DCD > White Paper

The Calm before the Storm?: Edge Computing in the United States

May 2017

BLACK & VEATCH

www.DCD.company

Contents

Main Findings & Analysis

Figures

DCD Comment

BLACK & VEATCH

DCD Comment

The current status of edge computing may be best described as the calm before the storm. Qualitative comment and research data indicate interest rather than urgency among most enterprise markets, with larger inter-connected colocation players and telecom providers laying the foundation for edge computing networks in preparation for increasing demand.

That demand will increase, and increase fast is difficult to argue. As with many mega-trends, the emergence of Edge is based on a combination of factors. Technological advances in processing, analytics and networking, in conjunction with the emergence of IoT, Big Data, mobile technologies and AI/ machine learning all enhance customer experiences and interaction from which revenue can be derived. Yet there are challenges to realizing that revenue, as many more data sources will congest already stretched data infrastructure and the dominance of the cloud for data storage. This means that analytics and processing need to be performed as close to the source of the data as possible.

Edge has been proclaimed by some as the end of Cloud. Yet the IT evolution is actually a process of previous technologies shifting and adapting to the newcomer rather than an outright switch. Colocation has evolved into a means of housing and accessing Cloud and enterprise facilities continue to exist due partly from the recognition that Cloud is not the best option for everything. Thus Edge will be the enabling trend for the next phase in the evolution of Cloud and of data centers rather than the end of Cloud.

Instead, it seems most likely and logical that organizations will continue to use existing networks and infrastructure while analysing core-to-edge value, operations and protocol. Ultimately, these current networks, even in combination with existing networks of content delivery networks [CDNs] will not be sufficient to support the edge computing requirement in the United States. This occurrence is projected to be a 'tipping point' in the process to Edge.

If these types of early adoption issues seem familiar, it's because the early adoption phase of Cloud was characterized by doubts regarding security, services levels, capability to deliver and compliance. These doubts re-emerge whenever a major cloud provider suffers an outage or each time enterprise or Government are forced to reveal that they don't know exactly where their (or their clients') data is located.

Greater understanding of the utility principle behind computing that is shared, available on demand and scalable has opened the mainstream to Cloud. Indeed, today's generation can hardly imagine the need to own or operate a server. What is the principle that will unlock Edge and open it to more organizations and to more people?

The heritage of Edge comes from a variety of places ranging from Content Delivery and Peer2Peer networks, from the analytic and processing capabilities of data centers at the Core, from the storage capacity and scalable utilization of Cloud and from the reach of telecom, power, cable and the Internet. With this varied heritage, the Edge is bound to be a bit complicated. The way to view Edge is as Smart Infrastructure that combines embedded sources of data collection connected to a network backbone, and which can make decisions locally and immediately on the basis of data analysis. Much infrastructure is 'almost smart' since it is automated and heavily fitted with sensors and controllers. However it is 'almost smart' because it lacks the capability of, for example, being able to slow cars traveling within range of traffic congestion.

Smart infrastructure should be viewed as total interconnectivity between sources, integration into a fully systematized network with seamless learnings running to and from the core. It should also facilitate the delivery of a better (safer, more organized, more relaxing and a more sustainable) customer experience. Just as Cloud took time to develop the hyperscale facilities with which its larger providers are associated, it will also take time for Edge to optimize delivery across the widely varying communities in the United States.

There is an emerging consensus that the design of edge computing units will be modular. This makes sense in terms of the physical housing unit given that Edge IT needs to operate from common software definitions and protocols. Otherwise what will emerge will be a variety of disparate systems which will compromise the aggregated value of edge computing.

In terms of units, what is most likely to emerge will be a family of edge processing units modular in design based on common standards that like servers, can be used alone or in racks. Some of the space for accommodating this infrastructure is already in place on sites operated by telecom or utility providers. The emergence of 5G and/or LTE will enable mobile data processing and transmission further empowering Edge. Throughout this discussion, when we refer to edge processing units or housing – this is what we are referring to – a modular family that can analyse and process different weights of data in different locations and of widely varying sizes as further described below.

Analysis indicates six different Edge usage topologies:

- **Mobile edge computing devices** in cars, planes, trains, trucks.
- **Home/ building computing unit –** maybe the form of a micro-modular data center the size of a utility cabinet.
- **A combination of home or building units –** may form a collaborative node on the way through to the core data center, reducing the need for larger edge computing devices.
- **Shared edge computing units –** shared by different businesses with a shared goal (for example, healthcare or education).
- **Distributive edge computing units –** these primarily will deal with the distribution of videos and other 'rich' and latency-sensitive content.
- **Linear/process edge computing units –** these will bond different companies or different parts of an organization to meet a common production goal from the beginning point ('requirement') to the end point ('delivery'/'re-engagement').

While there is increasing momentum behind edge computing, there are still questions about how Edge will become a major defining trend of the next few years. The importance to an organization (priority) assigned to core-to-edge bears some correlation to the organization's current dependence on outsourcing to meet its data hosting and/or processing requirements. This indicates that the first phase of Edge will be a service offered by third party providers whose investment will be based on demand from customer sectors.

Edge will not be the same at every place across the United States. It is inaccurate to speak of a United States core-to-edge network. Just as the fiber network does not reach every household in the United States, the development of the core-to-network edge will not either. Supply in downtown Los Angeles, CA will vary from that in Tuscon, AZ which will vary still further from that in Bethany, MO. This gradation of locations throughout the United States means that different solutions in terms of networking, processing and the costs of service will need to be established.

Executive Summary

BLACK & VEATCH

Executive Summary

'Edge computing' like most IT trends is not a revolution but more an evolution. Edge has its roots and heritage in the content delivery and Peer-to-Peer trends of the early millennium as well as in grid computing. Yet a combination of improved technological capabilities in networking, computing and analytics coupled with a demand based on huge projected increases in network data from 1 Zettabyte in 2016 to 180 ZB in 2025, means increasingly that where compute takes place matters.

This has led to the deployment of computing capabilities closer to the edge of the network and out of the hyperscale computing, memory and storage 'core' towards the places where data is first generated.

Edge computing is therefore a form of distributed IT architecture bearing a similar evolutionary relationship to grid or mesh computing much as Cloud does to utility computing. The evolutionary principle of both is based on the disruption to the core relationship between the producer and the consumer of data. This is similar to what both grid/mesh and utility computing initially explored and which Cloud has subsequently taken main stream - and which Edge will do in the near future. The user disruption that edge computing enables is in allowing local users to produce and perform analytics with data rather than just consuming it.

The Edge is not a fixed place. Rather, it exists as one part of a pair – Edge exists in conjunction with Core. Edge without Core negates the principle of its function. It is possible that as network speeds and capabilities evolve through the development and implementation of silicon phonetics, Solid-state drives [SSD] and parallel technologies, so the lines between Edge and Core will increasingly blur.

Edge computing is a major impact trend on the IT and data center industries and it will move forward in conjunction with other such impact trends – AI/machine learning, analytics, Software Defined Networks, advances in computing language and increased network performance and capability. The level of interest in Edge among delegates to the DCD Enterprise conference in March 2017 has increased by 50% over 2016.

The emergence of edge computing is based on a number of drivers.

- The growth in the generation of data, particularly in the form of the Internet of Things (i.e. data generated by machines and devices, rather than people). It is predicted that by 2020 there will be 50 Zettabytes of data generated by more than 25 billion connected devices.

- The increases in both the amount of traffic that is video and the proportion that is Cloud traffic (92% by 2020 as estimated by Cisco).

- As the amount of data running through networks increase, bottlenecks and interruptions will increase proportionately. If all data generated in 2020 were to be sent from source to core production facilities, the current system could not keep pace. Even if only the critical or most important data were to be sent, that data volume would still challenge the capabilities of the current network in the United States.

- Edge computing has gained traction due to the limitations of Cloud-based analytics systems and therefore Edge can act as a bridge into core Cloud production/storage facilities. This is most useful in situations where Cloud cannot offer the low latency required by some devices and users. Cloud is also not good at transmitting video which is one of the key content components of peer-to-peer sharing. Cloud is also less capable of offering the real time experience expected of augmented or virtual reality.

- Therefore in a situation where networks are likely to come under increasing pressure, efficiency in data transmission will become increasingly important. This is where edge computing and the process of data curation, selecting that which is important from that which is not priority, and organization come in. Estimates of the proportion of data that may be important range from 10% to as little as 0.0001%. ▶

Executive Summary

▶ Edge computing will in time be considered for any application where there is the need for immediate data processing based on one or more of the drivers described above. Currently the most common scenarios being considered include:

Automated vehicles

The self-drive car is perhaps the most public example of edge computing since one car will require an estimated 200+ CPUs so, in the words of Peter Levine of Andreessen Horowitz it is effectively 'a data center on wheels'.

Other Transportation related

Edge computing does and can perform a number of functions for commercial and public transportation. For highly complex vehicles such as spacecraft, airplanes and ships, the processing focuses on the most valuable information and allows only that to be transmitted for analysis. Local processing allows (as with the driverless car) information to be fed to the transportation system controls whether human or automated.

Media/content

Edge computing here represents the upgrade of content delivery networks. Edge computing will play a part in future content delivery and enable content providers to expand their geographical reach and to maximize the efficiency of their delivery networks as they introduce increasing numbers of value-added and interactive services. Online gaming is part of this category.

The Smart Home

The Smart Home would rely on extensive information collected around the home from sensors and controllers which can then be analyzed and acted upon within the home.

Smart Cities

The use of edge computing can be applied to the smart community or city. As the number of sensors and sources grow (individuals, traffic systems, healthcare, utilities, security and policing etc.) so the principle of storing and analyzing it all centrally becomes less feasible. A system based on the principle of edge computing allows the information of importance in meeting requests to be passed back through the network. Edge computing also reduces latency improving response times in situations where action is required quickly. It also, by the nature of edge computing, allows for geographic precision – information relevant to a particular intersection, suburb, or utility can be shared instantaneously.

Factories/production facilities

The use of robotics, AI and machine learning within some industrial and manufacturing organizations has taken the adoption of edge computing further than in most others. The objective is to weave production seamlessly into a lean and wholistic process from the predicting of demand, through to production based upon that demand to delivery and post-sale service. This means collaboration between data sources across a range of locations and of companies.

