Technology Trends

Edge Computing

Enterprise Architecture, Chief Technology Officer Branch

Version 0.1
Date 2019-7-26

Shared Services Canada Services partagés Canada

Canada

Table of Contents

Business Brief[1]

Edge Computing is a distributed information technology (IT) architecture in which data is gathered, triaged, analyzed, stored (short-term), and processed at the periphery of the network (i.e. as close to the originating data source or user as possible). Edge Computing is a part of a distributed computing/network topology that requires and includes a "core", either an enterprise data center or a centralized cloud service. Data may be processed at the point of origin by an intelligent device or sent to an intermediary server located in close geographical proximity. Data that is less time sensitive is sent to a cloud or data center for historical analysis, big data analytics, and long-term storage. According to Gartner, by 2022, more than 50% of enterprise-generated data will be created and processed outside the cloud or data center.[2] The move toward Edge Computing is driven by mobile computing, the decreasing cost of computer components and the sheer number of networked devices in the Internet of Things (IoT).[3]

The "edge" of Edge Computing, is usually identified where users and physical things connect and converge with the digital world.[4] There is no single edge, there are many edges in a network topology with multiple connections or hierarchy levels.

A fully autonomous, closed-loop system that does not interoperate with a core, does not meet the Edge Computing definition and should be referred to as Local Computing. If a sensor is only connected to a local computing device (personal computer (PC), a Programmable Logic Controller (PLC), an autonomous car, etc.), the sensor is not at the "edge". It is simply a sensor that sends data to a local computer, which only local users can interact with. Similarly, if a person is interacting with a computing device (for example, word processing or creating art using virtual reality tools with a headset and a PC) this also is not the edge of anything.

An Edge Device can be any device that produces data. These could be sensors, industrial machines or other devices that produce or collect data and are intrinsic to the IoT. The IoT refers to physical devices (also known as "connected" or "smart" devices) that connect to each other via the Internet or Mobile to Mobile direct communication. Much of the data produced by these devices and their users may not require in-depth processing or analysis as the interaction between them is often unique and fleeting. However, in that instance, a quick turnaround of data and action result is

---

[1] Further information will be added to this document based on additional research and consultation during the fiscal year 2019-20.
[2] Thomas Bittman, Bob Gill, Aapo Markkanen. How Edge Computing Redefines Infrastructure. Gartner. Published: 23 August 2018 ID: G00366225.
[3] https://searchdatacenter.techtarget.com/definition/edge-computing
[4] Thomas Bittman. 2018. "Digital Business Will Push Infrastructures to the Edge". Gartner, Inc. 2018. Presentation.

expected. Long latency (delay), as is typical across networks when data is shared to a cloud or data center, is undesirable for IoT devices.

Edge Computing offers an advantage over traditional data centers or cloud systems when latency, connectivity, bandwidth or cost is a concern and real-time processing is required. The purpose is to remove processing latency, because the data does not need to be sent from the edge to the core central processing system and back again. By eliminating the distance and time it takes to send data to centralized sources, data transport speed can improve, as well as the performance of the Edge Devices and applications. As the technology around machine learning and IoT advances, Edge Computing is predicted to grow.[5]

There are several related terms that have been used in place of the "Edge Computing" term. These can be seen as specific implementations or types of Edge Computing, focusing on different or particular topologies, use cases and imperatives: Fog Computing, which extends the cloud computing capability to the network edge; Multiaccess Edge Computing (MEC), which is focused on extending cloud computing to radio access networks (e.g. 5G); and Cloudlets, which is essentially "cloud in a box", bringing the cloud closer to users, enabling disconnected cloud computing.

Technical Brief

Edge Computing is a combination of hardware and software (compute, storage, networking, data analytics, data management, etc.) performing information processing at or very near the data source.

Information processing can be as simple as data filtering and storage, or as complex as advanced analytics and machine learning. Unlike cloud computing Infrastructure as a Service (IaaS), Edge architectures vary widely, based on use cases. A predominant type of Edge Computing is called Mobile Edge Computing (MEC), an architecture that utilizes the edge of the cellular network for enhanced operations.

