EDJX

THE NEW ROE

RETURN ON EDGE™

CONTENTS

THE NEW ROE: RETURN ON EDGE

Edge computing is an emerging new computing paradigm that some view as the evolution of cloud computing and some view as its antithesis, but is perhaps best described as a natural complement to it. In traditional cloud computing, resources are concentrated into hyperscale data centers and "rented" to customers on an as-needed basis. In edge computing, these resources—compute, storage, and network—are widely dispersed to locations close to where users, devices, and things are. This dichotomy is similar to centralized automobile manufacturing plants vs. distributed automobile dealerships, or hyperscale oil refineries vs. small, distributed gas stations.

Beyond the architectural distinction, there are numerous business benefits that generate the new RoE: "Return on Edge." These include improved economics for broad classes of applications and infrastructure, such as reducing backhaul data transport costs incurred to bring data from the edge to a central location. They include enhanced performance across many metrics, such as reduced latency and thus better response time for distributed applications. Ultimately, these lead to an enhanced customer, employee, and partner experience, in turn leading to greater productivity and customer satisfaction, as well as higher revenues.

DIGITAL TECHNOLOGY TRENDS & REQUIREMENTS

Today, more and more processes, resources, products, services, customer interactions, and innovation are digital, whether largely virtual, such as ecommerce, or largely physical, such as connected autonomous vehicles. In any case, there are a number of inescapable trends that are transforming the world that we experience, as well as requirements concerning the computers, storage, and networks behind these experiences.

- Cyber-physical Systems
- Endpoint Explosion
- Interactive and Real-Time
- Device Mobility
- Application Mobility
- Dynamic, Transient, Ephemeral
- Resource Constraints and Management
- Small and Lightweight
- Data Intensity
- Autonomy and Intelligence
- Fusion and Aggregation
- Integration and Optimization
- Trust, Privacy, Accuracy, Reliability, and Security
- Energy Use and Sustainability
- Diversity and Heterogeneity
- Shared Infrastructure, Dynamic Allocation
- Global Reach

CYBER-PHYSICAL SYSTEMS

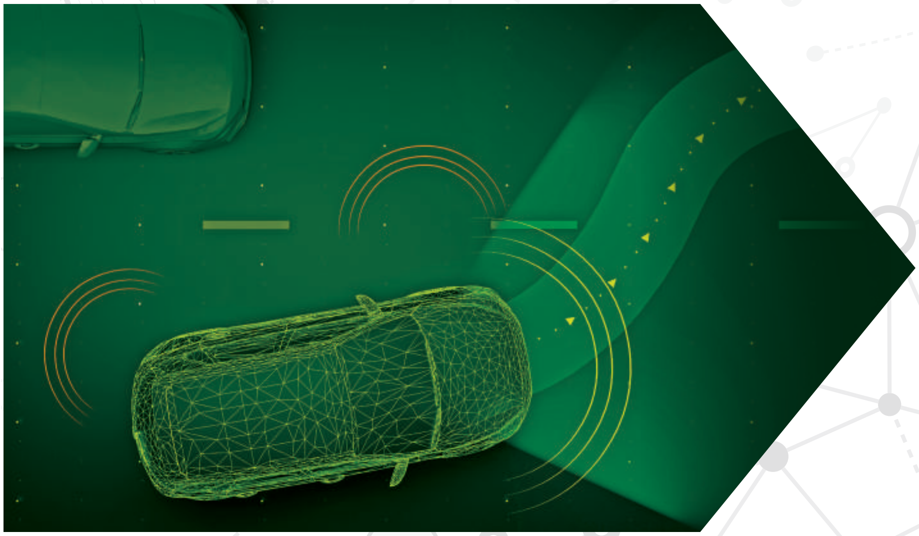
The first generations of the Internet dealt with web pages, email and other messaging, and ecommerce, existing purely in a virtual world. They were about transmission and sharing of data and information across digital networks. Now, the focus is shifting to hybrids of physical and virtual, generally referred to as cyber-physical systems, IoT, or "digical," or using terms specific to particular domains, such as "bricks-and-clicks" or "online-to-offline (O2O)" for digical retail.

ENDPOINT EXPLOSION

The number of connected cyber-physical devices and things is exploding. There are already tens of billions; by some estimates, it won't be long until there are a trillion. This is happening because there are not only billions of user devices such as smartphones, tablets, and laptops, but many more things that are becoming smart and connected: light bulbs, outlets, smoke detectors, smart speakers, drones, traffic lights, tires, and so on. Some devices, such as video cameras, act as sensors; some, such as traffic lights, act as actuators; others act as both.

INTERACTIVE AND REAL-TIME

The applications enabled by these devices have increasingly stringent latency and response time requirements. For example, an application that controls a connected autonomous vehicle has only milliseconds to capture images, LIDAR, or radar sensor data, receive telemetry from roadside infrastructure regarding pedestrians or bicyclists or pets, and determine which way to swerve to avoid casualties. A robot in a manufacturing plant similarly has only milliseconds to stop swinging its arm to prevent a worker from injury.



