The Edge Manifesto: Digital Business, Rich Media, Latency Sensitivity and the Use of Distributed Data Centers

31 July 2015 ID:G00290109

Analyst(s): Bob Gill

VIEW SUMMARY

The edge manifesto calls for the placement of content, compute and data center resources on the edge of the network, closer to concentrations of users. This augmentation of the traditional centralized data center model ensures a better user experience demanded by digital business.

Overview

Key Findings

Gartner's Nexus of Forces is driving the convergence and mutual reinforcement of mobile, social, cloud and information in support of consumerization and the democratization of IT, while raising user expectations, emphasizing content-rich data types, and creating increased volume of information and transactions.

Digital business takes these elements and applies them to support new business models stemming from information, and places them as critical pieces of core business offerings, with an emphasis on timely and accurate delivery.

Moving data centers' processing and content delivery/collection closer to the sources and sinks of this information, including cloud onramps and offramps, offers significant benefits and spawns new business models. This is the essence of the "edge manifesto."

The use of smaller, distributed, connected data centers (perhaps space in colocation centers), closer to concentrations of users and generators of content ("pushing things to the edge"), will be required for these workloads. In many cases, these will *augment, not replace*, the more traditional large data center.

Recommendations

Invest in architecture and technologies that help to address the increased criticality and user experience that digital business demands.

Cross-functional collaboration is required, so ensure that edge planning efforts fall under a cross-skilled engineering team or an architect who oversees compute, cloud and networking as well as business applications; "plan in silos, fail as an organization."

Ensure the architecture addresses the three "location" challenges, as outlined in this research.

Leverage best practices from vertical industries that use the Internet to deliver video or rich media.

TABLE OF CONTENTS

CONTENTS

FIGURES

Figure 1. The Digital Business Development Path

Analysis

We've begun the move to digital business, including rich content via mobile devices, where people, their devices and even unattended "things" become actors in transactions. To optimize the experience, Gartner believes the topology of networked data centers will push over the next five years from a centralized, mega data center approach, to one augmented by multiple, smaller, distributed sources and sinks of content and information, whether located in distributed, enterprise-owned data centers, hosting providers, colocation or the cloud (see Note 1, The Nexus of Forces and Digital Business).

The goal of the edge approach is keeping the heaviest of the new traffic *and processing* at the *edges* of the Internet, closest to the user applications and devices that are the sources and sinks of this traffic. Content delivery networks (CDNs) evolved to allow the timely *distribution* of at first static, and later, dynamic and rich content (including video) to millions of users by pushing the content to caches closest to those users (see "Technology Overview for Content Delivery Network Services"). Similarly, the edge manifesto recommends moving data, applications *and* their data centers (perhaps in the form of colocation centers) to the edge. The key difference between CDN and the "edge approach" is the degree to which each approach offers compute and management of bidirectional traffic.

How Can We Guarantee Acceptable Performance?

Several factors, when combined, drive the need for a shift in topology:

1. **Volume of content:** First, the sheer volume of content continues to grow as more users worldwide become connected and interact. Skeptics can argue that growth in the number of PCs is dwindling as tablets and mobile devices become more capable, but the number of devices and connected humans who operate them continues to grow.
2. **Nature of content:** Second, the nature of content, or media type, continues to trend toward richer media and the ubiquity of video, both streamed outward to and captured from this increased number of interactive devices. More devices handling richer content implies greater volume of traffic.
3. **Latency sensitivity:** Latency sensitivity refers to one of the effects of consumerization: User expectations are high and continue to rise. One side effect of digital business is that businesses interact with their customers digitally and directly; if an experience is poor, it reflects poorly on the business and its brand. Latency sensitivity drives the goal to reduce latency, complicated by the increased volume of content and content types (video with audio) in which latency is immediately obvious. In addition, latency can be particularly difficult as it is constrained by the laws of physics (speed of light) and cannot always be overcome by software.
4. **Data center location:** A fourth complication involves the trend toward placing new large data centers or substantial colocation deployments where energy costs are low and space is inexpensive. In the U.S., this might equate to placing centers in the Pacific Northwest or Mountain States, while in Europe, it manifests in data center buildings in Scandinavia and Iceland. While the reasons for this placement are many, they are not ideal for delivering latency-sensitive content. Despite the common image of huge "cloud farms" in remote locations, many cloud providers already augment these mega data centers with edge presence in colocation centers near concentrations of users. They have been exploiting the edge concept for years.