Data is processed at the edge, and all or a portion of it is sent to the central processing or storage repository in a corporate data center, co-location facility, or IaaS cloud.[6] Most data created and processed at the edge will never leave the edge. The edge proliferates wherever IoT markets grow and will intrinsically encompass a diverse set of use cases, technologies, requirements, and command standards and rules. Instead of IoT or Edge Devices needing to constantly call home to centralized cloud infrastructure

[5] Chris Gardner and Andre Kindness, Edge Computing Will Radically Alter Your Infrastructure Strategy, Forrester Research Inc., 2018
[6] https://www.networkworld.com/article/3224893/what-is-edge-computing-and-how-it-s-changing-the-network.html

for instructions or analysis, they are given the ability to accomplish these tasks on their own. Much of the data at the edge will need to be processed at the edge, acted upon at the edge, and filtered prior to data being forwarded to the core.

Edge Computing deploys data-handling activities or other network operations away from centralized and always-connected network segments and core, and toward individual sources of data capture, such as endpoints like laptops, tablets, or smartphones. An Edge Gateway is the buffer between where Edge Computing processing is conducted and the broader overall network. Edge Computing works in various ways, and contributes to IT architectures in different capacities. It is a frequent and popular means of enhancing networks to promote efficiency and more capable security for business systems.

The Public Cloud is unrivaled in its ability to store and compute data. However, data transport to and from the cloud is limited by the speed of light and the "size" of the internet "pipes". With Edge Computing, the Data is not sent to the cloud, but acted upon at the source, or as close to the source as possible, in order to create real-time insights. Physical distance increases the latency of communication between user and cloud. Additionally, the processing speed is largely dependent on the performance of the user's device.

Edge Computing Platforms work by allowing some application processing to be performed by a small edge server positioned between the cloud and the user. Crucially, in a location physically closer to the device. This allows some of the workload to be offloaded from the cloud or user's device at a location closer to the user for processing, while at the same time speeding up applications that require a low latency response on the user's device.[7]

Edge Computing Applications, such as web browsers, are one of the most widespread applications in use by businesses and private users. The majority of users on PC and mobile devices conduct activities on the internet via web browsers. Web browsers have been pushed to more and more devices, from set-top boxes and stick PCs to digital signage displays. Yet the usability and display speed of a web browser is heavily dependent on the device's performance and lower-capability devices are strained with the tasks. New web browsers can offload part of the workload to an Edge Server and improve everything from display speeds to data downloading. Websites with complex processing, such as Java scripts (and complex Javascript), and websites with a lot of images can be loaded faster (i.e. reducing latency). When a web browser's processing is performed on the Edge Server instead of the device itself, the user's device performance mostly becomes irrelevant. This allows the device to become free from the processing stress and can perform other necessary tasks for the user.

[7] https://www.youtube.com/watch?v=RjMS15V_7nQ

Industry Use

The Fourth Industrial Revolution (Industry 4.0) is characterized by technology that blurs the line between physical and digital (cyber-physical systems). The IoT, an important part of digital infrastructure, has emerged within Industry 4.0 as one of the fundamental technologies helping to create the “factory of the future”. Industry 4.0 transforms the traditional production system into the new model called “Industrial Internet of Things” (IIoT).

The IIoT involves accessing real-time data that allows manufacturing partners and their machines to share information accurately and quickly. The goal of Industry 4.0 is to achieve low-cost production efficiencies and create more reliable operations by leveraging IoT and automation. The IoT and immersive technologies are network-effect technologies that help speed the scaling of digital business with more data, interactions and business moments, as well as better and faster decisions. However, even in a modest Industry 4.0 project, the amount of data could overwhelm existing and new systems and add a significant cost of bandwidth, data storage, and computing and data science. Processing most of that data at the edge, to filter out signals from the noise, can help focus on the information that is most important and dramatically reduce the cost of data.[8]

In the context of IIoT, the edge refers to the computing infrastructure that exists close to the sources of data, for example, industrial machines (e.g. wind turbine, magnetic resonance scanner, etc.), industrial controllers such as SCADA (Supervisory Control and Data Acquisition) systems, and time series databases aggregating data from a variety of equipment and sensors. These Edge Devices typically reside away from the centralized computing available in the cloud.[9]

Edge Computing solutions today tend to be highly customized and vertical-industry specific. But as Edge Computing matures, more standardized markets are expected to emerge. Some of these markets will be massive, and others will be much focused and niche specific. Some may have varied solutions and computing topologies addressing the same requirements in different ways. Some will mature quickly, and others will take years — or even stay highly customized. A vendor's success in one Edge Computing market may have very little to do with its prospects in another Edge Computing market. At the same time, markets that mature faster (for example, consumer-oriented markets such as the smart home) will influence the technologies in other, adjacent markets.

