The Edge Computing Advantage

An Industrial Internet Consortium White Paper

Version 1.0

2019-10-24

Computing at the edge has grown steadily over the past decade, driven by the need to extend the technologies used in data centers to support cloud computing closer to things in the physical world. This is a key factor in accelerating the development of the internet of things (IoT). But, as is often the case with new technology, there is an abundance of overlapping terminology that obfuscates what the fundamentals are and how they relate to one another. We aim here to demystify edge computing, discuss its benefits, explain how it works, how it is realized and the future opportunities and challenges it presents.

Business Benefits of Edge Computing

Cloud computing in data centers offers flexibility and scale to enterprises. We can extend those benefits towards "things" in the real world, towards the edge. We accrue business benefits by connecting things and isolated systems to the internet, but there are risks that should be balanced against them. Disciplined security of the whole system is needed to make it trustworthy.

The flexibility to decide where to perform computation on data improves performance and reduces costs. Sensors are generally limited in terms of what they can do, while computing in data centers induces bandwidth costs as data is transmitted to them. Computing in the edge enables collecting data from multiple sources, fusing and abstracting it as needed, and computing on it right there. For example, the data from a surveillance camera can be abstracted into geometric features and eventually a face. If there is an issue requiring immediate action (detection of a known criminal, for instance), it can be acted on then and thereby reporting the location to the police or denying them access through a turnstile. Other information can be analyzed later.

Similarly, the flexibility to decide where to store data improves performance and reduces costs. Moving data uses potentially scarce bandwidth, costs money and increases the attack surface. Computing in the edge can segment data based on compliance boundaries imposed by regulation in different jurisdictions. Data segmentation also supports disciplined security and connectivity to the internet. If data is held on the premises with no connection to the internet, it less likely to be compromised by the proverbial hacker in his parents' basement.

Many industrial facilities are relatively secure today because they are not connected to the internet. In the jargon, there is an *airgap* between the facility and the internet. Once that airgap is breached, the facility is prone to many threats and risks hitherto unconsidered. There are many ways to cross airgaps (breaking down the doors, for example), but that requires physical access. Once connected to the internet, hackers worldwide can gain control of actuators and compromise safety.

Deployment flexibility also enables separating tasks that should be executed quickly from those that need more time. For example, training a machine-learning model may be executed in the data center, while the real-time inference phase of the algorithm to score the model for specific observations can be deployed in the edge “near” the device it controls.

Similarly, executing close to where the data is generated in the physical world, rather than passing data up to a data center and back down, reduces the time lag (*latency*) and indeterminate variation *(jitter)* between receiving data and acting on it. Faster is usually better, but determinism is essential for optimizing time-critical industrial processes. Critical control processes that experience too much latency or jitter can become dangerously unstable.

Localization of data and computation can improve privacy, security, reliability, resilience and safety, which, taken together, comprise *trustworthiness*. Keeping data local keeps it private within different security boundaries as defined by the particular application.

Computing in the edge using redundant, fault-tolerant systems enables critical services to continue in mission- and life-critical applications even as nodes or links fail. When connectivity to a data center or adequate computational throughput is unavailable, tasks can be emulated in the edge, or queued until connectivity is restored.

Computing in the edge enables new applications and features, which can increase efficiency, revenue and value for the customer. For example, smart grids are already feeding distributed energy resources into the power network, reducing energy costs and even providing discounts to customers.

Computing in the edge is fundamental to distributed applications such as connected cars. A car is already a network on wheels. Edge computing enables a platoon of cars traveling at high speed to communicate, making split-second decisions to avoid accidents and to travel closely together to save road space.

