451 Research® | PATHFINDER REPORT

Architecting Hybrid IT and Edge for Digital Advantage

COMMISSIONED BY

EQUINIX

JANUARY 2020

About this paper

A Pathfinder paper navigates decision-makers through the issues surrounding a specific technology or business case, explores the business value of adoption, and recommends the range of considerations and concrete next steps in the decision-making process.

ABOUT THE AUTHOR

CRAIG MATSUMOTO

SENIOR ANALYST, DATACENTER NETWORKING

Craig Matsumoto is a 451 Senior Analyst focusing on the confluence of CDNs, interconnect fabrics and cloud access. Craig has covered service-provider and enterprise networking since the dot-com bubble of 1999, including more than 10 years at Light Reading, where he covered broad topics including optical networking, routing and the then-new beat of software-defined networking. He also spent four years at SDxCentral, delving further into SDN, NFV and container technologies

Executive Summary

The enterprise network must evolve for the new era of hybrid IT. In order to meet new digital demands, enterprises increasingly rely on IT that is housed outside of the traditional datacenter, whether that means using a public cloud, a hosted private cloud, a SaaS provider – or edge computing. As the deployment options for hybrid IT continue to transform core business processes, enterprises are being pushed closer to the edge, spurred by oncoming trends such as edge analytics, 5G and IoT. These shifts have resulted in new distributed models, and as a result, connectivity across ecosystems, partners and SaaS applications has become the enterprise's lifeblood and cannot be left to the vagaries of the public internet.

In this new world, distance has become the silent digital-business killer. Enterprises can respond by turning to a new architecture based on interconnection and leveraging colocation in multi-tenant datacenters. Because these datacenters are already home to the strategic network, cloud and business ecosystems that enterprises rely on, this strategy can improve performance while simultaneously reducing transport cost. Think of this combination – colocation plus interconnection to rich ecosystems – as a way to regain control over all this networking. Not only does this architecture provide the lowest latency to the largest number of destinations, but since these connections are private, they bypass the internet and should provide a boost in reliability.

The strategy can be furthered by the advent of software-programmable interconnection that is customizable through APIs, globally. This fabric of programmable connectivity creates a geographic abstraction layer, allowing access to the new markets that business demands. The ability to reach markets globally and interact locally has become the standard for digital business.

The New Shape of the Enterprise Network

IT Has Dispersed

Enterprise IT has entered the era of the cloud and the edge. Traditional IT was based on a centralized philosophy – concentrate the majority of IT infrastructure with employees supporting the back-end business process. The modern IT organization does not have the option to backhaul traffic to a headquarters-based datacenter in this way. Organizations require proximity to the cloud for low-latency applications while simultaneously providing edge computing to support user experience and analytics at the edge. The days of the enterprise datacenter are giving way to a distributed model based on virtualized and hyperconverged infrastructure, consumed on a pay-as-you-go basis. For many enterprises, this transformation has evolved on its own, without central guidance or a long-term plan. As a result, some companies are only starting to understand the extent to which they depend on outside resources, including public and private clouds and SaaS providers.

The promise of the cloud was compelling, and it seemed to offer the fastest solution to this distributed problem. Early adopters quickly discovered that as their businesses scaled, unforeseen costs could break their financial models. Generic compute could not efficiently support high-transaction applications, resulting in unplanned licensing and support costs. As multicloud application demand grows, data transport dependency also grows, often resulting in unexpectedly high egress fees for data stored in the cloud.

For these reasons, enterprises are moving some workloads out of the cloud and gravitating toward a hybrid IT model. But backhauling all edge traffic to a centralized datacenter, as in the old model, is not sustainable. In addition, real-time user experiences, edge analytics, workforce enablement and growing customer demands require distributing work and data to more locations connected to a wider variety of business partners. Edge computing placed in multi-tenant datacenters with direct access to ecosystem partners can satisfy these requirements.

Latency looms large, especially for high-performance edge applications, IoT and 5G use cases. The distance to critical resources and users is becoming a crucial factor, one that enterprises can address by taking advantage of edge computing in ecosystem-dense colocation facilities.

