunications providers charge significantly more for IP traffic between networks than they do traffic that's entirely on-net. In other words, while it might be a challenge to afford an Internet connection big enough to shift all the bits you want, chances are better that a high-throughput network connection to a hybrid edge service provider would be affordable, because hybrid edge providers focus heavily on carrier interconnectivity.

Peering agreements and network congestion between carriers and backhaul providers can also be a roadblock. What may be a 100msec round trip for a sensor attempting to communicate with a major public cloud provider can be a 10msec connection to the nearest hybrid edge data center, which is itself only 10ms away from the desired public cloud provider.

The Internet is a complicated place. Hybrid edge computing doesn't simplify it, but it is a rational approach to coping with that complexity.

Use Case: Distributed Multi-Cloud Analytics

Two of the biggest advantages of hybrid edge computing: 1) The data centers of hybrid edge service providers are connected by astonishingly wide network pipes. 2) Hybrid edge service providers offer SDN



and SD-WAN capabilities. These two features can be combined to solve some tricky real-world problems.

Picture an oil pipeline that runs across North America. This pipeline is festooned with sensors, and in a perfect world, the data from these sensors would be analyzed in real time. Seconds – and perhaps even fractions of a second – can be the difference between cutting off the flow of oil when an irregularity is detected, and an ecological disaster that results in hundreds of millions of dollars' worth of fines.

Organizations like the hypothetical oil pipeline operator have latency-sensitive workloads that are reliant on data inputs spanning vast geographic areas. Reducing latency starts with having somewhere geographically proximate to send the data. As discussed earlier, hybrid edge service providers are an excellent place for that data to land.

Data can be sent from sensors to the nearest hybrid edge service provider Point of Presence (PoP). This data can then be ingested into a database instance that lives at that point of presence, and the data synchronized with all other database instances. Together, these database instances form a distributed data warehouse, with the minimum possible latency between geographic regions. The connections between the service provider's data centers will be faster than travelling over the unmanaged Internet.

Data from this distributed database can then be analyzed using proprietary services made available by the public cloud providers. The hybrid edge service provider will have high throughput, low latency connections between their network of data centers, and that of the major public service providers.

In fact, because the hybrid service provider is a centralized storage location with high quality connectivity to all the major public cloud providers, the data in the data warehouse can be analyzed using services from *all* the major public cloud providers. This allows hybrid edge service providers to offer value that wouldn't exist even if there was low latency, adequate throughput between all sensors and one of the major public cloud providers.

In this use case, the difference between using hybrid edge computing and the public cloud via the unmanaged Internet is somewhat like shipping a package via FedEx Next Flight versus FedEx Economy. The data will still get there, but there are some pretty big differences in the time it takes to do so.

Use Case: Retailer With Multiple Locations

For retailers with multiple, geographically dispersed locations, distributing data and workloads quickly and efficiently can be a challenge. Routing network communications from the centralized data center to a location thousands of miles away is often both expensive and slow. In addition, the difficulties of managing large retail operations are often magnified by the lack of onsite IT staff.

These issues are compounded by the economic and technologic pressures that retailers are facing. Retailers are no longer simply competing with other local retailers, but with the whole of the Internet. For those selling goods, Amazon, Walmart and other giants have changed that space forever, acting as both fierce competition, and a new outlet through which to sell goods.

Restaurateurs, similarly, have found themselves both challenged and emboldened by services such as SkipTheDishes, JustEat, and GrubHub. Even McDonalds, which exists on seemingly every corner, has given in. One can now order McDonalds via online food delivery applications. Change is inevitable.

Competition is driving retailers and restaurateurs not only toward embracing an online presence and delivery, but toward more automation within their physical locations. To accomplish this, these organizations are increasingly turning to Internet of Things (IoT) devices.

Putting aside any latency-sensitivity issues, both trends discussed earlier mean more IT widgets in every store and restaurant. These locations are unlikely to have on-site IT staff, so a reliable means must be found to manage and support the diversity of technological devices in use, while providing corporate IT services to those locations in a secure manner. The intelligence will have to be in the network and the edge, not in the store.

Many retailers and restaurateurs are turning to hybrid edge service providers to solve this problem. Locating corporate workloads at the nearest hybrid edge computing data center is often "close enough," without physically putting equipment onto the same network as an unknown (and growing) number of insecure – and unsecurable – IoT devices.

The IoT devices in question may be units that the organization itself may own, such as in-store kiosks and tablets to allow patrons to order food, or units allowing shoppers to research products they find in store. The IoT devices may also themselves be edge computing devices using the network extrusion model. Examples of this include systems provided by logistics providers to allow shipping labels to be created, or by meal delivery startups to allow restaurants to receive and respond to orders.

This is without even touching on all the many and varied technologies being used for inventory management, customer tracking and so on.

Retail stores and restaurants both face an edge computing trifecta of problems. They're increasingly deploying services and technologies into their locations that customers use, and where latency thus impacts the customer experience. They're deploying products that they don't have the ability to properly secure, and they don't have the margin to afford a sysadmin for each location.