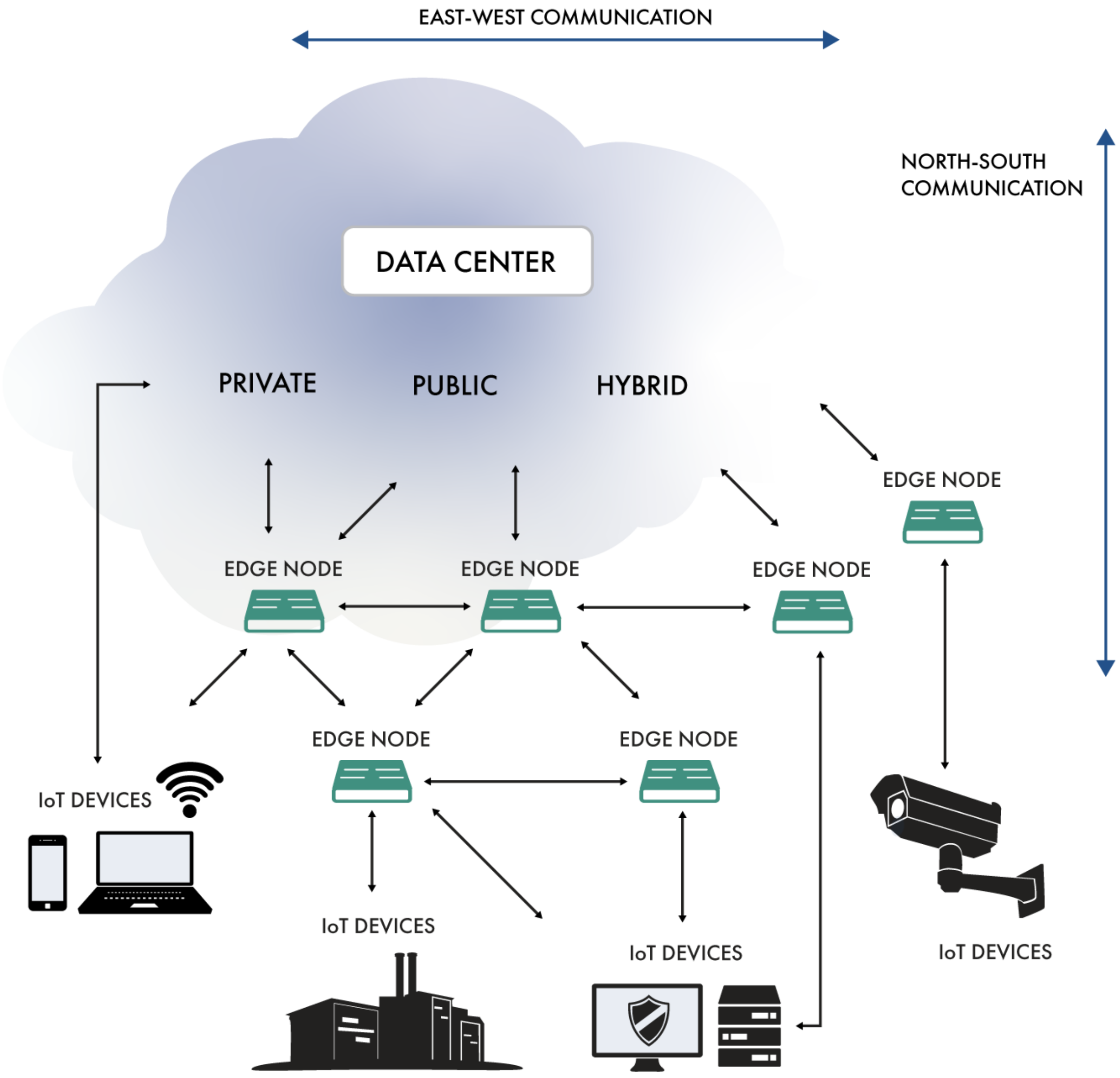
With these benefits in mind, we can now examine edge computing.

WHAT IS EDGE COMPUTING?

There is a plethora of words that have been used to describe edge computing and the technologies around it. Among them: edge, fog, edge computing, fog computing, mist, cloudlets, thing-to-cloud continuum and fog-to-cloud continuum. These names take certain perspectives and apply to specific technologies that will undoubtedly change. Here we capture edge computing as a set of core capabilities, and strictly limit the vocabulary used to describe them. We intend the term “edge computing” to embrace all of these perspectives.

The computing model is fully distributed and can support a wide range of interactions and communication paradigms. Edge computing lies between physical things in the real world as monitored and controlled by *IoT devices* (sensors and actuators), via layers of *edge nodes* to the *data center*. Operational technology production, supervision, and safety control can be implemented in edge nodes. This architecture also enables communication across subsystems.

To be able to support many vendors, legacy equipment and protocols, as well as avoid vendor lock-in, we need to be able to assemble hardware and software from multiple providers into a system that will interoperate seamlessly. In legacy systems, this may include multiple controllers using multiple protocols and gateways from different vendors, all of which need to be connected.