[8] https://iiot-world.com/connected-industry/4-0-reasons-why-edge-computing-is-relevant-for-industry-4-0/
[9] https://www.ge.com/digital/blog/what-edge-computing

The cloud still plays a critical role in enabling new levels of performance through the IIoT, where significant computing power is required to effectively manage vast data volumes from machines. But as more compute, store, and analytic capability is bundled into smaller devices that sit closer to the source of data — namely, industrial machines — Edge Computing will be instrumental in enabling edge processing to deliver on the promise of the IIoT.10

Edge Computing uses dedicated, on premise resources at the shop-floor level, rather than the remote servers that cloud computing relies on. This provides a significant increase in the rate and amount of data that manufacturers can process in real-time. There are several parts of the supply chain that benefit from the implementation of Edge Computing. The two main sectors are commonly broken up into near edge, and far edge. The near edge focuses on the automation of manufacturing processes, or the use of data analytics within a facility — to monitor for predictive maintenance needs, process optimization, and more. The far edge expands beyond the manufacturing floor and connects manufacturers with business partners and even consumers to improve proactive service for after-sales parts and services. For example, if a heavy equipment supplier sells an edge-enabled product to a construction firm, the manufacturer will be able to determine when the machine will need maintenance, what parts will need replacing, and more—before the machine actually fails. [11]

Bombadier, an aerospace company, opted to use sensors in its aircraft in 2016. That move offered an opportunity to generate more revenue by giving Bombadier real-time performance data on its engines so it can address problems proactively without grounding its aircraft to fix an issue. This eliminates the need to send engine sensor data back to a central server, either on the plane or in the cloud to determine more pressing tactical issues, such as if the engine is overheating or burning too lean.[12]

With autonomous vehicles – essentially a datacenter on wheels – Edge Computing plays a dominant role. GE Digital partner, Intel, estimates that autonomous cars, with hundreds of on-vehicle sensors, will generate 40 terabytes of data for every eight hours of driving. It is unsafe, unnecessary, and impractical to send all that data to the cloud as many of the actions and responses must be done in real-time with ultra-low latency to ensure safe operation for passengers and the public. An autonomous car sending data to the cloud for analysis and decision-making as it traverses city streets and highways would prove catastrophic. Additionally, if an autonomous car is in a snow storm and loses internet connection, it no longer matters as the processing and real-time actions do not require action information from an outside source when employing

[10] https://www.ge.com/digital/blog/what-edge-computing
[11] https://www.manufacturingtomorrow.com/article/2018/12/edge-computing-as-it-relates-to-the-manufacturing-industry/12659
[12] https://www.cisco.com/c/en/us/solutions/enterprise-networks/edge-computing.html

Edge Computing. However, the cloud is still an important part of IIoT equation. The simple fact that the car had to respond to an immediate and specific event might be valuable data when aggregated into a digital twin, and compared with the performance of other cars of its class.[13]

Mobile Edge Computing or Multi-Access Edge Computing (MEC) is a network architecture concept that enables cloud computing capabilities and an IT service environment at the edge of the cellular network. The basic idea behind MEC is that by running applications and performing related processing tasks closer to the cellular customer, network congestion is reduced and applications perform better.[14] MEC moves the computing of traffic and services from a centralized cloud to the edge of the network and closer to the customer. Instead of sending all data to a cloud for processing, the network edge analyzes, processes, and stores the data. MEC also offers cloud-computing capabilities and an IT service environment at the edge of the network.[15] Some common MEC use cases are: Data and video analytics, Location tracking services, IoT, Augmented reality, and local hosting of content, such as videos. MEC characteristics include: Proximity, Ultra-low latency, High bandwidth, and Virtualization.[16]

Canadian Government Use

There is a significant lack of documented Government of Canada (GC) initiatives and programs regarding Edge Computing. This may be due to the fact that the GC is currently grappling with the implementation of Cloud Services, and the majority of resources and efforts are occupied, as well as security concerns related to the protection of the information of Canadians.