Issues Requiring Resolution and Collaboration

While the requirement for edge computing and its role seems clear, there are a number of issues which need resolution:

- How data and therefore privacy will be protected. As part of this, the ownership of data and rights to data will need to be defined.
- The growth of the generation and transmission of data by machines referred to as 'IoT' and its enablement by edge computing will also complicate the concept of criminal or civil responsibility. If an accident happens for which a self-drive car is liable (due, for example, to meeting a situation for which it had not been programed or to system or network failure) where does the responsibility lie?
- The growth of edge computing will be based in part on the proliferation of devices that are included into the network universe as sources or as producers. A number of analysts question the ease with which large numbers of further devices can be easily incorporated.
- Resilience needs to be considered. What is the impact of the failure of a source or of a cell within the edge processing unit? Does it bring down the whole unit? How does it impact collaborative units? Shared programs? How can a system entirely dependent upon a totally fluid network continually protect itself against cyber- attack?
- How will priority be assigned to traffic on the network?
- As edge computing grows, so the shared language which enables multiple-edge systems and a variety of platforms and runtimes will need to evolve.
- Protocols that enable the systems and the network to work together are required and need to be developed. Dell and Intel are in the process of developing IoT gateways and routers that can support edge computing while software such as Apache Spark is also being developed and scaled.
- A naming system (like the Internet IP system) that can act as a means of identifying and communicating with devices and users on the Edge needs to be developed. Such as system needs to cope with the huge number of devices there will be, able to deal also with mobility and be secure.
- The means of enabling the huge network that a core/edge computing system will require is Software Defined Networking. This is the most established and mature of the Software Defined utilities and is already established in many core data centers, deployed as the most effective means of accessing Cloud systems. ●

Detailed Findings
& Analysis

BLACK & VEATCH

Detailed Findings & Analysis

Chapter One: The United States On The Edge?

The huge projected increases in network data from 1 Zettabyte in 2016 to 180 ZB in 2025, means increasingly that where compute takes place matters. This has led to the deployment of computing capabilities closer to the edge of the network and out of the hyperscale computing, memory and storage 'core' towards the places where data is first generated, the places McKinsey and Co. term the "digital edge". This process enables businesses to extend the advantages of business agility towards the Edge in terms of mining business-critical IoT, streaming-content data or through enhancing customer experience at the Edge.

What does this hastening trend mean for the planners, designers and builders of edge data center capacity? More of the same, or will the expertise and competencies required for developing edge facilities be different?

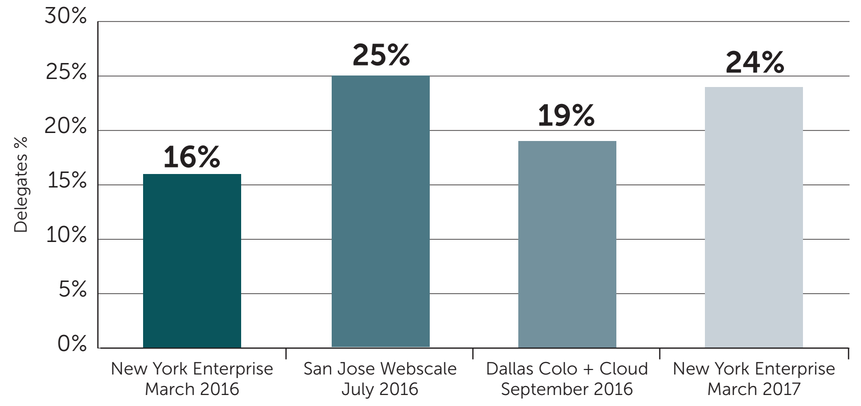
Since technology adoption trends globally are initiated in and have a significant impact on United States markets before any other, the course of edge computing will be set here first.

So, what currently are the status and trends in edge computing among the people responsible for IT and digital infrastructure across the United States?

To answer this in part, analysis has been performed across 2,515 responses from delegates attending the three DCD conferences across the United States in 2016, and the first 900 from recruitment to the Enterprise event to be held in New York in March 2017. These events provide a balanced representation of the upper end of the digital infrastructure industry across the United States.

Across the country, the proportion of delegates assigning priority to the core-to-edge track varies by the sample at each event. The sample from the enterprise focused New York event in March 2017 indicates a 50% increase on the level of priority from 12 months earlier. In the 2017 interim findings, 12% of the New York 2017 sample indicates that they are constructing or planning to construct an edge data center. Across all events the investment activity with the highest priority is 'Design + Build', with ratings of 35%, 56%, 40% and 67% reading across the events from left to right. ▶

Figure 1: Priority given to Core-to-Network Edge Content Tracks (% registered delegates DCD USA events 2016 & 2017)

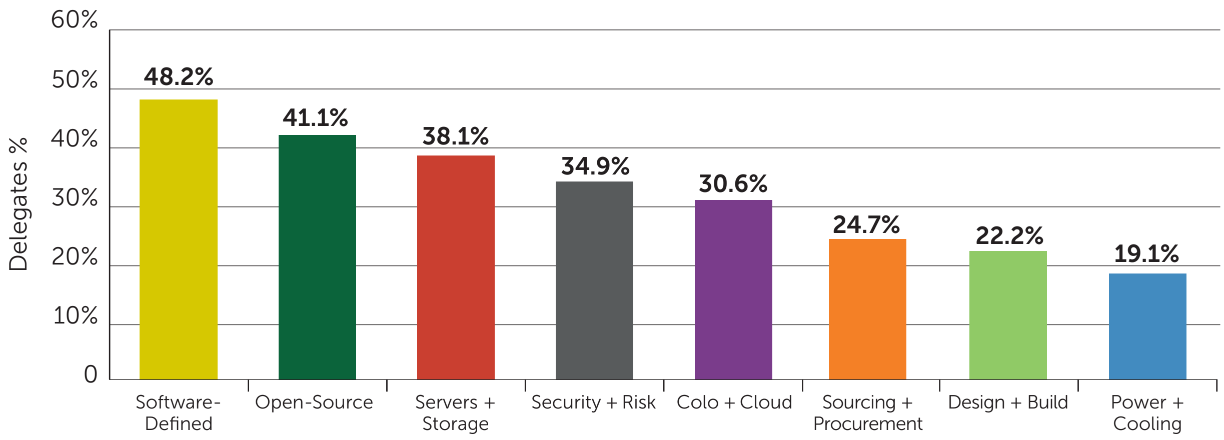


Source: DCD Group 2017

Detailed Findings & Analysis

▶ Core-to-edge is one of a number of investment priorities that the sample was able to nominate. Those giving priority to software-defined infrastructure, open-source software/hardware and to servers/storage are more like to combine these with a priority to Core-to-edge also. This positions those pursuing edge computing towards the adoption of upcoming IT and network systems and towards IT rather than facilities. This suggests that edge computing is positioned closer to IT investment streams rather than the facilities or outsourcing streams. Placing edge computing into a context where it builds on all the components that have created the present state of digital infrastructure is one of the key purposes of this White Paper (**Figure 2**). ●

Figure 2: Stronger Correlation between Core-to-Network Edge Content Track and IT/Systems than with Facility (% registered delegates DCD USA events 2016 & 2017)



Source: DCD Group 2017

Detailed Findings & Analysis

Chapter Two: What is 'Edge Computing'?

The familiarization of technological terms into IT markets is intended to simplify complexity, to make adoption less threatening, to reach beyond the adopters to their customers (think 'Intel Inside') and to reinforce hierarchies of understanding. 'Cloud', 'software-defined', 'virtualization', 'converged', 'hyper-converged', 'green' are all broad terms which take on greater meaning in direct correlation with growing market understanding.

The broad term 'edge computing' seems to date from around the millennium when 'edge servers' was a term used to refer to servers used by Content Delivery Networks [CDNs]. It has more recently been coopted to describe the practice of processing, analyzing and applying knowledge from data produced by sources at the edge of the network ('edge analytics'), in contrast to the traditional practice of transmitting the data to a 'core' processing unit. Note that the term is based on proximity of processing to source and it therefore covers a range of possible use situations in terms of numbers of sources (people, businesses, sensors etc.), volumes of data, processing requirements and specification, the extent of data reduction and even variations in distance or dispersion of sources.

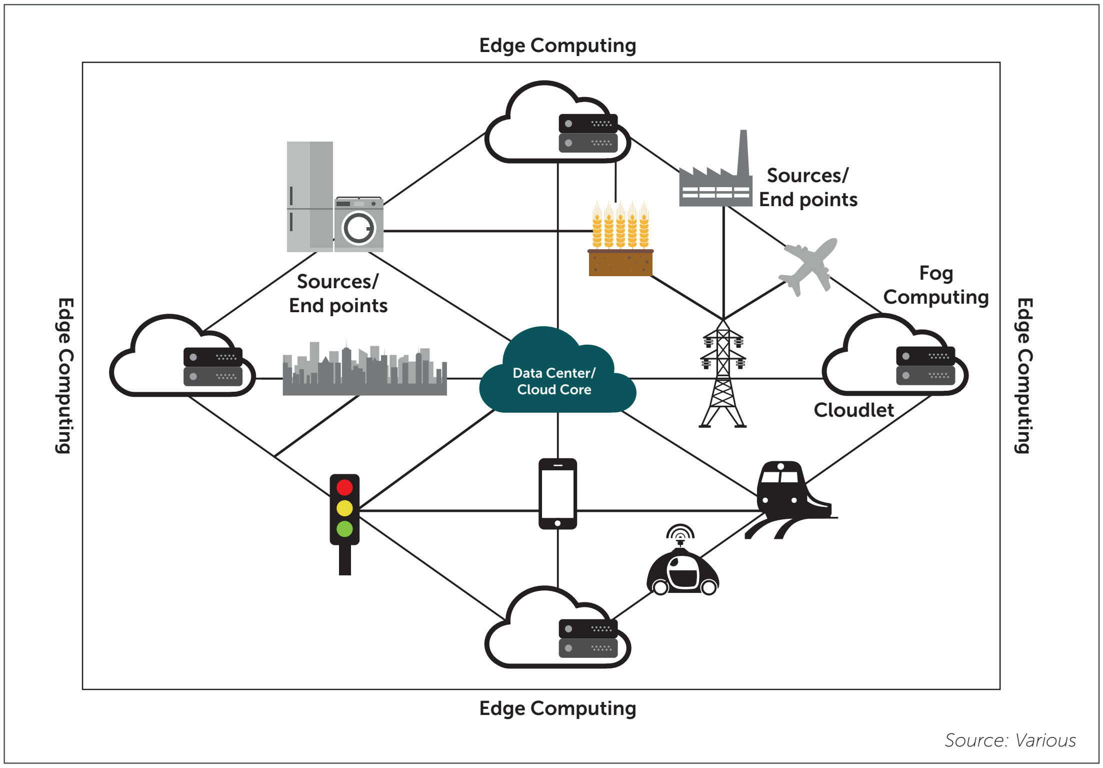
'Edge data centers' are processing units where edge computing takes place although a data center is only one option for such processing. 'The edge' is not necessarily a fixed place since, for example, the principle of mobile devices and environments is important to certain use cases.

Edge computing is a form of distributed IT architecture bearing a similar evolutionary relationship to grid or mesh computing as Cloud does to utility computing. The evolutionary principle of both is based on the disruption to the core relationship between the producer and the consumer of data that both grid/mesh and utility computing initially explored and which Cloud has subsequently taken main stream, and edge may do in the near future.