The Rise of the Digital Edge

As some infrastructure and services move closer to users – whether they are customers or employees – enterprises are forced to manage distinct classes of workloads. Core business workloads are often composed of multiple parts, including front-end applications that have to scale capacity; high-transaction data warehousing that has to scale up to achieve performance thresholds; and integration with ecosystem partners. For many businesses, this is true of their customer relationship management (CRM), enterprise resources planning (ERP) and other packaged core business applications. Pieces of these applications, then, are suitable to stay deeper in the network 'core' – inside the cloud or in a centralized datacenter.

Other workloads require more strategic placement outside the cloud, at some form of edge, often to gain the advantages of lower latency. Customer-facing functions that enhance the user experience, for example, might require near-real-time responses that are best achieved by removing the distance to the cloud. The same could be true of certain employee productivity functions, where 'productivity' depends on getting rapid responses from an application.

We included more examples, organized by vertical, in the figure below. As shown in these examples, certain verticals or use cases may lean more toward core or edge, but many industries will find a useful mix.

Optimal workload location varies by use case/vertical
Source: 451 Research

MANUFACTURING	
PRODUCTION/MANUFACTURING MONITORING	Edge
INTELLIGENT LOGISTICS	Core
FLEET TRACKING	Core
CONNECTED WORKER	Edge
TRANSPORTATION	
FLEET TRACKING/TELEMATICS	Core
SUPPLY CHAIN/LOGISTICS	Edge
PREDICTIVE MAINTENANCE	Edge
SEMI-AUTONOMOUS DRIVING	Edge
RETAIL	
FRAUD DETECTION	Core
RESPONSIVE CUSTOMER FLOW MANAGEMENT	Core
CUSTOMER FOOTFALL TRAFFIC	Edge
DEMAND-DRIVEN WAREHOUSING	Edge

Colocation and the New Enterprise Network

The enterprise datacenter, or on-premises enterprise IT, used to form a hub-and-spoke model, in a logical sense. Employees went to 'the network' to get information or to perform tasks, and the in-house IT group, or a managed-service equivalent, was responsible for keeping that infrastructure humming.

As IT distributes outward, it is obvious that the network should become distributed as well. One easy model is for employees to access the cloud over existing broadband – that is, over the internet – and in fact, this is how many enterprise cloud journeys began. But as clouds and edge locations become vital to the enterprise, the connectivity to those locations becomes equally vital. The open internet won't always be the best option for reasons such as performance, compliance, security and even cost.

One emerging strategy is to seek out private connectivity to replace the internet. In the case of public clouds, services such as AWS Direct Connect and Azure ExpressRoute are the most direct options available to most enterprises. These services work by establishing an on-ramp – a deployment of equipment inside a neutral, third-party datacenter – that acts as the gateway into a cloud datacenter. To use Direct Connect or similar services, the enterprise needs to connect into the third-party datacenter, possibly with the help of a telco or network service provider.

Rather than repeat this process for multiple services, the enterprise could choose colocation inside a multi-tenant datacenter and benefit from the ability to connect with the other tenants. In this scenario, the datacenter hosts the cloud on-ramps, providing an intersection for the networks of telcos, managed service providers and other large enterprises. These parties are attracted to the datacenter partly as a way to interconnect to one another, and colocation would be a way for the enterprise to join this game. Connectivity to fellow colocated parties would occur not across the internet, but within the walls of the datacenter (or at least within a private fabric, as we'll explain below).

The enterprise's colocation deployment would become the networking hub (logically, not physically) for the company. A connection into that hub could be extended, within the datacenter, to a cloud or SaaS on-ramp – or to service providers whose networks can further connect to the destinations not represented inside the datacenter. The hub would become the enterprise's Grand Central Station, with traffic arriving from the enterprise's offices and users (likely over Ethernet, MPLS or SD-WAN) and 'transferring' to reach other destinations.