Future in-depth interviews and research will need to be conducted with Shared Services Canada (SSC) Account Executives and with client departments in order to ascertain the level of Edge Computing capabilities that are current and planned for the GC.

[13] https://www.ge.com/digital/blog/what-edge-computing
[14] https://www.ge.com/digital/blog/what-edge-computing
[15] https://www.juniper.net/us/en/products-services/what-is/multi-access-edge-computing/
[16] https://www.juniper.net/us/en/products-services/what-is/multi-access-edge-computing/

Implication for Shared Services Canada

Value proposition

Edge Computing can enable the GC to quickly process large amounts of data where it is collected to improve service delivery timeliness and usefulness to Canadians. Potential benefits of this technology will be realized most at border crossing, airports, weather stations, etc. For example, this could be in the form of a video camera performing real-time facial expression recognition in an airport to alert border agents in case it detects suspicious attitudes or behaviours.

One of the major uses of Edge Computing is to improve network security. There is a lot of concern about security architecture in the IoT age, where more and more diverse devices are getting different kinds of access to a network. One strategy is to pursue Edge Computing to aggregate data further out, and encrypt it as it passes further into the network. If as much data as possible gets processed at the edge, rather than being sent to the cloud, there is a much lower risk of it being intercepted or tampered with. Robust Edge Computing systems will let organizations keep the bulk of IT and operational technology systems on secured parts of the networks.[17] Through Edge Computing, IT professionals hope to improve network security and enhance other network outcomes.

Edge Computing can also decrease the distance that data must travel in a network, or help with a detailed network virtualization model.[18] In some cases, latency and processing requirements can be satisfied by interactions with a central core — a cloud computing provider or an enterprise data center. However, as digital businesses seek more real-time business moments and interactions, compute and storage will need to be located physically closer to the people and things.

Sending data to a remote cloud data center for analysis gives long and unpredictable latency. The opportunity to act on the data might be gone for time-sensitive use-cases. Edge Computing gives a predictable and ultra-low latency ideal for time-critical situations, including any use-case where the operation is either mission-critical or where things are in motion. Moving vehicles, machines, parts, people or fluids are examples where delays could equal missed opportunities, high costs or security threats.

Edge Computing deployments are ideal in a variety of circumstances. One is when IoT devices have poor connectivity and it's not efficient for IoT devices to be constantly connected to a central cloud. Other use cases have to do with latency-sensitive

[17] https://iiot-world.com/connected-industry/4-0-reasons-why-edge-computing-is-relevant-for-industry-4-0/
[18] https://www.techopedia.com/definition/32472/edge-computing

processing of information. This is ideal for situations where latencies of milliseconds can be untenable, such as in financial services or manufacturing.

There are, however, several key drivers making Edge Computing a more viable reality today. The cost of computing and sensors continue to plunge, increased computing power can be executed in smaller footprint devices (such as a gateway or sensor hub), exponential data volume growth from machines and/or the environment (e.g. weather or market pricing), and improved modern machine learning and analytics.[19]

Challenges

There are five imperatives that are driving computing toward the edge: Latency/Determinism (Microseconds to sub-seconds, Consistency); Data/Bandwidth (Massive or moderate, Local value or aggregated value); Limited Autonomy (Occasionally autonomous or fully dependent, Self-organizing and discovering); Privacy/Security (Private/Sensitive or public, Regulated/Geo-Specific or unregulated); and Local Interactivity (Locally Collaborative or Locally Responsive, Horizontal or Vertical). The main challenge for organizations will be in assessing their networks, and evaluating how and when to apply Edge in order to accrue efficiencies and improvements to service delivery. Latency will be a factor (especially for mission-critical, safety-critical use cases), requiring some processing to be physically close. The challenge, again, is identifying mission-critical points on the network and giving them priority access to Edge Computing.

Edge Computing will not be a single market dominated by one vendor; there will be many types of edges and with many markets, addressed by various architectures, topologies, vendors and providers. The challenge for organizations is in knowing which part of the edge market to capitalize on. Bandwidth capabilities are expected not to grow evenly with the exponential growth in data from IoT and Edge Devices, and the cost of bandwidth will not decline exponentially either. This will be a challenge for senior management in choosing which applications and services will benefit from Edge Computing and which ones will remain with higher latencies.