There are a number of elements of definition that are important for understanding edge computing.

The first is that 'edge' is one part of a pair – 'edge' exists in conjunction with 'core'. Of the 12% of the delegate sample to Enterprise>New York 2017 who indicate they are investing or will invest in network edge units, over 90% will also invest in a new core production facility or upgrade an existing one. As with most network topologies, there may be data staging points between the edge and the core. Edge without core negates the principle of its function, just as the function of airplanes changes without airports, and the value of metals and minerals in the ground does without refineries, smelters or processing plants. ▶

Figure 3: Key Components of Edge Computing



Source: Various

Detailed Findings & Analysis

▶ It is the scalability in response to demand and in service distribution demonstrated by the hyperscale Cloud provision data centers that makes a disaggregated computing model possible at the level of data traffic that exists today and that will accelerate in the near future.

So, the two co-exist in mutual systemic dependence. It is possible that as network speeds and capabilities evolve through the development of silicon phonetics, SSD and parallel technologies, so the lines between edge and core will increasingly blur.

Grid computing is also used to indicate a group of dispersed computers that may be coopted to solve a problem or meet a goal. The network of over one hundred data centers across the world that were made available as needed to the CERN Hadron Collider as it wrestled with the staggering amounts of data generated by its experiments might be described as grid computing. While there are a number of possible delivery models for edge computing just as there are for Cloud, the principle is that the edge is the first point at which information collected from external sources meets the IT network. As such the principle needs to be one of inclusion so more sources can be added easily as required. If, for example, edge computing is used to improve remote security on a building through analyzing sensory inputs and responding immediately on the basis of their threat levels and an extension is added to that building then the whole remote security function is compromised if the data from that new extension cannot be analyzed and acted on at the Edge where the threat might be happening. When new engines are fitted to an airlines, the data generated needs to be the same (or updated) as that from the previous engine.

The grid principle is that the compute process stops at the included computers and since the input into grid computing comes through the IT system rather than from outside that system, the network is inward looking. In the case of Hadron, it is dedicated to processing and analyzing the vast amounts of data coming through.

In edge computing, the weight of processing and of investment is carried by the multitude of edge processing and data science units, in grid computing the weight of both is at the core.

For similar reasons, a small data center located in a small urban or suburban area is not an edge computing facility. It may be of a similar size and possibly of similar specifications to an edge data center but what is missing is the interactivity with data produced, processed and acted on locally. The edge facility enables local users to produce data not just consume it.

Edge computing has the status of a major upcoming impact trend on the IT and data center industries and it will move forward in conjunction with other such impact trends – AI/machine learning, analytics, software-definition, advances in computing language, increased network performance and capability. Yet, like previous disruptions, so long as its basis is sound technology to deliver user benefit, its emergence can be predicted even if the scope of uptake cannot. In this case it is part of the move of business enabled by IT and machine learning towards data-centricity, the practice of using data to make fast and personalized decisions to benefit customers and increase competitive standing, and the exemplar of scalability created by the cloud SaaS delivery model.

The emergence of edge computing can be added to a number of technological progressions based on the symbiosis between IT and telecommunications. There are parallels with the development of the mobile phone network and its evolution to support smart phones (and its further evolution into 5G). Within the data center, the increasing use of the LAN network to carry traffic east-west across the data center (between, for example, servers and storage equipment) rather than simply to and from users outside the data center has led to the development of more fluid topologies that increase the efficiency and reduce congestion through the data center. In an increasing number of facilities, the networks that enable this are software-defined.

The recent slowing of sector energy consumption in the United States based on the replacement of enterprise servers with Cloud (according to the 2016 Ernest Orlando Lawrence Berkley National Laboratory report) may have taken some steam from the debate about the energy sustainability of national IT dependence. However, the ever-increasing importance of the network to the delivery of Cloud indicates that the energy consumption focus of the industry will need to move outside the core data center, and that the addition of major new volumes of data traffic may severely test the capabilities, the latency and the reliability of existing networking resources.

> *"Cloud is taking over – they are creating a world in their own image- that is, as a giant data center."*

In a recent series of interviews conducted across the world on the planning needs of the colocation sector, one respondent wearily observed that *"Cloud is taking over – they are creating a world in their own image - that is, as a giant data center."* That may be an exaggeration as some observers view the data requirements of edge as a balance to the dominance of Cloud, yet the principle of extending IT outwards far beyond the walls of the data center and bringing in the (most important) data from there is a step in that direction.

So, what has led to the very high state of anticipation around edge computing? ●

Detailed Findings & Analysis

Chapter Three: Drivers to the Edge and Use Cases

The emergence of edge computing is based on a number of drivers:

- Changing consumer and business expectations and use of data
- The development of emerging technologies, particularly in the networking, processing, software and protocol areas that make edge computing possible
- Technical reasons where edge processing is either essential or desirable. These include the opportunity to learn from IoT and other forms of Big Data, greater efficiency in processing and transmission across the network, reduction of latency, the delivery of a better customer experience, and some advantages when it comes to data security.

The combination of consumption and technological factors creates in practice a more complex set of drivers which vary considerably between different categories of industry sectors and use cases.

Big Data and IoT = Opportunity

The growth in the generation of data, particularly in the form of the Internet of Things (i.e. data generated by machines and devices, rather than people) is the major driver of edge computing.

The evolution of Big Data is a headline trend. As at 2015, 2.5bn gigabytes of data were being created daily, enough to fill 10 million blu-ray discs which when stacked would measure the height of 4 Eiffel Towers.

The explosion of data in terms of volume, velocity and variety is known as Big Data.

- Volume refers to the amount of data generated. A decade ago, data storage for analytics was counted in terabytes. Now, organizations require at least petabytes of storage, and exabytes and then zettabytes are not far away.
- Velocity refers to both the throughput of data and its latency. The former represents the amount of data in movement (measured in terms of gigabytes or terabytes per second). The latter relates to the delay between the data ingestion and the data analysis (measured in terms of milliseconds).
- Variety refers to both the number of data sources and the heterogeneity of data (structured, semi structured or unstructured).

90% of the data ever created in the world as at 2015 was created in the 2 years preceding 2015 and according to the Cisco Cloud Index the amount of IP traffic per month will grow at a CAGR greater than 100% between 2014 and 2019. By that year, the number of Internet users will grow at 7% CAGR while the number of connected devices will grow faster (at 11.4% CAGR). The proportion of traffic that will be video (an important consideration for edge computing) will increase to 80% (**Figure 4**).

There were an estimated 3.4 billion people on the Net as of 2016. YouTube users downloaded 400 hours of new video every day, Instagram users liked 2.5 million posts every minute, and Facebook users shared 3 million posts per minute and liked more than 4 million posts per minute. Additionally around 4 million Google searches are conducted every minute of every day. Additionally 204 million emails are sent every minute, over 400,000 Apple apps are downloaded and 277,000 tweets are sent. To demonstrate the commercial value, Amazon sells an estimated $80,000 per minute.

A new set of terms has evolved to define this flood of data. An exabyte is a unit of information equal to one billion bytes. A zettabyte is one sextillion. And beyond these, yottabytes, xenottabytes, shilentnobytes and domegemenrottebytes lie in wait. ▶

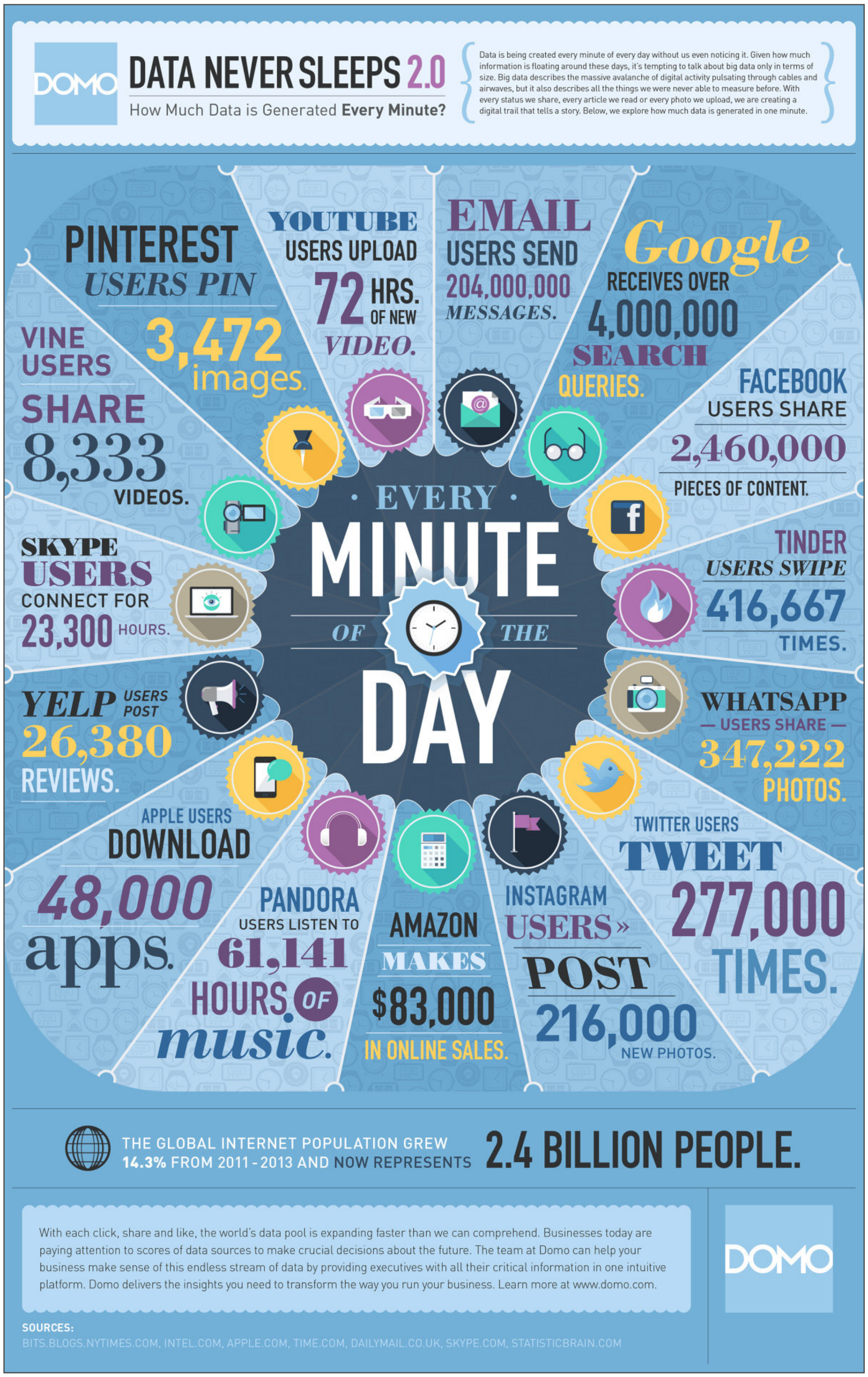
Figure 4: Growth in IP Traffic and Sources of Data