The topology of the network enables IoT systems to make use of layers of edge nodes and gateways to interconnect IoT devices and connected subsystems with various types of data centers. The cloud is the “highest-order” resource, and is usually implemented in large, protected data centers. It may be public, private or a hybrid to process and store data for specific vertical applications. Edge nodes perform local processing and storage operations.

Edge nodes can work alongside traditional IoT network elements such as routers, gateways and firewalls, or subsume those functions into a merged device that has computation and storage capabilities. North-south data communication links connect between layers, while east-west communication interconnects nodes on similar layers. Some of the nodes are public and shared, while others are private; some a combination. Processing and storage functions operate at whichever node and layer most efficiently meets the applications' requirements, which is how we reduce costs. Taken together, edge computing comprises:

- computation and storage resources, with layers of edge nodes between the data center and things in the physical world,
- peer-to-peer networking, such as security cameras communicating about objects within their scope, a platoon of connected vehicles or a fleet of wind turbines,
- distributed computation across IoT devices, edge nodes and the data center,
- distributed data storage for preserving data in IoT devices, edge nodes and the data center and
- distributed security functions, like data segmentation, authentication and encryption.

With edge computing, data, storage and computing are distributed throughout layers of edge nodes from IoT devices to the data center, distributing the economies of scale of the cloud throughout the IoT system.

To respond to new applications or varying workloads, we need to be able to add or remove resources quickly by reconfiguring the system to execute on more or fewer edge nodes. This *elasticity* supports, for example, first responder teams when computation and connectivity needs fluctuate in an emergency. It also enables *scale* so that small customers with modest needs can exist alongside web-scale operators with millions of users and provide this service at lower cost.

To be able to support many configurations, we need to be able to allow multiple independent entities to share the common infrastructure without interfering with each other or causing security and privacy problems. This is called *multitenancy*. Without this, a smart city, for example, would have to build out parallel networks and edge nodes for every government agency, logistics company, transportation carrier, mobile operator or smart grid in the city. That would be cost prohibitive and clutter every street corner.

HOW IS EDGE COMPUTING REALIZED?

Rarely does any digitalization initiative start from scratch; there are existing computation and storage resources that need to be integrated and that means we need *interoperability*. Interoperable edge computing is realized by agreeing on architectural and protocol standards, and by separating the logical (what the user perceives) from the physical (the actual computing machinery) to allow a single resource to be shared between several users. This is called *virtualization*, and it helps cuts overall costs for all users, because they need not buy that resource for themselves. It also enables scaling so that when one resource is inadequate, more can be added seamlessly, improving performance.

In addition to sharing physical resources, this integration enables a single software application to serve multiple customers, where you share the same application software with your own private dataset. This is called *software multitenancy*.

Multitenancy is also used for sharing the resources in the network that supports the edge. Its owner (the landlord) allows multiple independent entities (tenants such as companies, agencies or individual users) to share the common infrastructure without interfering with each other or causing security or privacy problems.

Managing the diversity found in deployments of edge nodes, virtualization and multitenancy is a gargantuan task, especially when resource availability and demand change rapidly. It cannot possibly be done manually. Automatic configuration, coordination and management of the elements comprising the edge is a must. This is called *orchestration*, which takes place throughout the lifetime of the system to include initial configuration, run-time resource allocation, fault recovery and managing abnormal situations such as a denial of service attack.

Taken together, integration, virtualization, multitenancy and orchestration provide *manageability* to ensure efficient configuration and operation of resources in the edge and enable edge-computing users to scale elastically. We need to be able to change system structure as the volumes and requirements change. Edge computing enables distribution of computation workloads to any part of the infrastructure: control systems, sensors and actuators, the data center, and anywhere in-between. This enables distributed data management for both data at rest and data in motion, defining where and what data is to be stored, in what form, and for how long. Edge computing supports diverse data governance models, including quality, discovery, segmentation, usability, privacy and security.

That last is key. When you connect edge nodes or IoT devices to the internet, you must secure them. We know how to make IoT devices secure and what needs to be done, but not all devices are secure. An IoT device may have been built without security in mind, or be a part of a legacy system. Devices are being hacked every day and the problem will only worsen. Moreover, edge devices may be deployed in unsupervised and remote locations prone to attacks ranging from physical tampering to modifications of the code. Current security mechanisms cannot scale to the rapid growth of the IoT. Many IoT devices don’t have the computing power or energy reserves to perform strong security processing on their own. Edge computing can extend security improvements to edge nodes and allow the management of complex security policies on behalf of less capable IoT devices.