As with the real Grand Central, the key here is the network effect of being able to connect to many destinations. The single 'doorway' into the colocation hub leads to multiple networks. This could be particularly useful for edge computing, where resources placed closer to the user would need to find connectivity back to a cloud or to an enterprise location. And with pathways available to reach many destinations, the enterprise would have flexibility, making it easy to accommodate any changes to the clouds or services being used.

This strategy could also include connectivity to other enterprises. Partners and suppliers are beginning to interact by opening their networks to one another, taking advantage of private, controllable connectivity. Consider the sharing of data between healthcare-industry partners, for instance; ideally, this would be done over secure, private connections. And if enterprises could use this colocated hub to connect to one another, then by default, this could also become a way for a B2B enterprise to connect with customers or suppliers.

The Grand Central Station analogy has a catch: a train station, no matter how grand, serves only one metro area. Grand Central, in New York, offers no benefit to a commuter in Tokyo or London. By the same token, our discussion of cloud connectivity has implied that destinations such as cloud on-ramps are present in the same datacenter as the enterprise's colocation deployment – or at least within the same campus of interconnected datacenters. The datacenter-hub concept becomes more powerful if we can stretch it across geographies, bringing more destinations into the datacenter in a virtual sense. We need to connect distributed facilities, including remote edge-computing sites, in a seamless way that makes them feel like one solution – the digital edge.

Software-Programmable Interconnection

This is where network virtualization and software-defined networking (SDN) come in. The technologies differ under the hood, but for our purposes, what's important is that they make the network pliable. Through software commands, virtual cross connects can be assigned within minutes – deploying a connection between, say, an enterprise's colocation hub and an AWS Direct Connect on-ramp – and the connection can be turned down when the task is completed. The key is the automation behind this process. Connectivity that previously took weeks to provision can now be satisfied on demand.

Much of the talk around network virtualization and SDN has concerned the interior of the datacenter, where a leaf/spine switch architecture forms the physical fabric, atop which we can create these virtualized network connections. But what happens when we move this connectivity outside the datacenter? We could knit datacenters from various geographies into one large-scale fabric. Tenants on opposite sides of a country could be connected just as easily as if they were inside the same facility, for instance. This kind of connectivity also creates cloud adjacency because resources can seek out cloud on-ramps located elsewhere on the fabric.

That adds heft to the idea of colocation as an enterprise hub. To go back to the train analogy, it's really the *fleet* of datacenters that acts as the enterprise's Grand Central Station. A branch office connecting into a Paris datacenter could provision a connection to a private cloud hosted in a Singapore facility. Assuming the two datacenters are part of the same networked fabric, this connection could be generated almost immediately.

At 451, we refer to this connectivity as **software-programmable interconnection** (SPI). The term applies whether the connectivity comes from SDN or simpler network virtualization – they serve the same goal of creating (or decommissioning) virtual connections quickly, through software. SPI abstracts away geography so that to the end user, it feels as though a datacenter in another country is a next-door neighbor.

SPI and Automation

Think about how the cloud differs from ordinary hosted infrastructure. Among other things, it's about automation – the ability to call up infrastructure only when needed. Similarly, SPI differs from the interconnection services traditionally offered by datacenter operators, putting connectivity into users' hands via a self-service portal. For the customer, networking becomes nearly frictionless and open to experimentation.

This is a natural product of the SDN discussions that have been taking place for a decade, but the technology has matured enough to offer a new perspective for datacenter operators. Colocation alone is no longer the primary mission; it's about interconnection, especially in the Grand Central Station scenario. Combining datacenters into a unified fabric puts more destinations within arm's reach. Software-programmable interconnection can enable this wider net.

This could be particularly important to edge computing. Consider the trend of mobile banking. We would want customers' devices to connect to processing at an edge computing site in order to keep latency low, but eventually, we will want to backhaul that data to a destination such as a cloud. If the edge computing node is inside an SPI-enabled multi-tenant datacenter, the desired connectivity can be achieved with a virtual cross connect – no bandwidth charges would be incurred, and the connection can be moved to different clouds or other enterprise nodes as necessary.

