# The Drivers and Benefits of Edge Computing

## White Paper 226

Revision 0

by Steven Carlini


## Executive summary

Internet use is trending towards bandwidth-intensive content and an increasing number of attached “things”. At the same time, mobile telecom networks and data networks are converging into a cloud computing architecture. To support needs today and tomorrow, computing power and storage is being inserted out on the network edge in order to lower data transport time and increase availability. Edge computing brings bandwidth-intensive content and latency-sensitive applications closer to the user or data source. This white paper explains the drivers of edge computing and explores the various types of edge computing available.


RATE THIS PAPER ★ ★ ★ ★ ★

APC by Schneider Electric White Papers are now part of the Schneider Electric white paper library produced by Schneider Electric’s **Data Center Science Center**
DCSC@Schneider-Electric.com

Schneider Electric

# Edge computing defined

Edge computing places data acquisition and control functions, storage of high bandwidth content, and applications closer to the end user. It is inserted into a logical end point of a network (Internet or private network), as part of a larger cloud computing architecture.

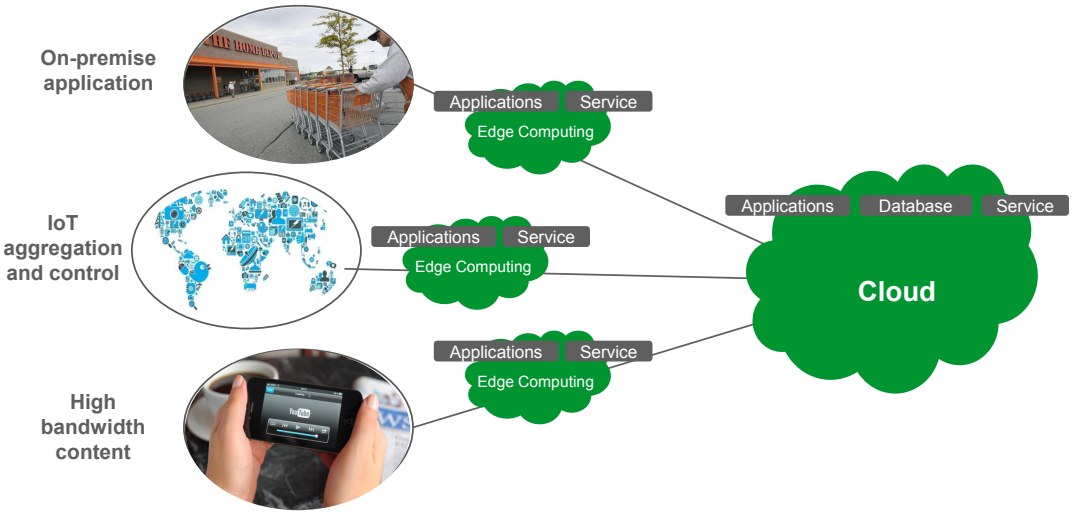



**Figure 1**

*Basic diagram of cloud computing with edge devices*

There are three primary applications of Edge Computing we will discuss in this white paper.

1. A tool to gather massive information from local "things" as an aggregation and control point.
2. A local storage and delivery provider of bandwidth-intensive content as part of a content distribution network.
3. An on-premise application and process tool to replicate cloud services and isolate the data center from the public cloud.

But before we discuss the applications and solutions let's define how networking and the Internet works

# How the Internet works

## "East-west" data transmission

Source data is converted into packets that are transmitted across the network via the networking protocol called IP (Internet Protocol). The Internet routing is handled by another protocol called BGP (Border Gateway Protocol). The Internet was designed to survive massive outages and route around problems. BGP does not factor in timing for data routing. It only looks at the number of hops between two networks trying to communicate. Those hops may be really congested or the route may take a physically long route with fewer hops instead of a very short route with multiple hops. **Figure 2** shows a map of the many long-distance hops in the U.S.[1] While BGP works really well in terms of reliability, and is a fundamental technology on which the Internet is built, it is really sub-optimal from a latency (delays, jitter, and image freezing) performance standpoint.