We have to know what every connected IoT device *is*, and we have to know it is correctly configured by properly authorized staff using authenticated code and data. What you don’t know can hurt you. Once authenticated, we can trust the data coming from them. Without that, an IoT device could send false data, and with false data you cannot trust any part of your system, from how you act on that data to training algorithms for machine control. Edge computing enables disciplined security at all levels of an IoT solution from sensors and actuators to the data center, and this creates trustworthiness.

OPPORTUNITIES

The edge computing model provides opportunities, and there are some trends in the industry that are likely to continue to influence it.

In the future, it will be less expensive to create significant business and social value. As bottlenecks are discovered, fixes will be applied. In the 1960s and 1970s, memory was slow and computing power relatively faster. That drove the industry to improve access to memory and then the obverse was true. This cycle will continue. There will be pressure on network management costs, necessitating edge manageability improvements and ceaseless pressure on total lifecycle cost of ownership from purchase price, to installation, management, energy usage, updates and decommissioning. Techniques currently in the news, such as artificial intelligence, were known decades ago but there was neither enough computing power nor data capacity to enable those techniques to be applied in reasonable time.

We expect these pressures to continue to move computation and data around. Computations will be executed in the “best” location; data will be stored where it can be accessed most efficiently. Computation will be drawn towards the data (*data gravity*) at the optimal location for it. And that will continue too. This suggests needs for architectural standards and innovation in scaling, depending on the specific application. Some edge nodes will scale down, running on inexpensive, low-power, compact hobbyist-class computers; others will scale up, running on computer clusters with equivalent power to hundreds of rack-mounted servers, depending on where the data pulls them.

The centers of data gravity depend to some degree on required response times—how quickly the data needs to be acted upon. For example, training a machine-learning model requires a good deal of data and time. This can be done in the data center, and the trained model deployed closer to the IoT devices, possibly in the physical plant that needs to be controlled and optimized. This reduces latency and improves user response. Consequently, we expect to see artificial intelligence, machine learning and deep learning to become more widespread and provide new insights and intelligence to optimize decision-making.

Data gravity also applies to services as they migrate closer to where they need to be, or further away for broader data spans. This will enable new business models and new services.

Entities need to share information; they also need to keep it private. Distributed ledger technologies, such as blockchain, can be used as authentication providers. This means that *more* data can be shared because the provider has more confidence that the shared data will be restricted to the preselected groups. This could be used to provide attestation of edge elements and software, and track the provenance and completeness of the critical edge-hosted data.

Reliability can be improved by adding redundant edge nodes, but this increases costs. Depending on the application, different patterns will emerge for certain types of applications, taking full advantages of edge computing architectures and cutting costs.

CHALLENGES

There are always challenges, many of which we have mentioned above. But there are more.

When the internet of things first appeared on the hype curve, there were ever-increasing estimates of the number of IoT devices that could be connected to the internet (and ever-increasing time scales). We face several scaling challenges to realize the potential for IoT. First, enormous numbers of devices, and the connections between them, increase complexity. Provisioning edge nodes between the IoT devices and the cloud can reduce it. Second, clusters of edge nodes, as in data centers, require large amounts of energy and cooling. Consequently, some data centers are located near hydro-electric power generation in cooler climes. But how can the energy consumption issue be addressed when the edge nodes need to be close to the IoT devices?

Conversely, when IoT devices (and their associated edge nodes) need to be physically located in remote, environmentally challenging places, how can they be maintained, provisioned and secured? For example, an oil field in a remote desert location in a politically unstable country will need reliable power (generators from well by-products?), cooling, physical security and an internet connection. How are we to do that? More generally, how are we to address new applications of the technology in various environments?

Demand management is a challenge. When a natural disaster occurs in a smart city, for example, the demand curve shifts dramatically owing to a huge number of IoT devices reporting changes in data. Moreover, when infrastructure is damaged, regardless of the location, we need to stand up new infrastructure rapidly. How are such changes in demand to be handled?

Disconnection is an issue. Because a connection cannot be guaranteed all the way from IoT devices to the data center all the time, we need to store information and execute critical control algorithms locally until the connection is available again. This requires more processing power and storage and that affects scalability. This also implies that you can’t just take legacy hardware, without such processing capability or storage, and make it work in edge computing scenarios without some newly deployed hardware.