There is also the risk that organizations may place too much data processing at the edge. It is easy to overwhelm the smaller processor and storage platforms that exist on or near the edge. In some cases, storage could be limited to a few gigabytes and processed using a single CPU. Power and size restrictions are what set the limits.[20]

[19] https://www.ge.com/digital/blog/what-edge-computing
[20] https://www.cisco.com/c/en/us/solutions/enterprise-networks/edge-computing.html

Another pattern is failure to integrate security from concept to production. Security is systemic to both edge and fog computing architectures and centralized processing. Security needs to span both and use mechanisms such as identity and access management. Encryption is not a nice-to-have, but rather a requirement for device safety. Imagine if a jet engine could be hacked in-flight.[21]

Additionally, while interoperability and standards are important for markets to develop, the various Edge Computing markets will have very different standards and requirements for interoperability.

Considerations

With the deployment of Edge Computing on IoT devices, SSC will need to look at the impact on end-state data center compute requirements once these devices start to process information locally. As more and more data is processed on IoT devices, a sizeable reduction of computing needs is expected in end-state data centers. SSC will also need to assess the security and privacy implications of computing data locally on IoT devices outside its data centers.

SSC should consider action plans that prepare the network for disruption among vendors by creating an exit or replacement strategy for Edge Computing systems. According to Gartner, more than 80% of non-embedded Edge Computing deployments implemented in 2018 and 2019 will be replaced with new hardware and/or software technologies by 2022.[22]

SSC should also consider ensuring that all future infrastructure strategies include Edge Computing. Different IoT requirements will drive different Edge Computing and storage requirements. At one extreme, processing and storage strictly in a central data center will be sufficient. At the other extreme, the processing and storage will need to be local to the data source. However, there will be a wide variety of choices in between, based on latency, bandwidth, connection to the core, autonomy, security, privacy, regulatory requirements, etc. It may not be valuable enough to implement Edge Computing or it may not be technically possible. SSC will need to have a plan in place to manage the future infrastructure and network needs.

SSC should consider targeting any process connected to an IoT device or sensor that can manage small amounts of data and small amounts of processing, and assess if it can benefit from Edge Computing. The ability to deal with these processes at the edge

[21] https://www.cisco.com/c/en/us/solutions/enterprise-networks/edge-computing.html

[22] Thomas Bittman, Bob Gill. The Future Shape of Edge Computing: Five Imperatives. Gartner Inc. Published: 28 June 2018 ID: G00334442.

will take care of about 90% of most required IoT-based processing, and the data and computation requirements are typically small. [23]

SSC may want to consider setting up tiers of processing to centralize where in-depth processing should occur and be stored, and where tactical local processing that does not require as much computing horsepower at the edge. The Local placement of resources becomes imperative when the conversation turns to IoT. For serious IoT implementations, analytics capacity needs to be nearby. Processing data over a vast distance just is not going to get it done when near-instantaneous reactions are expected.[24]

It is very important that SSC senior managers and employees understand that Edge computing and centralized cloud computing are complementary. Edge Computing does not compete with cloud computing directly, it complements and completes cloud computing, ensuring that the cloud remains capable of in-depth big analytics work and storage. SSC may want to collaborate with business units and clients to understand current and future edge-related infrastructure requirements, and to co-develop an Edge Computing strategy to ensure appropriate service delivery continues.

Edge Computing locations will develop organically in enterprises and enterprise-owned locations, spread out from cloud providers and carriers, and expand from the footprints of smartphones, appliances, game consoles, smart speakers, and gateways. SSC should be cognisant of these organically created Edge points and leverage them for network efficiencies and cost savings. There will be many edges and many competing ways that edge capability can be delivered for different locations and agility will be required to capitalize on them. The choice for enterprises will be specific to use cases, and the typical enterprise will build and/or leverage many forms of Edge Computing.

Most enterprises will need skills both in building and managing their own Edge Computing capabilities, and skills in leveraging third-party edge capabilities as they emerge. SSC should consider the talent it requires for the custom work required with Edge Computing, as well as aligning Edge Computing sourcing strategies (i.e. private and public) to use cases by evaluating requirements for privacy, security, regulatory requirements and maturity of public edge offerings.

[23] https://www.cisco.com/c/en/us/solutions/enterprise-networks/edge-computing.html
[24] https://searchdatacenter.techtarget.com/ehandbook/An-edge-computing-architecture-upends-the-data-center