We expect networking in general to gravitate toward SDN, and in the long term, it would not be surprising to see SPI become nearly ubiquitous among datacenter operators. Until that day, those operators pursuing SPI can wield it as a competitive advantage.

Governing Workload Placement with Software-Programmable Interconnection

As workloads shift to a variety of cloud, datacenter and edge locations, enterprises are now asking new questions to govern their process for workload placement. Traditionally, workloads have been optimized for the classic three-tier architecture to support the *presentation*, *business logic* and *data access* layers from a central location. Today, another set of tiers needs to be considered: whether a workload, or portions of it, should reside in the *cloud* (or a datacenter that we could call the 'near edge'), at the *edge*, or at the *end device*. As we saw in the figure, the workload requirements across industries can vary, but the analysis for workload placement can be distilled down to two foundational questions: Where will enterprises place the workload, and how will they deploy the infrastructure supporting the workload?

Where You Deploy Matters

Distance can hamper or hobble digital communications. As the resources that enterprises need to interconnect become more distributed, the key is identifying the strategic locations that offer the shortest possible distance to the largest number of resources. Identifying these critical exchange points can reduce latency and dramatically drop networking costs. What was once an expensive backhaul between two metros can become a simple cross connect across a multi-tenant datacenter.

Maximizing the value of this setup will likely involve distributing resources across cloud availability zones and across regions. For workloads being placed nearer to the cloud, inside datacenters, the list of candidate locations can be refined by identifying those with the highest density of business, network and application ecosystem partners. Edge locations, on the other hand, are often at a significant distance from these cloud on-ramps. Here, ideal locations would allow for aggregation across the largest number of users along with support for a large variety of network providers to ensure flexibility in the most efficient last-mile options. At the edge, many enterprises are relying on a combination of SD-WAN, MPLS and Metro Ethernet to meet the wide range of requirements.

How you Deploy Matters

As workloads are broken down into their components, the analysis comes down to a simple evaluation of scale vs. performance. The one-size-fits-all solutions have proven to limit business at scale, and now enterprises are analyzing specific workloads to assess whether these need to be placed on optimized infrastructure for performance, or if they can be scaled more effectively in repeatable virtualized blocks on demand.

Hybrid-IT Analysis

Many applications are designed for the presentation and business logic layer to be scaled on virtual resources. Cloud deployments can simplify these layers, but the underlying data layer is different. Typically, these database-driven workloads are designed to support high transaction rates and require very specific resource profiles, and they do not efficiently scale in the cloud. With databases placed in cloud-adjacent locations and data warehousing that can be optimized on hyperconverged infrastructure, the application layer can be scaled as a cloud deployment. This combination can offer lower costs per transaction while still keeping multiple clouds within low-latency reach and offering the agility required at scale. This analysis of strategic locations often offers compounding benefits that include local control of data to support compliance initiatives, efficient license utilization, performance and access to business partners for new workload initiatives.

Edge Computing Analysis

As shown in the figure above, edge workloads are becoming increasingly critical to enterprises across all industries. The trends to support the workforce, offer real-time analytics and transform the customer experience continue to be accelerated by the compounding effects of IoT, AI and 5G. Bringing services closer to users while simultaneously increasing performance to support real-time needs requires a completely new model. Strategic locations should be identified based on their proximity to large populations of users and the ability to colocate the ecosystem partners required across the supply chain. Workload functions should be analyzed based on whether they can be virtualized at scale and what benefits they can gain from residing at different ‘edge’ locations.

Conclusion and Recommendations

As enterprises increase their dependence on clouds, SaaS providers and other outside entities, they can also rethink their networking profiles to regain control of connectivity. A multi-tenant datacenter already hosts many commonly required destinations and can transparently provide connectivity over private links. For hybrid IT and edge computing, this essentially removes the distance to business-critical ecosystems. A cloud on-ramp or a supply chain partner can be just a cross connect away.

The advent of software-programmable interconnection extends reach, giving a colocation site access to many destinations around a fabric of datacenters. This creates economies of aggregation, as the enterprise can enjoy the network effect of being surrounded by connectivity options.