Complexity also bedevils software, especially as we move data between verticals. Each vertical tends to develop its own application interfaces and standards, each exposing different features. Moreover, significant domain knowledge is needed to integrate different domains. An engineer conversant with oil and gas, for example, may know nothing about energy generated from diesel and solar to supply the edge nodes at the oil wells. To facilitate software development, common platform infrastructures and standardized application programming interfaces are needed. We also need to abstract data further and find ways to converse between different data abstraction models. *Semantic interoperability* is a general name for techniques that address this challenge. How can we develop generalized patterns that can be applied in many verticals?

Skilled technical experts for edge computing and its applications are needed. That means relationships with academics and universities to develop training materials and certifications for individuals. How else will we find people who can train the next generation?

The Industrial Internet Consortium's mission is to accelerate the development of the internet of things. This paper addressed a key challenge we have encountered: market confusion. We have groups of member representatives working on defining edge computing (of which this paper is just a part) and innovating to address the challenges outlined above. By reducing confusion, especially over terminology, we help to develop the marketplace and achieve our mission.

What Does This All Mean?

There are many evident technical and business advantages that drive edge computing to become a keystone of IoT's evolution. In response to the high-compute, high-performance and low-latency requirements of IoT, edge computing moves critical data, storage and network functions "closer" to the things, co-located in the IoT device or physical plant. When properly designed and administered, this enables otherwise unattainable performance, efficiency, reduced operational costs and opens opportunities for new IoT applications.

Edge computing provides many of the qualities that will be required in IoT settings in the coming years. Edge computing is inherently scalable and elastic—it flexibly supports multiple communications models, layers and software programmability features. Edge computing security supports trustworthiness of the entire IoT solution from sensors and actuators to the cloud.

We will continue to see blurred lines from the edge to the data center, as cloud-computing and edge-computing architectural models merge and emerge. Here we have striven to demystify these technologies, identify some benefits, define edge computing and how it is realized, and then look ahead at future opportunities.

For these reasons, we believe edge computing will have a profound effect on the future of IoT. That's what this means. We are only at the beginning of a multi-decadal revolution.

Next Steps

There is a lot to learn. To get started, here are some references you can check out:

- Introduction to Edge Computing,
- Industrial Internet Reference Architecture,
- Industrial Internet Security Framework,
- Industrial Internet Connectivity Framework,
- Industrial IoT Analytics Framework,
- OpenFog Reference Architecture and the
- Industrial Internet Vocabulary.

Note that these documents are also located at our global Resource Hub, where you can find many other pieces of information.

You can also work with us to learn more and address this exciting challenge. Please join us.

The Industrial Internet Consortium is the world's leading organization transforming business and society by accelerating the Industrial Internet of Things. Our mission is to deliver a trustworthy Industrial Internet of Things in which the world's systems and devices are securely connected and controlled to deliver transformational outcomes. Founded March 2014, the Industrial Internet Consortium catalyzes and coordinates the priorities and enabling technologies of the Industrial Internet. The Industrial Internet Consortium is a program of the Object Management Group® (OMG®).

Visit www.iiconsortium.org.

IIC members gain experience they could never have as a non-member. They experience member meetings unlike any local meet-up groups. Here are some key benefits of membership:

- ***Networking***: Make the connections; find the needed expertise.
- ***Information & News***: A fast pass to newsworthy industry developments.
- ***Competitive edge***: Stay ahead of the competition or take advantage of changes and developments that might otherwise have passed you by.
- ***Create a market***: Join a collective voice supporting a single mission; create the disruption in the market and develop the business opportunities.
- ***Success***: Members are building businesses and dedicating their professional lives to IIoT. They want to be successful, and they want others to succeed.
- ***Professional development***: Grow your career, meet mentors and mentees, career prospects.
- ***Solve important problems*** and help your partners and customers.
- ***Events***: Capitalize on opportunities for continuous exposure to industry developments.

AUTHORS AND LEGAL NOTICE

This document is a work product of the Marketing Working Group.

Authors: The following persons contributed substantial written content to this document: Chuck Byers (IIC staff), Ron Zahavi (Microsoft), John K. Zao (National Chiao Tung University).

Contributors: The following persons contributed valuable ideas and feedback that significantly improved the content and quality of this document: Evan Birkhead (IIC staff), K. Eric Harper (ABB), Brett Murphy (RTI), Mitch Tseng (Tseng Infoserv).

Technical Editor: Stephen Mellor (IIC staff) oversaw the process of organizing the contributions of the above Authors and Contributors into an integrated document.