[1] http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p565.pdf

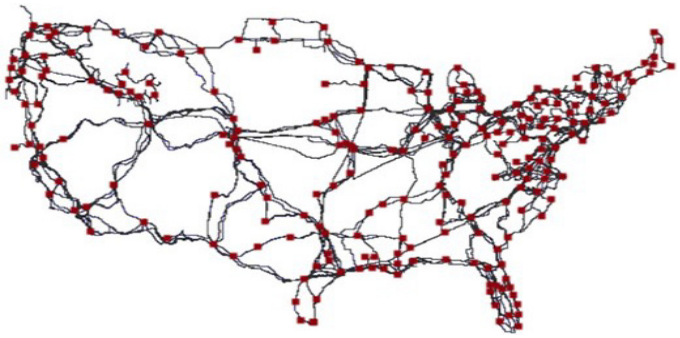


**Figure 2**

*Map of various network hops in the U.S.*

### "North-south" data transmission

As illustrated in **Figure 3**, from the inside of the typical cloud data center network to the outside, data flow goes from a physical server interface through top of rack switches or end of row switches. From each top of rack switch, data goes through an aggregation switch and the aggregation switches route data through a core switch that is the main input and output of the data center. Each one of these switches transfers data and is considered a network hop with its associated data slowdown and possibility of network congestion. If there is an oversubscription in any network layer (i.e., bandwidth is not sized for peak output), there is further potential of additional slowdowns during these periods of heavy usage.

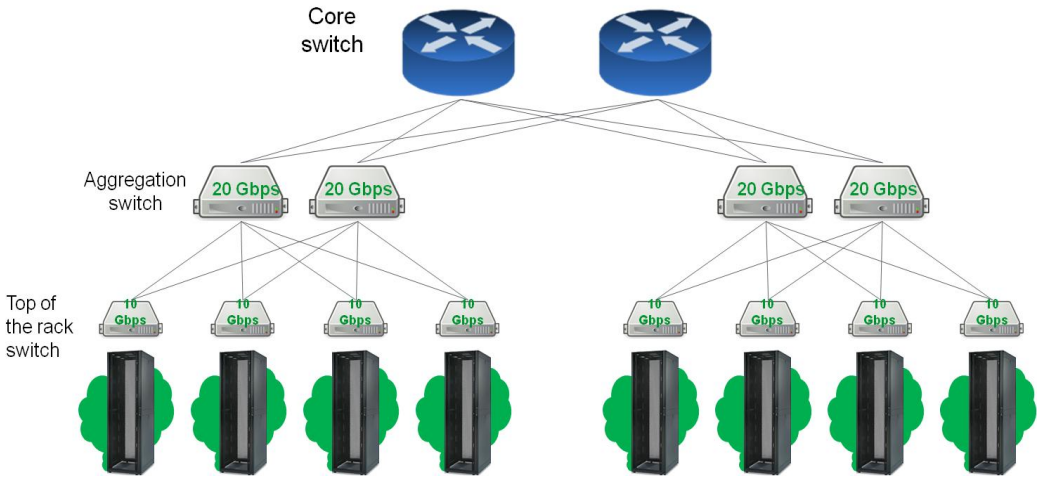



**Figure 3**

*Data center network*

## Application #1 : high band-width content distribution

Latency is the time between the moment a data packet is transmitted to the moment it reaches its destination (one way) and returns (round trip). Even though most data travels only one way, it is almost impossible to measure. That is why round-trip time from a single point is the most common latency measurement. Round trip latencies of less than 100 milliseconds (ms) are typical and less than 25 ms desired.

Bandwidth refers to the transmission speed of data on the network. Networking equipment maximum speeds are published by their manufacturers. However, the actual speed obtained in a given network is almost always lower than the peak rating. Excessive latency creates traffic jams that prevent data from filling the network to capacity. The impact of latency on network bandwidth can be temporary (lasting a few seconds) like a traffic light, or constant like a single lane bridge. The greatest probability of network congestion is from high band-width video content. As we see from **Figure 4**, VoD, 4K TV, and video streaming are the fastest growing high bandwidth applications[2].