What Can We Do?

> Deploy assets where they encounter the fewest bottlenecks between content and consumer.
>
> "Push it to the edge."

How We've Solved Content Delivery Challenges Before

Content delivery networks were developed to solve network latency and bandwidth constraint problems for the delivery of content, usually from one origin to many thousands or millions of users. The model was to position many hundreds or thousands of servers in locations around the globe.

Despite the potential of using these servers for compute, a CDN today is fundamentally used as a delivery platform. But rather than impose a computing load on a delivery platform, the edge manifesto asks, "What if we borrowed from the CDN model, and distributed *data center capacity* in the form of micro data centers and colocation sites, rather than just delivery caches, at the edges?"

In the past few years, enterprises and data center providers have taken the CDN concept and coupled it with brick-and-mortar data centers to gain the benefits of pushing content to the edge. In some cases, the ideal scenario is a hybrid of both. For example, the meshed model proposed in "Use Colocation Networking to Provide New Connectivity Paradigms and Drive Business Transformation" offers the potential for general-purpose compute, as well as public cloud onramp and offramp services, to more efficiently connect enterprises at the edge.

The Edge Manifesto

In essence, the edge manifesto borrows from the real estate slogan of "location, location, location." The concept is to avoid issues of latency and congestion by placing compute resources and content closer to concentrations of users and sources of data. Rather than simply a technology, it's also a concept of topology.

The Edge Manifesto

Topology and Technology

- Location
  - Geography, topology
- Location
  - Gravity, interconnection
- Location
  - Access, last mile
- Master plan
  - Elevated awareness in the organization
  - Reference architectures
  - Implementation services

Location Consideration No. 1

The first notion of geography centers on location from a broad nationwide/worldwide perspective:

- Where are the concentrations of users?
- How can they be identified, stratified, ranked and prioritized?
- Can we differentiate among these users in terms of application types, latency sensitivity and volume?

We can use these answers to create models ranging from one centralized location to many interconnected locations, along with the required networking, to determine the optimal mix. For example, one Gartner client found that by eliminating a hub-and-spoke model with dozens of point-to-point Multiprotocol Label Switching (MPLS) circuits, and instead deploying a meshed network of smaller data center footprints in colocation centers interconnected by private lines, they could reduce both networking costs and latency by significant amounts. From a high-level perspective, this first "location" attribute addresses the question of topology. We are not questioning the need of the centralized data center, with its proven efficiency and economies of scale. We are augmenting the centralized model with edge presence. While some operational costs may rise serving the edge, others will fall, and client satisfaction is improved.

Location Consideration No. 2

The second notion of location describes the elements of gravity and interconnection. "Gravity" refers to the presence and concentration of desirable connectivity partners, infrastructure-as-a-service providers, software-as-a-service providers, and ecosystem or peering partners. For example, an enterprise may find that spreading out its onramp and offramp's infrastructure and service provider to various colocation facilities around the country may balance traffic and provide better performance. Interconnection is just the technical aspect of how the involved parties connect, ideally directly over fiber, for reasons of security, low latency and Internet-free communications at low cost.

Location Consideration No. 3

The third and final consideration of location concerns is often called "the last mile," or more accurately, the actual connection between end users or data centers or sensors — in other words, the sources and sinks of information. More geared toward colocation-based edge scenarios, locations that have a broad selection of local fiber providers or perhaps high-speed tethers to carrier-dense "carrier hotels" can extend the high-speed interconnection benefit to the edge and beyond.

Master Plan

Finally, all of these concerns must be considered in the context of a formally structured master plan. There should be criteria, thresholds and pricing considerations that consistently guide location choices.

A master plan starts with an elevated awareness of the edge concept in the planning organization and architect level, above the individual silos of networking, applications, data center or even individual cloud projects. It is critical to look at these topologies with an overall goal in mind, not optimizing specific subsystems (and perhaps missing out on an optimized overall solution).