Traffic	2014	2019	CAGR
IP Traffic per month	287 Exabytes	10.42 Zettabytes	105.1%
Internet users	2.8 bn	3.9 bn	6.9%
Number of connected devices	14 bn	24 bn	11.4%
Average broadband speed	20.3 MBps	42.5 MBps	15.9%
% traffic that is video	67%	80%	112.6%

Source: Cisco Cloud Index 2014-2019

Detailed Findings & Analysis

Figure 5: 'Data Never Sleeps'


DOMO
DATA NEVER SLEEPS 2.0
How Much Data is Generated Every Minute?
PINTEREST USERS PIN 3,472 images.
YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.
EMAIL USERS SEND 204,000,000 MESSAGES.
Google RECEIVES OVER 4,000,000 SEARCH QUERIES.
VINE USERS SHARE 8,333 VIDEOS.
FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.
SKYPE USERS CONNECT FOR 23,300 HOURS.
TINDER USERS SWIPE 416,667 TIMES.
• EVERY • MINUTE OF THE DAY
YELP USERS POST 26,380 REVIEWS.
WHATSAPP — USERS SHARE — 347,222 PHOTOS.
APPLE USERS DOWNLOAD 48,000 apps.
TWITTER USERS TWEET 277,000 TIMES.
PANDORA USERS LISTEN TO 61,141 HOURS OF music.
AMAZON MAKES $83,000 IN ONLINE SALES.
INSTAGRAM USERS » POST 216,000 NEW PHOTOS.
THE GLOBAL INTERNET POPULATION GREW 14.3% FROM 2011 - 2013 AND NOW REPRESENTS 2.4 BILLION PEOPLE.
DOMO

Data is being created every minute of every day without us even noticing it. Given how much information is floating around these days, it's tempting to talk about big data only in terms of size. Big data describes the massive avalanche of digital activity pulsating through cables and airwaves, but it also describes all the things we were never able to measure before. With every status we share, every article we read or every photo we upload, we are creating a digital trail that tells a story. Below, we explore how much data is generated in one minute.

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. Learn more at www.domo.com.

SOURCES:
BITS.BLOGS.NYTIMES.COM, INTEL.COM, APPLE.COM, TIME.COM, DAILYMAIL.CO.UK, SKYPE.COM, STATISTICBRAIN.COM

▶ Data is generated many ways, such as Internet browsing, smartphone activity and movement, digital business processes, social media activity, and sensors in buildings, products and people. Yet all analysis indicates conclusively that the growth in data will not be based on human activity but instead on the Internet of Things.

The need to capture, process, store and analyze data to generate corporate value has generated the emergence of a new breed of technologies, including NoSQL data stores, HADOOP, Massive Parallel Processing (MPP), in-memory databases and distributed file systems.

Although Big Data is still an emerging area, its value has already been proven. Examples from companies such as Google (applied to its search engine), the McLaren F1 team (to manage real-time streaming data from an F1 car during a race) and Amazon (applied to its real-time recommendation engine) have shown that the use of Big Data solutions can transform and optimize business processes. From these frontrunners, the opportunity to capitalize upon knowledge derived from the data generated has now reached more 'ordinary' businesses and consumers.

As the Internet of Things, Smart Cities, Big Data and Cloud drive more and more data processing requirements, it is predicted that by 2020 there will be 50 Zettabytes of data generated by more than 25 billion connected devices. The growth in data is exponential and far surpasses Moore's Law as an expression of growth for the IT sector (**Figure 5**). ▶

Source:DOMO

Detailed Findings & Analysis

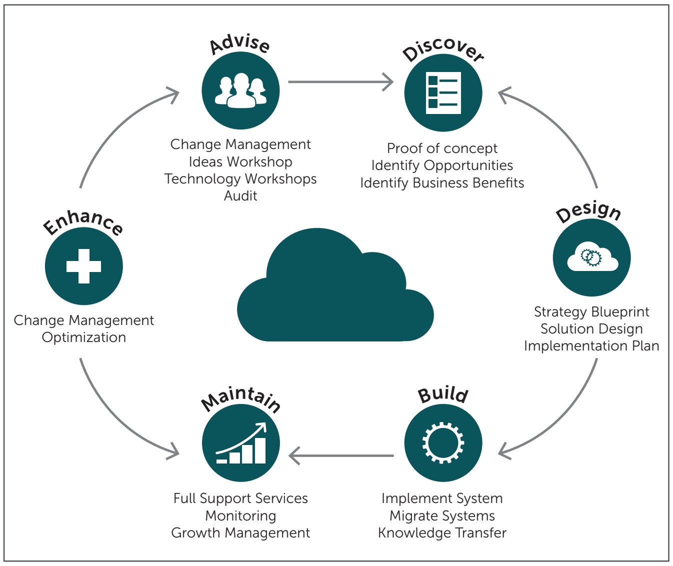
▶ The growth in IoT comes from many sources, almost anything that can house a sensor. Oyster and salmon farmers in Tasmania produce data to measure water temperature and quality in order to effectively manage threats to the health of their stock. The list of IoT devices is potentially endless – according to the Cisco Global Cloud Index 2015 to 2020. The Cisco data are indicative of a city of one million but they indicate that these sensors and devices will contribute enormously to the data that will overall be generated. What is also interesting is that most of the data generated is discarded (**Figure 6)**.

Figure 6: So Much Data Generated, So Little Used

Source	Data generated per day	% Transmitted
A connected plane	40 TB	0.1%
A connected factory	1 PB	0.2%
Devices to improve public safety	50 PB	Less than 0.1%
Weather sensors	10 MB	5.0%
Intelligent building	275 GB	1%
Smart hospital	5 TB	0.1%
Smart grid	5 GB	1%
Smart car	70 GB	0.1%

Source:Cisco Cloud Index

Figure 7: The Analytics and Cloud Cycle



Source:Quoted in DCDi: Big Data & the Data Center

The plethora of information represents a major opportunity for organizations as well as a major challenge. The opportunity is that by using data in an active/interactive way, understanding people or machines on the basis of immediate readings and analysis that enable the recipient to respond. A company can text information about a new product once a known (consenting) customer gets close to a store (physically or online), a 'driverless' car can slow or accelerate, remote monitoring of serious medical conditions can be linked through to medical services.

The Role of Analytics

The difference between a mass of data and actionable outcomes or strategies is analytics, part of data science. Analytics is the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions. Analytics may be the input used for human decision-making or may drive fully automated decisions. Increasingly analytics requires a parallel understanding of the programs that process and analyze data. In short, analytics enables decision-making based on data and facts.

The key requirement of defining outcomes and action standards from data before it is collected is important because it means that the collected data is actionable. That means defining specific business objectives, determining the scope of the analytics strategy, creating a team of experts, defining an improvement process methodology, and selecting and implementing the right tools, technology and data integration.

This is not a new process – historically, research has developed action standards based on risk parameters so that the data can be fed into a cycle of knowledge and action. However it is process with risks – how do you know if you are looking in the right places? How do you know if you are throwing away valuable data? Or keeping junk?

As depicted in **Figure 7**, The Analytics and Cloud Cycle process beings with identifying the corporate requirement ('Advise'). Then, it evaluates how knowledge gleaned from data can meet requirements ('Discovery'). It then moves through the stages of 'Design' and 'Build' which refers to systems as much as buildings. Finally, the process moves back through a process of 'Maintenance and Improvement' as the knowledge gained is applied and refined to the process. ▶

Detailed Findings & Analysis

▶ Data Curation = Efficiency

In a situation where networks will come under pressure, efficiency in data transmission will become increasingly necessary. This is where 'edge computing' and the process of data curation, selecting that which is valuable from that which is not, comes in. As the amount of data running through networks increases so will the number of problems caused by bottlenecks, latency, and interruptions. If all data generated in 2020 were to be sent from source to core production facilities the current system could not effectively handle the traffic. Even if just the 'useful' data were to be sent, this would still challenge the capabilities of the current (2017) network in the United States. This means that data curation is a key. Edge computing and efficient data curation are both necessary to allow networks to keep pace with demand.

While volume is an issue, so is the cost of networking in terms of both capital investment and ongoing operation costs. Of all the items tracked in the DCD Census cost of networking is the fastest growing (CAGR of 9.1% Capex and 6.7% Opex 2012 to 2015). In fact, it is growing faster than expenditures on facility or IT equipment. While the costs of WAN have decreased considerably (one study indicates a decrease from $100 per Mbps per month to $5 per Mbps per month), this may change in a situation where there is intense competition for connectivity. An academic study of [The US Long-haul Fiberoptic Infratructure, Durairajan, Barford, Sommers, Willinger] mapping fiber based on data from Tier-1 ISPs and major cable providers indicates that the USA has around 113,000 miles of long-haul fiber-optic cable ('the physical Internet'). The study also highlighted risks to latency and connectivity associated with shared infrastructure and traffic congestion. There is little published regarding the cost of getting the network to a state to deal with 2020 traffic volumes although it it is known to be substantial. For instance the cost to upgrade Boston has been estimated at $300 million while Google's investment...be considerable – to upgrade Boston is quoted at $300 million.Google's investment in Kansas to create a Google Fiber City was estimated at $94 million in 2013. This led Goldman Sachs to estimate that (at 2012) it would take $140 billion to extend the network across the USA.

Key here is the process of determining which data is valuable as opposed to the data that needs to be discarded. (Estimates of this range from 90% to over 99%). Cutting down on the data that is transmitted it will save on bandwidth, and thereby save on the power and maintenance requirements of the associated networking equipment.

The importance this process of data selection and organization is the reason that the investment pattern of edge computing is the reverse of grid computing. The edge data center or edge processing unit commands a high share of investment costs, since the edge is where the work is largely done. This changes the role of the core production facility to one where it acts in an Enhancement and Advisory role focusing on modifying the processes and analytics to improve outcomes. The core production facility acting in the above referenced Enhancement and Advisory roles, modifying the processes and analytics to improve outcomes.

The Limitations of Cloud = Latency

Edge computing has gained traction due to some of the limitations of a Cloud-based analytics system and therefore acts as a bridge into core Cloud production/storage facilities:

- Cloud cannot offer the low latency required by some devices, which require immediate response to data to take immediate action

- Cloud is also not good at transmitting video which is one of the key content components of peer to peer sharing

- Cloud cannot offer the real time experience expected of augmented or virtual reality

- Academic tests* have indicated improved latency times for edge compute functions over Cloud, from 900 to 169ms. The use of cloudlets for computing tasks where the source is wearable cognitive assistance reduced response time by 80ms, to 200ms.