Faced with this conundrum, a geographically proximate, secure data center, manned by experienced systems administrators, with a focus on high-speed, low latency interconnection, starts to look like a good idea.

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.

Our personal devices — smartphones, watches and fitness monitors — can track heart rate and other health factors and report the compiled metrics to providers. Even at-home medical devices such as CPAP machines and pacemakers regularly phone home with data.

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, and do so in different ways. A few extra seconds calling up patient records may not impact the

Even a second or two matters if your pacemaker is the medical IoT device calling for help; and it won't always be doing so while you're in a hospital that's connected to the rest of the world with redundant fibre connections.

average doctor's visit, but it can be life-and-death in the operating room.

Similarly, while a few seconds may mean nothing to either patient or doctor during the yearly physical exam, this discussion changes as it scales. A few seconds, multiplied by tens or hundreds of millions of patient visits, starts adding up to real money.

Once the discussion moves away from the well-managed confines of medical facilities, the importance of edge computing to the healthcare sector becomes even more apparent. Even a second or two matters if your pacemaker is the medical IoT device calling for help; and it won't always be doing so while you're in a hospital that's connected to the rest of the world with redundant fibre connections.

Healthcare data management — including sharing, privacy, security and backups — is an important topic. Medical facilities tend to generate a lot of imagery that not only needs to be stored, but needs to be easy to retrieve when doctors want to look at it. The doctor requesting an image is often in a different facility than where the image is generated, and they may rely on third-party services to perform analysis on that data in real time.

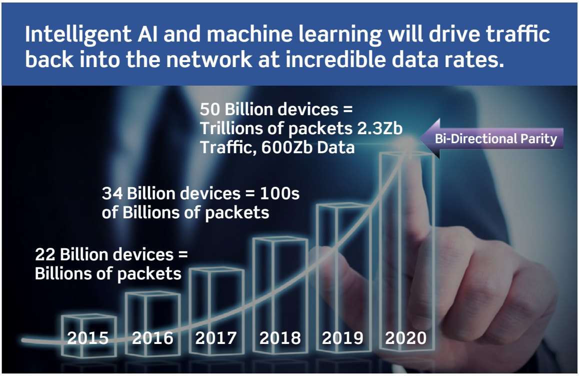
All this data has to be secured, backed up and monitored for unauthorized access. This all has to occur in facilities that meet local, national – and potentially international – regulatory requirements. It's becoming increasingly unrealistic to meet these requirements from the local hospital's server room.

CHAPTER 3

The Challenges Driving Edge Computing Adoption

It's easy to understand the utility of edge computing when broken down into specific use cases, but there's also value in looking at high-level trends. Technological progress waits for no one, and as technology marches on, so too must organizations adapt to an ever-changing reality.

To remain relevant as the world figures out how to make use of edge computing, IT teams and even cloud providers need to adopt the same architectural principles as distributed computing. Small, automated, self-service IT environments will be required near the end user. These must offer dynamic, secure, regulatory-compliant storage, and have the ability to communicate back to the more centralized and larger-scale IT environments.



Digital Experience Expectations/ Mobility

In what won't come as a surprise to anyone, individuals and organizations are predicted to increase consumption of data through mobile devices at an explosive pace. Our insatiable desire for data is likely to increase as 5G deployments offer increased bandwidth. Cisco predicts that wireless and mobile devices will account for 63% of IP traffic by 2021.[1]

As wireless data usage increases, so to do the challenges that mobility brings. While some devices will remain

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

fixed in place, using mobile networks only because they're the most convenient, an unknown number of existing and novel data sources will move.

The movement of these data sources may be practically irrelevant. For example, despite their inherent mobility, IoT-tagged cows aren't likely to move far enough to switch cell towers. On the other hand, this mobility may be significant for activities like parcel tracking, or managing fleets of delivery trucks which move regularly between cities.

Data Growth and Bidirectional Traffic

That our demand for both data storage and network throughput perpetually increases is both an overused trope of tech marketing and one of the daily challenges IT practitioners face. Cisco predicts that the amount of traffic transiting the Internet and IP WAN networks will reach 3.3 zettabytes by 2021.[2] That's a huge, almost absurd number. It's difficult to wrap one's mind around it, except to say that it's rather a lot more than today, even though 2021 is no longer all that far into the future.

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

INTERNET TRAFFIC

HAS INCREASED **5x** in the past 5 years

EXPECTED TO GROW **2.3ZB** by 2020

21B

Active devices powering people, machines and the internet of things **will generate approximately**

600ZB

of data by **2020**

MOBILE DEMAND WILL INCREASE

51% of millennials can't go more than 3 hours without checking their phones

55% of millennials watch streaming video several times a day, on various devices

MOBILE VIDEO STREAMING ACCOUNTED FOR

60%

of all mobile device traffic in **2016**

EXPECTED TO RISE

78%

in **2020**

**Data obtained from Vertiv*

As we progress into the world of greater mobility and 5G capabilities, bidirectional traffic will become a more serious consideration than the streaming-dominated Internet of today. More devices will be sending data to the hybrid edge and to the cloud. 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.