To ensure consistency and replicable results, establish reference architectures to allow known, successful configurations to be adopted and implemented across the organization. These can include recommendations on the data center footprint, cloud connectivity, data center interconnection and even local access types of configurations.

Finally, a standardized approach to providing the actual implementation, such as via a specialized team or provider, can greatly streamline deployment across a number of locations.

Bottom Line

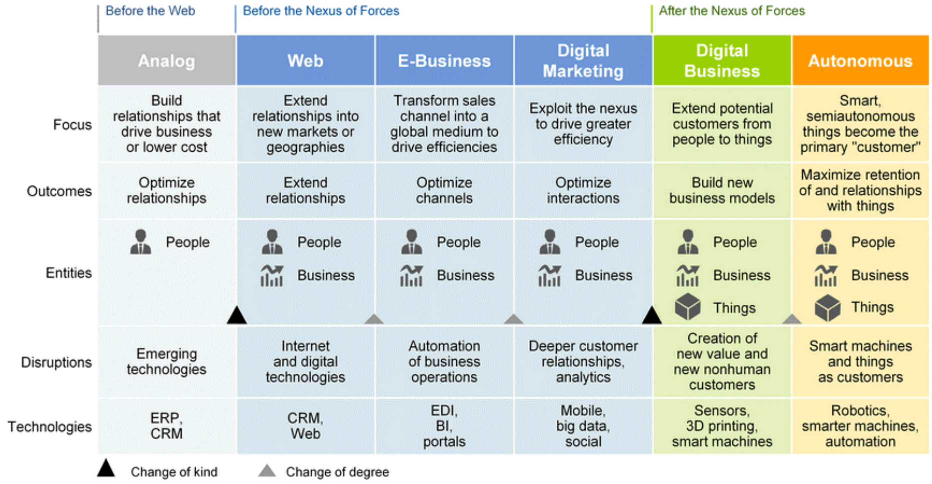
The notion of edge computing is not novel, but conducting business directly and digitally involving rich media calls for spreading the deployment model out, not consolidating into one geographically centralized source. Cloud providers, content-offering cable TV providers and large industrial companies such as GE are proving the concept valuable. Whether for reasons of data sovereignty, avoiding latency, business continuity/resilience or simply lowering costs, the edge computing model will continue to grow in practice.

Appendix

The Nexus of Forces and Digital Business

The Nexus of Forces accelerated the move from e-business to digital business as consumers shifted to mobile devices and integrated social interactions, creating information-driven relationships with global reach. The transition from digital marketing to digital business occurs as *things* become *actors* in transactions (see Figure 1). With the emergence of the Internet of Things, with potentially billions of endpoints all creating information to be used by digital business, and perhaps using edge-based analytics to take action, the increase in traffic, storage and computing rises still more. In the emerging world of digital business, relationships are by their nature bidirectional, as the content and data types continue to get richer.

Figure 1. The Digital Business Development Path



	Before the Web	Before the Nexus of Forces			After the Nexus of Forces	
	Analog	Web	E-Business	Digital Marketing	Digital Business	Autonomous
Focus	Build relationships that drive business or lower cost	Extend relationships into new markets or geographies	Transform sales channel into a global medium to drive efficiencies	Exploit the nexus to drive greater efficiency	Extend potential customers from people to things	Smart, semiautonomous things become the primary "customer"
Outcomes	Optimize relationships	Extend relationships	Optimize channels	Optimize interactions	Build new business models	Maximize retention of and relationships with things
Entities	People	People Business	People Business	People Business	People Business Things	People Business Things
Disruptions	Emerging technologies	Internet and digital technologies	Automation of business operations	Deeper customer relationships, analytics	Creation of new value and new nonhuman customers	Smart machines and things as customers
Technologies	ERP, CRM	CRM, Web	EDI, BI, portals	Mobile, big data, social	Sensors, 3D printing, smart machines	Robotics, smarter machines, automation

▲ Change of kind △ Change of degree

Enlarge

Electronic data interchange (EDI); business intelligence (BI)

Source: Gartner (July 2015)

About Gartner | Careers | Newsroom | Policies | Site Index | IT Glossary | Contact Gartner