Dependence on the Cloud computing model reduces the user experience for virtual and enhanced reality. As noted in the previous chapter the variety of different edge use cases does not cut Cloud out of computation at the edge of the network altogether.

Data Security = Compliance

Data privacy and security at the Edge follows the same rules as data privacy and security anywhere. Possibly a large amount of data generated from a small geographic area or from a few individuals may make them more vulnerable to cybercrime (or, for a Smart Home user, to physical crime also). It is estimated, for example, that half of all home Wi-Fi systems are unsecured or set on default passwords [Wi-Fi Network Security Statistics: BotRevolt]. The protocols for data protection on the Edge are still in development, but the risk is dependent on the effectiveness of those protocols and adherence to security and privacy procedures at the edge. Yet the principles of the Edge does put additional pressure on ensuring that data that is not used and returned to the core is suitably destroyed.

Changing User Role = Expectation and Delivery

IT has created performance expectations among users, whether they are individuals, businesses or governments. This means that IT has to continually upgrade and improve the user experience it delivers. All of the expectations below build the case for edge computing, for sharper connectivity and for enhanced customer delivery:

- Immediacy – connections that work instantly, games where the illusion of 'real' is maintained by actions and reactions that are real-time, algorithmic trading, content production etc.

- Mobility – that virtual experiences can be enjoyed or used

- Interactivity – the flood of social media has taught consumers that they don't have to just consume but that they can produce and share data, and use this as a means of shaping their world. Meanwhile, business and government have learned to capitalize on this need for interactivity. ▶

* As quoted in Shi, Cao, Zhang, Li, Xu IEEE Internet of Things Journal, October 2016

Detailed Findings & Analysis

▶ Use Cases

Reference is made in this section to 'Edge processing units'. These are the devices where edge computing takes place, usually through processing the data that is collected from the immediate environment and making or recommending actions to be taken on the basis of that data. They will vary in processing capacity, size and configuration according to the processing requirement and the options for housing it.

Edge computing will in time be considered for any application where there is the need for immediate data processing based on one or more of the drivers described above. Currently the most common scenarios being evaluated include:

Figure 8: The Google Self Driving Car

Source: Google

Automated vehicles

The self-driving car is perhaps the most public example of edge computing since one car will require an estimated 200+ CPU's. In the words of Peter Levine of Andreessen Horowitz it is effectively 'a data center on wheels'. The car needs to process live video and streams of photos and make immediate decisions based on that input. At a time when self-driving cars become more common, they will be able to share information to collaborate on decisions. Data will be sent to the Cloud for updating the protocols for all relevant cars (**Figure 8**).

Other Transport

Edge computing does and can perform a number of functions for commercial and public transport. For highly complex vehicles such as spacecraft, airplanes and ships the processing allows only the valuable information to be transmitted for further analysis. Local processing allows (as with the self-driving car) information to be fed to the vehicle controls whether human or automated. The second function is informational – to build data patterns into traffic systems that can be used on a unitary basis to improve travel efficiency and safety.

Media/content

Edge computing here represents the upgrade and refresh of content delivery networks. CDNs were originally developed 20 years ago to deal with the increasing richness of content (from text to graphics to video) and to increase the quality of the user experience. While CDNs were initially unable to keep up with the increase in demand for content that emerged with the growth in broadband internet services, the growth in video content, together with ubiquitous internet access on a wide array of devices and the availability of an increasingly sophisticated range of services, mean CDNs became a vital tool in delivering internet content to businesses and consumers globally.

The data center was a key component of a CDN, because the CDN point of presence (PoP) needed to be housed in a physical location. Edge computing will play a part in future content delivery to enable providers to expand their geographical reach and to maximize the efficiency of the delivery networks, particularly as they introduce increasing numbers of value-added and interactive services. The CDN already brings content closer to the user; edge computing therefore would represent a logical further step in that process by shifting more of the operational applications closer to the user. A review of the Internet suggests that, to date, this is one of the areas most active in developing edge computing.

The Smart Home

The claim by a number of data center OEMs that every home in the United States will soon become a data center, is closer to becoming a reality, although possibly not in the sense intended by the claim as edge computing links the home back to the core production center rather than creating it as a data center in its own right as the edge will likely do.

The home already has a number of 'smart' devices that respond to stimuli within the environment – Fit Bit and wearable technologies and Amazon Echo, for example. There are refrigerators that are able to suggest what food might need to be purchased or go directly to a pre-set online retailer to access needed products. Home security has a need to develop a smart component to help distinguish between true problems from false alarms.

Yet these advances don't represent a smart home that uses edge computing since the examples above tend to represent single lines of activity (excercising, shopping, and security) rather than building into a home that is enabled and managed by edge computing. It takes more than an internet connection linked to the Cloud to create a smart home.

A smart home would need to rely on far more extensive information collected around the home from sensors and controllers which would then be analyzed and acted upon within the home. In a situation where domestic power costs are likely to escalate, particularly as the cost of investment in sustainable sources and 'cleaner' energy is recouped, the smart home will be configured not merely to switch lights or heating on or off but to sense and switch between grid and local renewable energy sources as well based on cost and sustainability. The analysis might also include the interaction with local providers, with weather predictions, possibly in collaboration with the local community to make the process more collaborative and interactive. ▶

Detailed Findings & Analysis

▶ Smart City

The use of edge computing can be applied more broadly to the smart community or city. As the number of sensors and sources grow (individuals, traffic systems, healthcare, utilities, security.) so the principle of storing and analyzing it all centrally becomes less feasible. A system based on the principle of edge computing allows the information of value in meeting requests to be passed back through the network. Edge computing also reduces latency delays in situations where action is required quickly (medical emergency, criminal activity, traffic congestion, train delays etc.). It also, by the nature of edge computing, allows for geographic precision – information relevant to a particular street, block, or suburb can be shared instantaneously. For example, in the event of a local crime, edge computing facilities in a neighborhood can be fed an image or video of the perpetrator which video analytics can search for and link to police systems.

Factories/production facilities

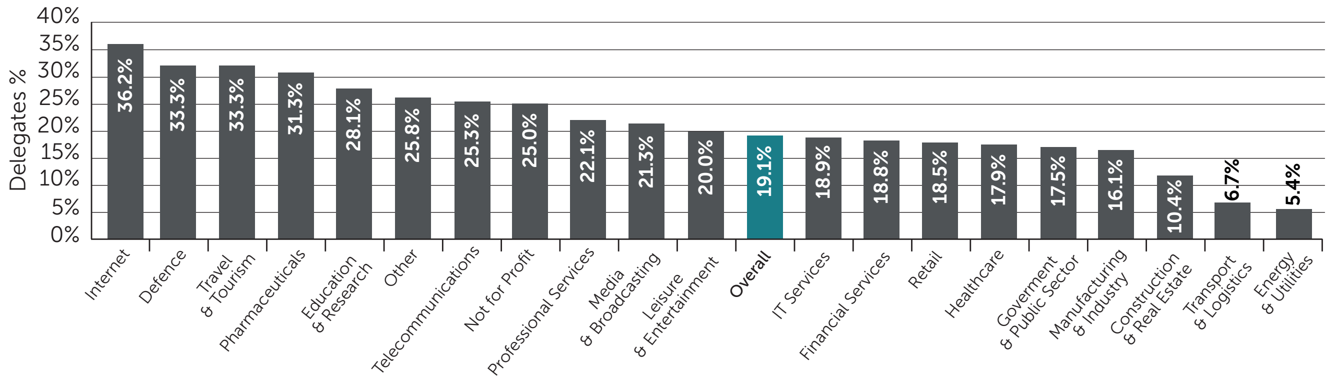
The fourth era of industry and production relies on digitization to manage and improve processes in terms of efficiency and performance. The three previous eras of industry relied on water to generate power (1), electricity (2) and the operation of computers (3). The use of robotics, AI and machine learning within some industrial and manufacturing organizations means that adoption of edge computing has been taken further in this sector than in most others. Part of the principle of the fourth era is to weave production seamlessly into a whole process from the reading of demand, through producing based on that demand to delivery, post-sale service. This requires the very type of collaboration between data sources across a range of locations and of companies that edge computing provides.

There are many other possible applications of edge computing which will continue to increase as, following the IT paradigm, the expectations of what is possible in terms of the user experience also increases.

From the 2016/2017 DCD delegate samples, the sectors which assign the greatest priority to core-to-edge are diverse, and include ISPs, the defense industry, travel and tourism, pharma, education and research. This indicates that the principles and benefits of edge computing are perceived to have relevance to a number of different user cases. Perhaps surprising is the status of IT services at just below average but this is a heterogeneous category and if separated out, colocation indicates higher priority at 28% (**Figure 9**).

The decision to engage edge computing in reality will present a series of practical options for companies. That is why we look next at strategy for the Edge. ●

Figure 9: Priority Assigned to Core-to-Network Edge by Sector: USA 2016/17



Source: DCD Group 2017

Detailed Findings & Analysis

Chapter Four: The Edge Strategy

As with any major IT investment, investment into edge computing needs to be based on a strategy that meshes with business objectives and outcomes. There also needs to be some element of caution as any major trend to edge computing will represent the latest swing of the computing pendulum. Remember, this trend started from the centralized topology of mainframe, through the distributed topology of servers, back into the centralization of Cloud and now into distributed edge computing. This means that any form of investment needs to carefully consider how the computing model may re-centralize (or whether computing will ever find an optimal point of balance between aggregation and disaggregation).

The key components of developing an edge strategy and delivering on it are as follows:

1. Defining the objectives and requirements of such a move in terms of business, brand, customer and ROI requirements.

2. Mapping the network topology from the edge back to the core.

3. Defining the systems, protocols and programs that constitute the processing, abstraction and communications capabilities of the unit.

4. Defining the networks that link the unit to data sources and back through to the core processing facilities.

5. Developing strategies for the supervision, maintenance and protection of the edge computing system.

Since the initial foray into edge computing in the United States has been largely leveraged utilizing the digital infrastructure on digital infrastructure provided by colocation and telecom providers, the profile for an edge strategy can be used as the means of evaluating a suitable outsourced approach to capitalize on the opportunities that edge provides.

Defining Objectives and Requirements

The first component of the strategy is, essentially, to decide whether edge computing is the correct option for the organization. With the enormous increases projected in IoT and the expectations of immediacy and accuracy that this creates, questioning the need may at first seem unlikely. Edge computing would appear to present major benefits as the number of connected devices proliferates. Yet the extent to which it rolls out quickly and widely will depend on cost, on maturity, and on the speed with which those few organizations which have the necessary infrastructure, configure that to edge computing purposes. Possibly only after 'Cloud ready' 5G capable networks arrive will the shape of the deployment become more apparent. Maturity will reflect the state of programing, security and naming protocols for the source units and traffic systems of the edge computing universe.