[2] ACG Research, The value of content at the edge, 2015, p.4

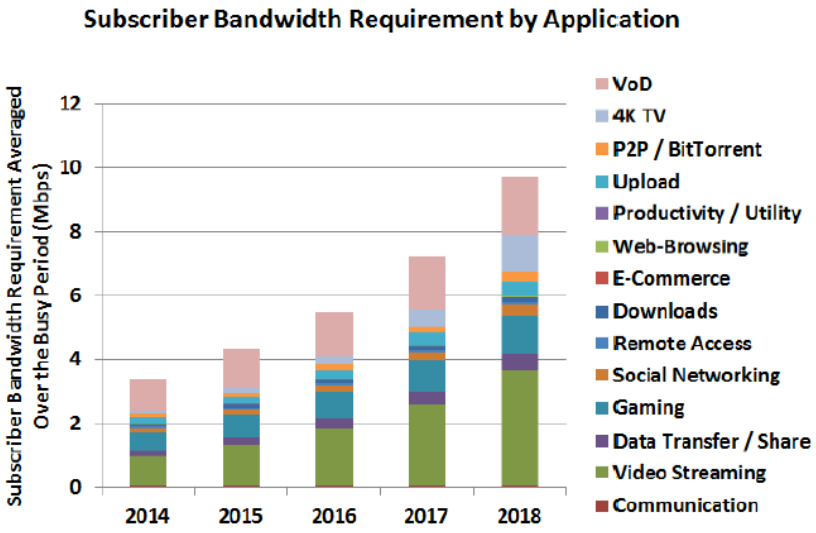



**Figure 4**

*Growth of high bandwidth applications*

In order to relieve network congestion to improve streaming of high bandwidth content now and in the future, service providers are interconnecting a system of computers on the Internet that caches the content closer to the user. This enables the content to be deployed rapidly to numerous users by duplicating the content on multiple servers and directing the content to users based on proximity. These computers caching content are an example of Edge Computing (**Figure 5**).

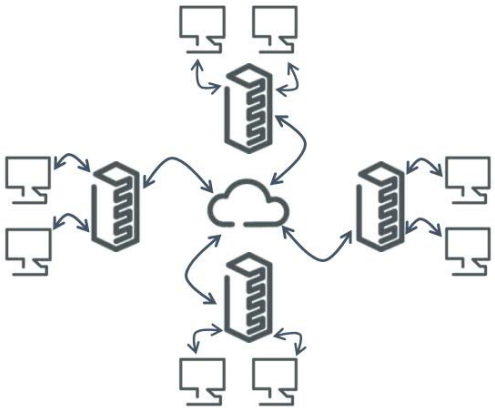


**Figure 5**

*Simple content distribution network (CDN) diagram*

## Application #2: Edge computing as IoT aggregation and control point

The technologies that will enable "smart" everything – cities, agriculture, cars, health, etc – in the future require the massive deployment of Internet of Things (IoT) sensors. An IoT sensor is defined as a non-computer node or object with an IP address that connects to the Internet.

As the price of sensors continues to decline, the number of connected IoT things will sky-rocket. Cisco estimates the IoT will consist of 50 billion devices connected to the Internet by 2020[3]. IoT can automate operations by:

- Gathering information automatically about physical assets (machines, equipment, devices, facilities, vehicles) to monitor status or behavior.
- Using that information to provide visibility and control to optimize processes and re-sources.

[3] Dave Evans, The Internet of Things: How the Next Evolution of the Internet Is Changing Everything, Cisco Internet Business Solutions Group, p. 3

Machine to Machine (M2M) refers to technologies that allow both wireless and wired systems to communicate with other devices of the same type. M2M is considered an integral part of the IoT and brings several benefits to industry and business in general as it has a wide range of applications in the Smart City.

The Industrial Internet of things (IIoT) which includes the harnessing of sensor data, machine-to-machine communication control and automation technologies generate large amounts of data and network traffic. Proprietary industrial IT systems and networking technologies are migrating to mainstream commercial IT systems that are communicating over IP (Internet Protocol) networks.