Software-Defined Everything

Software-Defined Infrastructure (SDI) is quickly becoming a reality. Infrastructure automation and orchestration are being married to new advances in AI. The result is that everything from infrastructure monitoring to network configurations are becoming not only automated, but responsive to change in that automation. SD-WAN in particular has proven to be a valuable tool in delivering certain types of connectivity.

Software-defined everything is the norm for public clouds, and the ease of use they demonstrate is setting expectations for all other IT services. A particularly important consequence of this is that manual network and security configuration is no longer considered acceptable.

While other IT operations teams, such as storage, have had the past decade to adjust to a world where they're required to provide instantaneous provisioning and change request fulfillment, networking and security have remained stubbornly hands-on until recently. Agility is the name of the game today, an expectation complicated by the increasing complexity of today's hybrid multi-cloud networks.

CHAPTER 4

Being Successful at the Edge

Surmounting the mountain of challenges presented by modern IT, and the adoption of edge computing in particular, requires a specific set of tools. In addition to the automation and orchestration of software-defined everything, edge computing providers need to support – and integrate with – various public, private and hybrid clouds.

Connectivity is essential for edge computing, but interconnection with as many providers as possible is critical to hybrid edge computing in particular. Managed network services, including layer 2 extensibility, QoS and DDoS protection are important features to make complex, geographically distributed networking easy to use.

Last-mile connectivity, traditionally done through, MPLS, VPN or private carrier will increasingly need to incorporate SD-WAN. As the number of locations where data is ingested grows, the importance of SDN and SD-WAN increase.

> Connectivity is essential for edge computing, but interconnection with as many providers as possible is critical to hybrid edge computing in particular.

Organizations should ensure that the hybrid edge providers they select have a network of partnerships with both carriers and other data center providers which allows them to create a broad, geographically distributed presence. Getting the data processing closer to the data source is the point of edge computing, and that's easiest when there are numerous of points of presence to choose from.

Regulatory compliance certifications and approvals, backup and disaster recovery services in all the colors of the rainbow, and "one-throat-to-choke" support and billing are also nice-to-haves. They help discriminate the serious hybrid edge service providers from those simply attempting to find a new name for classic hosting.

Flexential FlexAnywhere

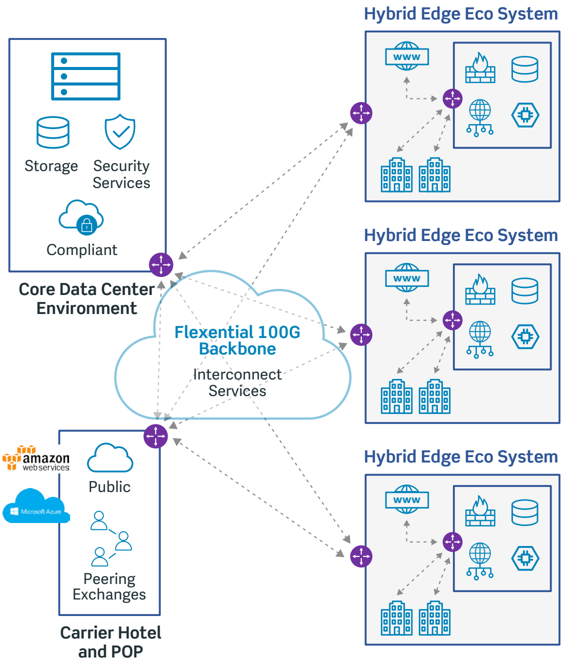
Developing a full fabric of edge data centers to distribute or receive data from multiple locations is an important part of a successful edge strategy. Managing services at the edge, while connecting them back to a centralized compute or cloud environment, allows for the delivery of low-latency applications and services. Flexential FlexAnywhere is a hybrid edge computing fabric that delivers performance, low latency and reliability without sacrificing the ability to scale.

With a broad portfolio of interconnected edge data centers across the country, Flexential's FlexAnywhere delivers high-capacity, low-latency and secure connectivity to the edge. FlexAnywhere connects the edge to corporate data centers, carrier hotels and public clouds.

The private and secure 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.

FlexAnywhere offers access to a comprehensive range of domestic and international carriers and cloud

FLEXENTIAL FABRIC



services. Flexential has numerous managed private cloud nodes located across the country. Flexential also offers colocation, peering exchanges as well as public cloud access, including Amazon Web Services, Microsoft Azure and Google Cloud.

FlexAnywhere has direct network connections to Asia and the South Pacific via the New Cross Pacific and Hawaiki subsea cables, and the FlexAnywhere SDI 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 and FlexAnywhere provides the infrastructure framework for hybrid edge computing fabrics that will connect organizations to sensors, people, public clouds, other organizations… and the future.

For more information on the FlexAnywhere fabric, visit **https://www.flexential.com/flexanywhere**, or call **888-552-FLEX** to speak with one of our experts today.