The strategy at this point in the evaluation of edge should describe the targets for edge, the benefits, the objectives for the company and the return on investment. In particular, in a situation where a company is going to invest in its own edge processing unit and what forms this should take, the financial strategy needs to identify a model of how investment in these small, networked units will produce a return and over what time period.

Mapping the Network Topology

At the edge, the topology will be defined by the number and variety of sources, the complexity of their linkages and the commonality of the edge computing requirement. This commonality helps develop different types of edge computing, broadly:

- **Latency reduction:** where the overriding benefit is to reduce the latency of transmission and thereby improve user experience.

- **Video transmission**: where the content transmitted is video, for consumption or analytics.

- **Multi-edge:** where the source moves seamlessly between edge nodes. For example, a customer using a mobile device from their car to their home security system. The customer sees video playback that the alert received was a package delivered and left by the front door. He next receives an alert from his refrigerator that he is low on milk and butter so next moves onto a traffic site to determine whether there is time to stop before a scheduled meeting to pick up these items. The items are ordered via an automated shopping application and picked up via grocery drive through. The customer then is able to quickly check on the health of his elderly father and reserve a parking spot for the meeting as his car negotiates the most efficient route to the meeting.

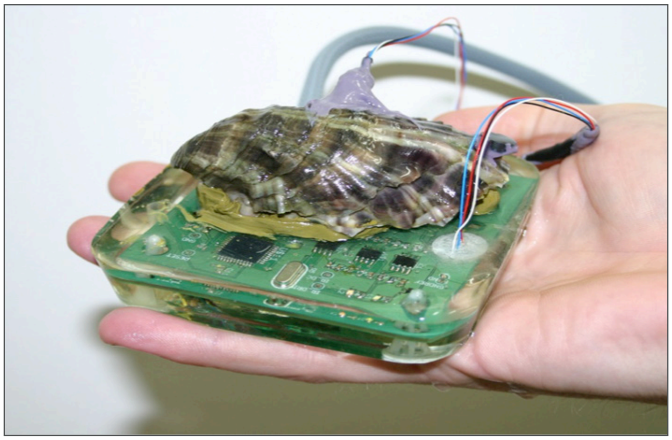
- **Collaborative edge:** where stakeholders can share a virtual view of a source from a number of related perspectives, for example, healthcare. With an aging population and growing number of managed health conditions such as diabetes, a collaborative edge can combine data on hospitalization, medication, diet, exercise, management (blood sugar, ketosis), health insurance and also locational information in the event of an emergency.

- **Shared edge:** an edge processing unit which will be shared by companies needed access to data sources in a particular area or locality.

- **Linear edge:** where the edge is developed for linear process such as logistics or manufacturing that has a set start-point and a set end-point and where the edge is engaged to make the process more efficient or geared to a customer requirement. Thus, sensors in the soil are used to transmit key chemical analysis and information on crop growth. From this information, an automated facility sorts food waste from local supermarkets to be fed to worms who process it to provide the chemical mix required for each crop. There is collaboration with transport and agricultural service companies to lay the fertilizer, although in a Smart Crop facility this too can be done automatically.▶

Detailed Findings & Analysis

Figure 10: Edge Computing Oyster-Style



Source: CSIRO

▶ The topology, the number of nodes, and the linkages will determine the configuration of the edge processing unit and the extent to which it requires power, processing capacity and bandwidth/connectivity. This will determine the design and topology of edge processing units which we discuss in the next chapter.

The greater the processing capacity of the individual edge processing units, the more direct the network connection to the core production unit is likely to be. Large numbers of edge processing units are more likely to require a greater number of nodes to reach the core.

Defining the Systems, Protocols and Programs

Edge computing systems will require a number of complimentary IT systems and processes:

- First, analytics which determine the valuable data, the data which can be discarded, and the actions that need to be taken on the basis of what is learned. This form of analysis relies heavily on algorithms – sequences of instructions for data processing. Since each company will have different data requirements, the instructions that take the raw data and abstract that data for further analysis may need also to be unique. These systems are already relatively mature in core data centers, therefore the key task is to establish the mirror versions within the far smaller processing capabilities of the edge unit.

- Operating systems which manage the process whereby data is relayed from the sensors and controllers to the edge processing unit.

- A shared language which enables multiple-edge systems across variety of platforms. These languages are evolving with edge computing, just as the access to Cloud developed via programming language. It is possible that edge platforms running different languages can be made to speak to each other using computing streams that enable interprocess communication in a manner deployed previously across grid computing systems.

- Protocols that enable the systems and the network to work together. Dell and Intel are in the process of developing IoT gateways and routers that can support edge computing while software such as Apache Spark is also being developed and scaled.

- A naming system (like the Internet IP system) that can act as a means of identifying and communicating with devices and users on the Edge. Such a system needs to cope with the huge number of devices there will be, able to deal also with mobility and security. It will need also to be user-friendly so a user can interact with a number of different devices.

Edge processing units will need to be lean in terms of their use of resources and that this will constrain the resource requirements of some of the software present at the Edge. It is possible that, for identification, a system that is based on a hierarchical framework (such as the IDD>SDD>number for fixed phones) will be more efficient than one based on a flat universal ID system (such as an IP address).

Defining the Networks

One of the most likely means of enabling the huge network that a core/edge computing system will require is software defined networking. This is the most established and mature of the software defined utilities and is already established in many core data centers, SDNs are currently the most effective means of accessing Cloud systems and the backing of virtually all the major suppliers of networking equipment and fabric systems in the industry. The computing stream generated by a software defined network would enable data to be processed at the Edge, through intermediate nodes and, where this mode is used, back into a Cloud core. The basis of a software defined network would allow also some of the connectivity issues (such as the multi-edge) to be resolved. A software defined network basis presents the only viable option since any less flexible alternative would be inefficient on a very large scale, and more prone to congestion. ▶

Detailed Findings & Analysis

▶ Strategies for Supervision, Maintenance and Protection

It is not going to be enough to build an edge computing capability. It will also need to be operated, policed and maintained which is one strong reason for a roll out based on supply by integrated IT/telecom providers. Once edge computing begins to experience exponential demand, decisions will need to be made as to which organizations are responsible for its overall supervision and which decide, on what data gets priority across the network. Or will it evolve in the more anarchic manner of the Internet bound only by rules of technology and connectivity? This seems unlikely since an edge computing capability requires a far higher level of direct investment and because guidelines will be required for items such as the location and physicality of edge computing units.

There are a number of issues that cannot yet be fully answered but which will need special or additional consideration during development of the edge computing growth phase:

- How data and privacy will be protected. One of the key drivers of edge computing is the personalization of response - from anticipating a consumer's choice of media to producing a pharmaceutical based on the direct identification, analysis and trialing of individual health needs. The downside of this when compared to the traditional means of making the above decisions on the basis of more anonymous data is that more identifiable information moves across the network. It is probable that edge computing will therefore spawn its own data sovereignty debate. As part of this, the ownership of data and rights to how data can be used will need to be defined.

- The growth of IoT and its enablement by edge computing will also complicate the concept of criminal or civil responsibility. If an accident happens for which a self-driving car may be liable (due, for example, to encountering a situation for which it had not been programed or due to system or network failure) where does the responsibility lie? With the human for not overriding the self-driving function? With the vehicle designer? With the programmer? With the network provider if it can be proven that the accident can be attributed to latency in transmission?

- The growth of edge computing will be based in part on the proliferation of devices that are included into the network universe as sources or as producers. A number of analysts question the ease with which large numbers of further devices can be easily incorporated. Therefore, as much as any IT investment, the investment into edge computing needs to be future-proofed, probably at a network rather than a processing unit level. Again, this is why the driving force behind roll out will be third party providers. Just as when customer buys a desktop or a tablet, that customer does not have to build the infrastructure that enables the device to connect.

- Resilience needs to be considered. What is the impact of the failure of a source or of a cell within the edge processing unit? Does it bring down the whole unit? How does it impact collaborative units? Shared programs? How can a system entirely dependent upon a totally fluid network continually protect itself against cyber attack? While security will most likely be provided from the core, it will need a zero-latency response against attacks that have the capacity to spread very rapidly. Just as cell phones offer the capability of making a 911 call without needing to unlock the phone, will an edge computing network offer a similar over-ride capability to protect against unwanted intrusion? Again, this suggests the role of an overall network and unit provider that can build such protocols into the system.

In the next chapter, we look at how the key component – the edge processing unit - will evolve. ●

Detailed Findings & Analysis

Chapter Five: What will the Edge processing unit look like?

Quotations in this section have been taken from a series of interviews conducted with colocation and data center service providers.

While the logic of the need for edge is convincing, the means by which it will be delivered is less clear. Will its roll out require major new investment in infrastructure? How much can be done with existing infrastructure? And can one size fit all?

How the edge computing unit as a mass proposition will eventually be physically deployed depends on a number of factors:

- Site considerations including zoning and compliance.
- Its position in the edge to core network. There may be nodes between the source and the core with larger numbers of smaller processing units at the edge, larger cluster nodes gathering information from the edge units and transmitting this to the core. This hierarchy has been represented in ascending size as mist computing, fog computing, then cloudlet nodes transmitting to a cloud core.
- The capability of the core.
- The type of edge computing to be performed – latency-sensitive, collaborative, shared edge, linear edge. This will impact key operational considerations such as reliability, efficiency, latency, and fail over strategies. These requirements will define the capability of the processing unit.
- Cost – as has been estimated a number of times and by a number of reputable sources, the costs of just providing connectivity across the United States to a bandwidth that would support projected data traffic is very high even without considering the additional costs of edge processing units.

It seems logical that the core-to-network edge system should where possible utilize existing networks and infrastructure. This would suggest a combination of the companies that run the estimated 110,000 miles of the United States' long haul / backbone fiber network and those that run the colocation facilities that can bridge between edge and core are well positioned to provide this core-to-edge network. The history of acquisition over the past 5 to 6 years suggests that industry players have been alert to this possibility - CenturyLink and Level 3 Communications, Equinix and Switch & Data, NTT and RagingWire, Verizon and XO Communications and even AT&T's acquisition of Time Warner. The infrastructure that telecom companies have traditionally developed is used as one way into edge computing:

> *"The hundreds of telecom routing nodes represents a huge advantage but they are of differing quality. One is a bunker and it is very impressive but it goes all the way down to fiber hubs. Telecoms were perhaps the first edge, using pedestals and fiber in the ground .... Being able to really leverage those assets provides a big advantage."*

Will existing networks of CDNs be sufficient to support the United States' potential edge computing requirement? In their current form, CDNs will still exist on the networks that edge will use but as units they are closer to the small data center on the edge of the grid than to the edge computing unit. The CDN node does not have the capacity to interact with the numbers of devices and sensors at the edge.