Oil & gas exploration is an example of this IIoT application. Multiple flying drones called "aerial data collection bots" examining job sites during oil exploration is generating large quantities of data in the form of high definition video. These job sites are difficult to coordinate with fleets of massive trucks, cranes, and rotary diggers. Legacy methods of traffic management have used manned helicopters for surveillance video. Self-piloted drones can photograph job sites 24 hours a day providing site managers an up-to-the-minute view of how their resources are deployed. Relying on edge computing allows the drones to transmit the data in real time and receive instructions in a timely fashion.

**Figure 6**

*Oil & gas exploration: Drones collect massive amounts of data on oil fields and use edge computing to enable real-time data transfer and movement instructions*

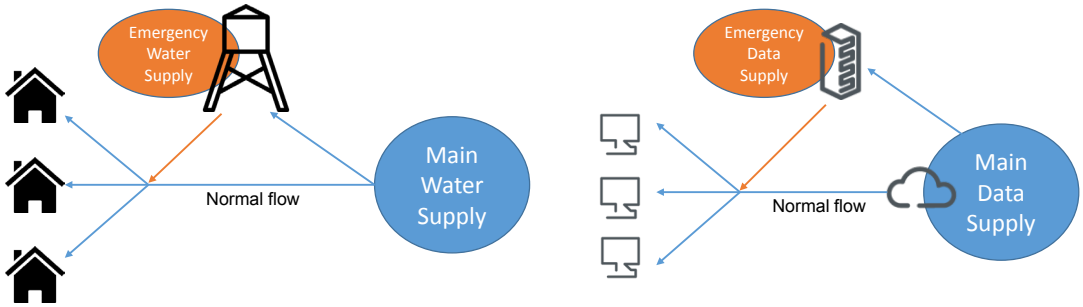
## Application #3: On-premise applications

The need to maintain or increase availability of IT and its networks is almost always a top of mind concern. Cloud computing has always been a centralized architecture. Edge computing transforms cloud computing into a more distributed computing cloud architecture. The main advantage is that any kind of disruption is limited to only one point in the network instead of the entire network. A Distributed Denial of Service DDoS attack or a long lasting power outage for example would be limited to the edge computing device and the local applications on that device as opposed to all applications running on a centralized cloud data center.

Companies that have migrated to off-premise cloud computing can take advantage of edge computing for increased redundancy and availability. Business critical applications or applications needed to operate the core functions of the business can be duplicated on-site. A way to think of this is a small town using a very large shared water supply as a main source as illustrated in **Figure 7**. Should this water supply be interrupted due to a disruption caused in the main supply or distribution network, there is an emergency water tank located in the town.

**Figure 7**

*A town water supply system as a metaphor for edge computing.*

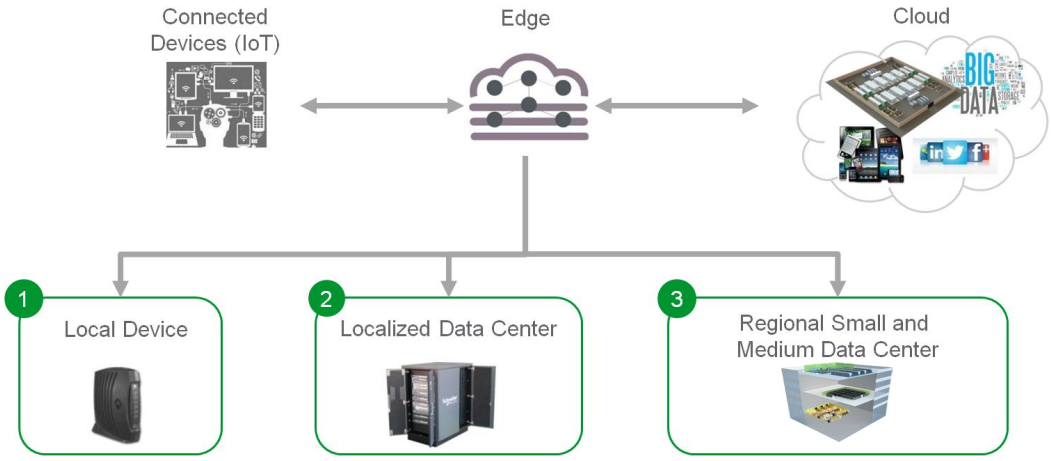



## Types of edge computing

In general, there are three types of edge computing as illustrated in **Figure 8**.