451 Research has shown how digital business has become synonymous with the seamless flow of IT traffic. Data volumes continue to grow exponentially, and customers are demanding real-time experiences across more diverse ecosystems. These compounding forces are straining connectivity and traditional networking architectures. We have seen how distance has become the silent digital business killer and how digital leaders are shifting to a new architecture based on colocation and interconnection.

This architecture, known as Interconnection Oriented Architecture (IOA), allows enterprises to increase data exchange by an order of magnitude, gain strong security and control, and achieve the lowest latency possible, all at a dramatic cost reduction. However, achieving these benefits requires building on the right digital platform with the combination of capabilities, locations and ecosystem access that digital business demands.

In the traditional model, colocation was a tool to house latency-insensitive data and applications. Distance was an afterthought, and the cost of power, space and cooling was a primary consideration. Today this has completely shifted. The priority has now shifted to identifying the strategic locations that can offer the shortest possible distance to the richest ecosystems accessible through interconnection and edge services. Leaders across industries have discovered that for applications that require access to more than one cloud or require exchange with 2 or more business partners, the cost of colocation is insignificant when compared to the value that can be built by being in the right location directly adjacent to the ecosystems they need.

Criteria for selecting a digital platform include:

- Strategic locations adjacent to multiple cloud providers across the cloud availability zones
- Global reach across the metros that compose the digital edge
- Direct access to the ecosystem partners including networks, SaaS providers, technology partners, industry business partners
- Software Defined Networking with global reach and enabling local agility
- On-demand edge services
- API-enabled access

Top question to ask your colocation and interconnection provider

- How many locations does your platform support?
- In the locations we need to support today, how many Cloud and SaaS providers can you provide direct access to?
- What edge services can you offer to help us quickly expand our market presence?
- What availability standards do your multitenant datacenters meet?

Platform Equinix was designed as a global platform for digital business. Platform Equinix offers dynamic, secure and low-latency connectivity to dense ecosystems of cloud and industry partners and providers. Thousands of digital leaders have used Platform Equinix to establish their digital edge in the world's most strategic locations, leveraging ecosystem choice to create a global fabric that aggregates last-mile access to people and things that directly interconnects services from the widest variety of dense ecosystems. With over 200 datacenters globally, industry leading 99.9999% availability and access to over 2400 network providers, 2500 clouds, 750 content providers and over 3500 enterprises to directly peer with, Platform Equinix has become the best platform to build your digital advantage.

Getting Started

Transforming Hybrid IT and Edge computing solutions can appear challenging. For many Enterprises the process starts with optimizing the network to meet the requirements for the new strategy. The process starts by identifying the locations with the largest volumes of data transfers across multiple clouds and business partners. By using colocation and interconnection in these locations, businesses can transform their network by directly peering, resulting in a reduction in latency, increased application performance, and network saving that can be used to unlock the capital for the next phase of transformation.

Contact Equinix to learn more about how the right digital platform can help you transform your digital presence, gain direct access to rich ecosystems, gain new levels of agility through on-demand edge services and help you architect Hybrid IT and Edge for your advantage.

Equinix.com/Contact

CONTENT PROVIDED BY:

EQUINIX

About 451 Research

451 Research is a leading information technology research and advisory company focusing on technology innovation and market disruption. More than 100 analysts and consultants provide essential insight to more than 1,000 client organizations globally through a combination of syndicated research and data, advisory and go-to-market services, and live events. Founded in 2000 and headquartered in New York, 451 Research is a division of The 451 Group.

NEW YORK
Chrysler Building
405 Lexington Avenue,
9th Floor
New York, NY 10174
+1 212 505 3030

SAN FRANCISCO
505 Montgomery Street,
Suite 1052
San Francisco, CA 94111
+1 212 505 3030

LONDON
Paxton House
30, Artillery Lane
London, E1 7LS, UK
+44 (0) 203 929 5700

BOSTON
75-101 Federal Street
Boston, MA 02110
+1 617 598 7200

451 Research®