Similarly, the edge will take on some of the distributed principles of peer-to-peer but, like CDNs, the form of technology evolved before the Cloud and IoT eras. Therefore, they can be seen as a precursor but are not robust enough for the IoT requirements in terms of flexibility or scalability or for a model where core is as a necessary part of the network structure. The evidence is that CDNs are beginning to move into edge and shaping it to their own requirements:

> *"CDNs love companies such as Cogent, Light Tower, Zayo for content. They will use them for low cost storage and use someone else for transport. CDNs shop heavily based on price so won't always use one provider for all."*

Much of the academic literature assumes that investment on the edge will be major and will outweigh investment at the core. Yet, conversations with providers indicate that customers may have a more pragmatic and cost-conscious attitude, based on the principle of redundancy as based on a paradigm of load-shifting across the network. This cost-saving approach may be adopted as an organization tests what works before committing to more major investment allowing time for the weak points of the system become apparent:

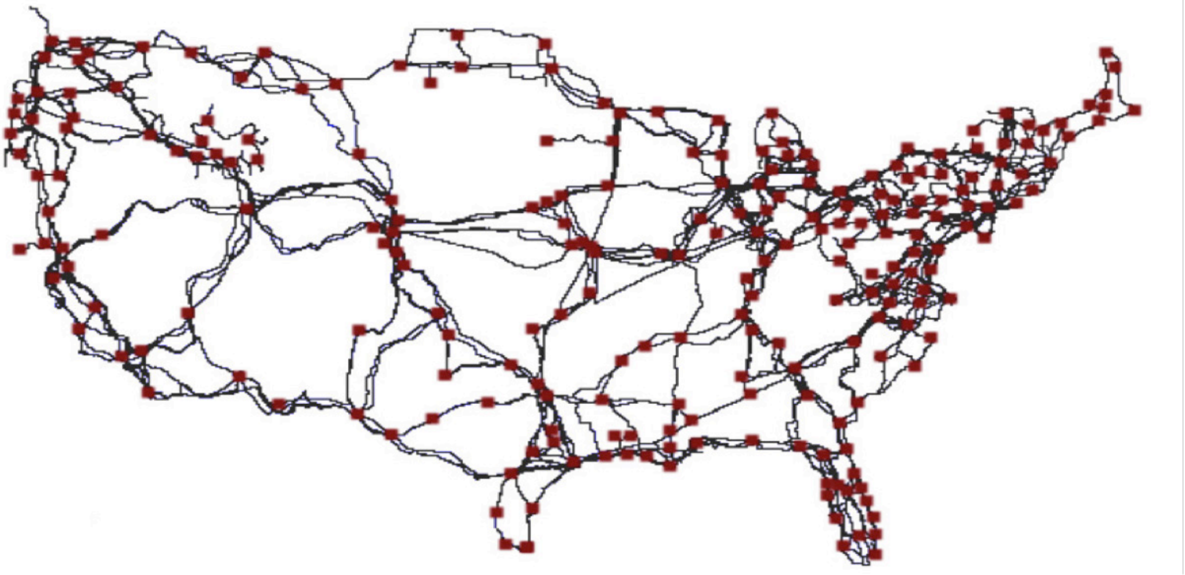
> *"Sometimes, these units are not pretty. Central Office with just power, run as 'lights out' but clients don't care. They don't look at the facility, they are instead focused buying the edge point of connection. As it is just one of many connection points on the network, if it fails, they can just skip to the next spoke – the core stuff is kept in the data center which offers real redundancy."*

When costs were estimated for updating the fiber optic network in Boston against those for a small town in Minnesota. The cost of the first was estimated at only six times that of the second. This indicates that, just as the fiber network does not reach every household in the United States, so the development of core-to-network edge will not either. Edge computing in downtown Los Angeles, CA will vary from that in Tuscon, AZ which will vary from that in Bethany, MO. This gradation of locations throughout the United States means that different solutions in terms of networking, processing and the costs of service will need to be established:
(**Figure 11**) ▶

> *"The industry still hasn't come to a consensus about what 'edge' means and how it will work differently for metro versus rural and for mobile devices. So a facility where you can colocate in major city and then interconnect with networks serving the end user such as the Time Warner and MSO models, that's 'edge'."*

Detailed Findings & Analysis

Figure 11: Location of Physical Conduits for Networks in the continental United States



Source: A Study of the US Long-haul Fiberoptic Infratructure, Durairajan, Barford, Sommers, Willinger

▶ As with 'core' data centers, the design of the edge processing unit needs to meet the requirements of three main elements – IT for processing, analytics and operations/communications protocols; networks both inward (information from the sensors and controllers) and out towards the core, and the edge unit supply systems – power, cooling, racking, monitoring, security/access.

There is emerging consensus that the design of edge computing units will be modular. This makes sense given that edge IT needs to operate from common software definitions and protocols. Otherwise what will emerge will be a variety of disparate systems which will compromise the aggregated value of edge computing.

In terms of edge processing units, what is most likely to emerge will be a family of edge processing units modular in design based on common standards that' like servers, can be used alone or in racks. Some of the space for accomodating edge infrastructure is already in place on sites operated by telecom, colo or utility providers. The emergence of 5G and/or LTE will enable mobile data processing and transmission.

"Albuquerque is not going to be the same as Los Angeles metro. Maybe if there's a building block standard of, 200 kW per unit, that might be the only one block in Albuquerque but would be one building in LA. There's demand for micro-modular DCs one or two two racks in a unit about the size of a refrigerator. You can fit that next to a cell tower, in an office building, in a residential building, anywhere where there is connectivity to devices."

"A large metro market Edge location will be near the network hub. Companies will be willing to pay a premium to be at the hub. This will divide between the 'deep internet' in small metro areas versus the 'broad internet' in large metros. Coverage will come down to what the wireless providers are going to do. But 100 million IOT devices in a place like Denver means that volume will have to drive the data way to the Edge. That could mean building Micro data centers as analytics need to go as far to Edge as possible."

"We may see more containers at the edge when 5G arrives - we don't see a lot of interest in those today." ▶

Detailed Findings & Analysis

▶ There seems to be an emerging consensus that the key to the design of edge computing units will be modular. This makes sense in terms of the physical space / edge processing unit given that the edge IT needs to work from common software definitions and protocols, otherwise what will emerge will be a variety of disparate systems which will compromise the aggregated value of edge computing.

Mobile edge computers – in cars, planes, trains, trucks, etc. These will necessarily take their form from their host, and be linked closely to their operating systems. Power to these units can be provided by the host with batteries providing temporary back-up for IT systems.

Home/ building computing unit - These may take the form of a micro-modular data center the size of a utility cabinet which can then collaborate with other local units to form a smart community or a Smart City. Power to these units can be provided by the host with batteries providing some IT back up.

This may form a collaborative node on the way through to the core data center, reducing the need for larger edge computers. This may be the natural place in the core-to-edge network for the containerized half- or full-rack unit sited close to telecom facilities.

Shared edge computing units – Those shared by different businesses with a shared goal (for example, healthcare or education). These may need to deal with sufficient data volume to justify the use of containers.

Distributive edge computing units – These may be shared by different businesses or operated as part of IT/data center services (where the core will be located in a colocation facility). These primarily will manage and deliver the distribution of videos and other 'rich' and latency-sensitive content.

Linear/Process edge computing units – these will bond different companies or different parts of an organization to meet a common production goal from the beginning point ('requirement') to the end point ('delivery'/'re-engagement'). In the situation of finite and 'private' use, the edge computer located somewhere close to the point of production or manufacturing may take the role of the core as information is shared along the line to manage and deliver the best possible product at greatest efficiency. It is possible that among high-spec manufacturers, some of the edge computing functions may be shared within existing computing resources including private cloud. Of all of the types of edge computing, this form is the least likely to utilize third-party providers, for reasons of security and protection of intellectual property. ▶

Detailed Findings & Analysis

▶ While there appears to be increasing momentum behind edge computing, the jury may still be out in terms of whether this is a major defining trend of the next few years. Interviews conducted with colocation/data center service providers indicate a mixed reaction, particularly from more 'traditional' colocation providers. The 'case against' is based on the growth in network capacities and speeds which are seen to have already reduced latency, some perception that the services that would benefit most from lower latency are not 'essentials' and a perceived lack of demand:

> *"The concept is not new – it has been around for 15 to 20 years now when ISPs were offering cache processing for Hotmail. Obviously the networks were slower then so as demand for bandwidth grows we hope that will drive demand for local data replication and caching but we are really yet to see that. It depends on higher bandwidth services. We are yet to identify where opportunities for it exist - I have seen it hyped but no one has come to me looking for an edge solution. The replication and caching of data closer to the user may improve the user experience and reduce latency but fat pipes are now available. In the office here we notice the greater latency in services from the cloud rather than from an on-premise server but the cloud is a lot cheaper. Maybe other database-intensive stuff like Salesforce which requires a lot of data crunching, there would be value in having a proportion of that closer to the users. But as for demand, I haven't seen it."*

> *"Edge makes sense but it's still niche stuff. It's CDN stuff which needs low latency so people don't skip your website but the network now goes further, the pipes get bigger, fiber gets more robust and it's going from 4 in bandwidth to 10 to 40 to 100. There are now 3674 strands in a dark fiber bundle."*

The more positive view recognizes the potential of edge, and it is noticeable in part that this is based on demand from clients. Edge here might be developed on the basis of property availability – edge at the central city where there is little space and core in the suburbs where there is greater potential for expansion:

> *"It makes a lot of sense – our clients look for leverage, cheaper power costs, higher network density and more interconnectivity. It will be impossible to add much capacity downtown but there's a lot of land and power available here in the suburbs. So it is possible to create an edge DC at the network edge where there's little room to expand."*

Location may also be selected on the basis of costs of installation and operation – this may leads to the separation of computing and networking nodes: ▶

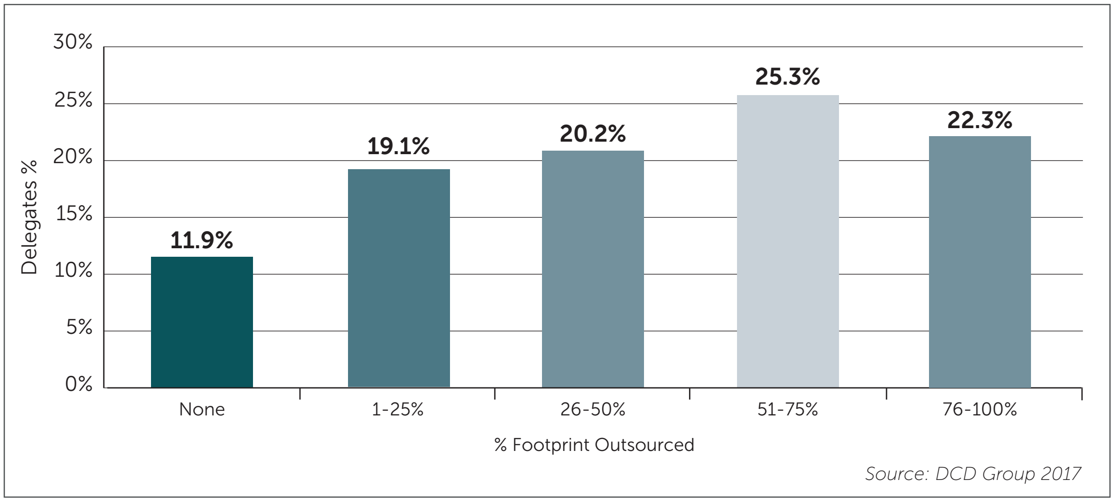
> *"Decisions about where to locate will be based also on the cost of the area. There seems to be two approaches. The first is to amass a consortium of MSO's and content providers, and build facilities. These will be pivoted to building where the cloud providers are highly latency sensitive otherwise it is not as important. The public cloud can act as a demand aggregator, for example in places like Ireland and Amsterdam. This may lead to the separation of compute and connectivity nodes - keeping connectivity in the more expensive areas but pushing compute farther from the edge."*