**Figure 8**

*Types of edge computing*




**Local devices:**
Devices sized to accommodate a defined and specified purpose. Deployment is "immediate" and they are suitable for home or small office applications. Running the security system for the building (Intel SOC appliance) or storing local video content on a DVR are examples. Another example is a cloud storage gateway which is a local device and is usually a network appliance or server that translates cloud storage APIs such as SOAP or REST. Cloud storage gateways enable users to integrate cloud storage into applications without moving the applications into the cloud itself.

**Localized (1-10 racks) data centers:**
These data centers provide significant processing and storage capabilities and are fast to deploy in existing environments. These data centers are often available as configure-to-order systems which are pre-engineered and then assembled on site, as shown in **Figure 9** (left). Another form of a localized data center is prefabricated micro data centers which are assembled in a factory and dropped on site, as shown in **Figure 9** (right). These single enclosure systems can be equipped in rugged enclosure types – rainproof, corrosion proof, fire proof, etc. or normal IT enclosures for an office environment. The single rack versions can leverage existing building, cooling, and power, thereby saving on CAPEX vs. having to build a new dedicated site. Installation requires picking the location in close proximity to the building power and fiber source. The multi-rack versions are more capable and flexible due to scale, but require more planning and installation time and need their own form of dedicated cooling. These 1-10 rack systems are suitable for a broad base of applications requiring low latency, and/or high bandwidth, and/or added security or availability.


**Figure 9**

*A configure-to-order (left) and a prefabricated micro data center (right) example*

**Regional data centers:**
Data centers that have more than 10 racks and are located closer to the user and data source than centralized cloud data centers are called regional data centers. Due to their scale, they will have more processing and storage capabilities than localized 1-10 rack data centers. Even if they are prefabricated they will take longer to construct than localized data centers due to the likely need for construction, permitting, and local compliance issues. They will also need dedicated power and cooling sources. Latency will be dependent on the physical proximity to the users and data as well as the number of hops in between.

## Conclusion

Edge computing can solve latency challenges and enable companies to take better advantage of opportunities leveraging a cloud computing architecture. Workloads generated from bandwidth intensive streaming video are causing network congestion and latency. Edge data centers bring bandwidth intensive content closer to the end user and latency-sensitive applications closer to the data. Computing power and storage capabilities are inserted directly on the edge of the network to lower transport time and improve availability. Types of edge computing include local devices, localized data centers, and regional data centers. The one that provides the deployment speed and capacity in-line with future IoT application demands are the localized 1-10 rack versions. These can be designed and deployed quickly and easily with either configured-to-order or prefabricated variants.

## RATE THIS PAPER ★ ★ ★ ★ ★

## About the author

**Steven Carlini** is the Director of Marketing for Data Center Solutions at Schneider Electric. He was behind some of the most innovative solutions that changed the data center landscape and architecture throughout his career. He hold a BSEE from the University of Oklahoma and a MBA in International Business from the University of Houston. He is a recognized expert in the field and a frequent speaker and panelist at data center industry events.

# Resources

**Cost Advantages of Using Single-Rack Micro Data Centers**
White Paper 223

**Practical Options for Deploying Small Server Rooms and Micro Data Centers**
White Paper 174

**Browse all white papers**
whitepapers.apc.com

**Browse all TradeOff Tools™**
tools.apc.com

## Contact us

For feedback and comments about the content of this white paper:

Data Center Science Center
dcsc@schneider-electric.com

If you are a customer and have questions specific to your data center project:

Contact your Schneider Electric representative at
www.apc.com/support/contact/index.cfm