Detailed Findings & Analysis

▶ **Companies that already Outsource are more likely to be interested in Core-to-Edge.**

The priority assigned to core-to-edge bears some correlation to the current dependence of the organization on outsourcing to meet its data hosting and/or processing requirements. Organizations that prioritize core-to-edge outsource an average 29.4% of their footprint compared to 23.3% of those that don't. There is an increase in priority as the proportion of footprint outsourced increases before the proportion drops back among those closest to total dependence on outsourcing. This is probably because the responsibility for the functions associated with edge computing will be seen as the responsibility of the colocation, hosting or possibly the cloud provider and the organization will consider them if offered (**Figure 12**).

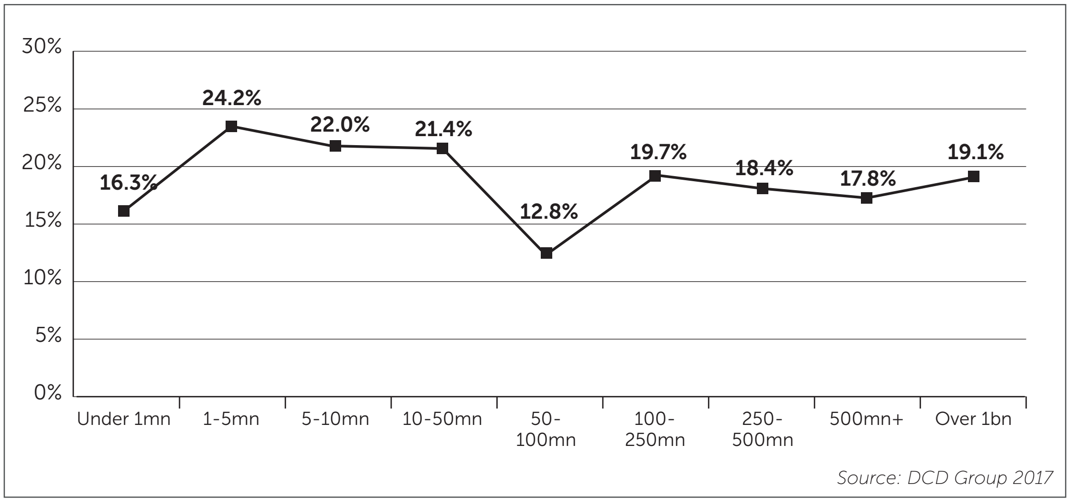
Figure 12: Companies that already Outsource are more likely to be interested in Core-to-Edge



Source: DCD Group 2017

There is little variation in priority assigned to core-to-edge by the size of overall IT budget. Organizations that prioritize core-to-edge have an average annual IT budget of $144 million compared to $155 million among those that don't. The drop among organizations which have a budget between $50 and $100 million appears to be a sample idiosyncrasy. (**Figure 13)**

Figure 13: 'Core-to-Edge' is given Priority by all Sizes of Organization (based on IT budget)



Source: DCD Group 2017

So, in terms of the future there are some aspects of the development of edge computing that appear reasonably certain, some that appear likely and some where the path is unclear. Accordingly, we conclude this paper with the thoughts and insights of Brad Hardin, the CTO of the Paper's commissioning organization, Black & Veatch. ●

The Black & Veatch Viewpoint

BLACK & VEATCH

The Black & Veatch Viewpoint

Brad Hardin, Chief Technology Officer, Black & Veatch

The White Paper indicates that despite the considerable technological inevitability driving edge computing, there is uncertainty about the form and the speed with which it will roll out. To add a Black & Veatch perspective on the opportunities and challenges that edge computing represents to the American market, DCD talked to Brad Hardin, Chief Technology Officer at Black & Veatch.

Hardin is responsible for Black & Veatch's program of software and hardware development, for the cultivation of innovation within the company and for investment decisions relating to these processes. In addition, Hardin works closely with leading providers in the data center and technology sectors and has a keen professional interest in where technologies are heading and how this impacts Black & Veatch and their customers.

Edge Computing is Inevitable

Hardin has no doubt about the emergence of edge computing, based as he sees it on the incessant generation of more data by humans and by machines. He argues that this increase in data generation and traffic will require new solutions and deployments, just as Cloud emerged to meet the increased storage and processing requirements of IT earlier in the decade:

> *"The need is real because we have created such a considerable volume of data, and that led to the shift towards Cloud for batch and data processing. There will be something after Cloud. We require so much storage and processing capacity that we cannot build enough data centers. So this is a roadblock – either someone will invent a new solution or we do it through better management and by being more efficient."*

Hardin describes the driver behind edge computing as a mixture of both new technological solutions that make Edge possible, and changes to behavior whereby users (and devices) are configured to local, peer-to-peer data sharing and usage rather than sending increasing amounts of data over the network: ▶

> *"What becomes interesting is that where we perform compute becomes important – it is pretty apparent that sending data is not the most efficient way of doing things but that we should follow more of a peer-to-peer route, a more direct way whereby we share and store data differently to cope with the rise in devices. This will mean that human and mechanical behavior will change and become more local."*

The Black & Veatch Viewpoint

▶ Edge Computing as 'Smart Infrastructure'

Just as enterprise is working out the possibilities of edge computing, IT providers across the industry are also doing so. For colocation and data center service providers, it is a question of balancing network capability against demand, for vendors of solutions and equipment, the issue is one of design based on new and developing common standards to meet demand.

The mixture of influences and antecedents for edge computing may make it difficult for supply companies to determine what they are able to offer. Hardin suggests that specific discussions of IT or telecoms may narrow the debate and provide edge computing with a wider discussion about infrastructure and how the concept of 'Smart' can be applied to it:

> *"There is huge opportunity in the shift from a utilitarian view of infrastructure to the idea of 'smart infrastructure'. If might just build just with utility in mind without a deeper look, you might just build a sidewalk. With a smart view, you see the sidewalk as maybe housing heat panels or traffic sensors. Huge possibility exists in the public infrastructure space and people are hungry for data; they have shifted from a utility view to a smart view. Therefore our biggest opportunity is that we are an integrator rather than just a provider of devices."*

Hargin illustrates this point by using the example of the work that Black & Veatch has carried out for alternative fuel vehicle infrastructure including hydrogen fuel cells and electric vehicles.

> *"Black & Veatch has a unique ability to be able to design and build to scale. With alternative fuel infrastructure, you have to be able to create infrastructure that is dynamic, scalable and easily accessible. Our vehicle charging infrastructure portfolio is approaching more than 200MW of capacity, including support of the deployment of the largest contiguous DC fast charger network in the world. Alternative fuel isn't just EV, though. We also helped developed True Zero's California Hydrogen Network for fuel cell vehicles (FCVs) - the first hydrogen network for FCVs in the U.S. "Smart" infrastructure goes beyond personal vehicles. Black & Veatch is looking holistically at clean transportation ecosystems, from personal cars to transit and commercial vehicle fleets. ."*

Future Paths?

When thinking about future paths towards edge computing, Hardin reiterated his earlier point that innovation needs to act as a circuit breaker to deal with unrestricted demand:

> *"Thinking of where we are today and where we are headed, we will either be reliant on what we are currently doing - more of the same – or innovation will create new choices and disruption will happen. Once that has happened, then people will try to monetize from that disruption. A lot of those technologies out there are having real impact - Blockchain has the finance sector still reeling and nimble start-ups like that can have a sudden and impactful result."*

There are a number of possible sources of such innovation, those represented by quantum computing and machine learning particularly excite Hardin:

> *"I think that it will be a case of computing better and faster. I think that quantum computing is closer than we realize and offering the capacity to store data on particles and overcome the limitations of silica means that quantum is the next advance. Quantum has to happen just as Amazon saw the numbers of their servers that were under utilized, as a way of making money from those assets which led to the AWS Cloud."*

> *"It doesn't matter how much data is produced – it will be served up, filtered, and deleted by machines."*

The Development of Black & Veatch Capabilities

The Black & Veatch approach is based on considerable experience of not just designing and building infrastructure but rather from analyzing it to deliver at a 'smart infrastructure' perspective:

> *"The answer lies in our view of infrastructure. It's not hard to find yourself on a road, railway, using utilities or some other form of infrastructure. That's what we build but we have evolved deeper domain knowledge. We have a Data Center group and a Smart City group and we know also how power plants work. We are not experts on all things but we see IoT as critical to the development of infrastructure and we are well positioned to do it first. We see ourselves as an integrator in that role ...connecting the dots."*

The perspective has enabled Black & Veatch to achieve very high ratings (1st or 2nd) globally for performance in power, water and telecommunications projects.

> *"It's the way we think about things - our way of looking at the world is different and we are a privately owned company which is unique for an engineering firm of our size. This allows us to develop a longer term view of client relationships and also enables us to be more innovative."* ●

DCD > White Paper

Black & Veatch delivers engineering, procurement, and construction (EPC) solutions for data centers and high tech projects worldwide. We are routinely ranked in the top five in the world in Energy, Water and Telecommunications. With over 110 offices and over 10,000 professionals, we are uniquely positioned to provide data center EPC services at a whole new level.

At Black & Veatch, we are Building a World of Difference ®

www.bv.com/markets/data-centers

BLACK & VEATCH

Black & Veatch Today

110+ offices

AND

12,000 global workforce

$3.2B 2016 revenue

Founded 1915

projects on 6 continents

MARKETS

Energy

Telecom

Water

7,000 active projects WORLDWIDE

Black & Veatch provides end-to-end data center services

- Planning and Global Site Search
- Consulting Services and Project Definition
- Architecture
- Engineering for Data Centers and Infrastructure
- Power Supply and Distribution
- Water Cooling/ Water Supply
- Networks/ Communications
- Renewable Energy
- Upgrades, Expansions and Enhancements
- New Data Center Construction
- Global Procurement and Sourcing
- Modular