DEVICE MOBILITY

Many of these devices and things, such as light bulbs, stay in one place. Increasingly, however, they are mobile: automobiles, smart bikes, delivery drones, in-factory mobile robots / pallets, etc. From a connectivity perspective, this means not only that the devices must be wirelessly connected, but that strategies are in place behind the scenes to ensure that the devices "home" properly to nearby compute and storage resources. In addition, not only are endpoints mobile, but the infrastructure that interacts with endpoints may be mobile as well, e.g., an edge node that resides on a bus.

APPLICATION MOBILITY

Because users and things are mobile, and low latency is often key, the applications and data that they need to access must move around as well. Various technologies have been emerging to enable those applications to move, such as virtual machines, (VMs) containers, and container orchestration. The latest such technology is a W3C standard called Web Assembly. It uses a high-performance execution run time that can run on virtually any platform and whose apps can be written in virtually any language. It loosely resembles Java or JavaScript for browsers, but broadens the execution environment to include servers, including those at the edge. As a result, the same code can run in an endpoint device, at the edge, or in the cloud, and move among them as needed.

DYNAMIC, TRANSIENT, EPHEMERAL

In turn, application components needed by the endpoint devices and things need to migrate their location appropriately to compute and storage resources that are "close" to the endpoints needing their services. As users, things, applications, and even infrastructure become mobile, they become like ships passing in the night—only in contact with each other for brief amounts of time. This means that the connections between user devices and things and the infrastructure they need to use may be short-lived, and applications and relevant data for those applications must migrate to one of the edge locations near to the device or thing. The connection must be established, the application and its data acquired, loaded, and launched, and then safely saved and migrated to the next location in exceedingly brief periods of time. There are many application scenarios where a static configuration won't work. Picture an electric utility repair crew that shows up after a storm, or firefighters who show up at a wildfire. The team members may have never met each other, and they may be in a remote area. No one has built up fixed infrastructure and manually and securely configured team-member access ahead of time, yet it all must work.

RESOURCE CONSTRAINTS AND MANAGEMENT

Problems such as resource allocation and contention now must be solved as workloads ebb and flow in any given geographic area, and at any given processing node. Nodes may run out of resources. If this happens, they can either refuse new tasks or workloads; suspend existing ones; or turn to nearby neighbors for help, possibly migrating existing in-process tasks to the new node. Any of these approaches leads to additional challenges. For example, the first approach can be dangerous if a vehicle is asking for help to prevent a collision, or in the near future, when a drone is asking for help to navigate through a congested air route.

Suspending existing tasks may lead to additional work to repeat intermediate processing steps or even liability. Turning to nearby neighbors requires determining the best alternate node to use, determining the best route to get there, and moving applications, connections, microservices, and/or data to that node.

SMALL AND LIGHTWEIGHT

This degree of dynamic reconfiguration means that various elements must be lightweight to be able to move from node to node, and to be able to be rapidly invoked and turned up when needed. In other words, rather than a monolithic application with a million lines of code, the application needs to be made up of small, on-demand microservices, each of which can be rapidly (re-)started and rapidly migrated across nodes, and each of which only uses processing resources when it is running.

DATA INTENSITY

While it is true that some sensors are extremely low bandwidth, such as, say, a smart water meter that generates a few bits every 15 minutes, there are also broad application categories requiring high-bandwidth sensors such as video capture devices, such as for security, entertainment, inspection and quality control, or collaboration.

AUTONOMY AND INTELLIGENCE

Sometimes, these sensors operate in a closed loop with nearby actuators. They use machine learning and artificial intelligence to exhibit complex and advanced behaviors. For example, a thermostat can operate in a closed loop with an HVAC unit, and can learn patterns of an inhabitant's desired temperature though the day. The intermittent use of intense processing power, however, often means that it is more resource efficient to move such processing power up one layer from the endpoint sensors and actuators to a nearby edge processing location.

FUSION AND AGGREGATION

Often, the rationale for moving processing from the endpoint to the edge isn't resource utilization, but functionality. Imagine a number of video sensors that are being used to track inventory, in part to prevent unwanted shrinkage, i.e., theft. It is only when the feeds are fused together by a central agent that movements into and out of the field of view of any given camera can be stitched together and understood. Sometimes this fusion can be based on federation and aggregation, where no particular sensor, such as a vibration detector, can decide if there is a problem; it's only when, say, five or more are triggered that a threshold is crossed and an alarm generated.

INTEGRATION AND OPTIMIZATION

By fusing together multiple sensors and actuators, a system can be globally optimized. As an example, by integrating roadway sensors, vehicle-to-roadside communications, traffic lights, roadside infrastructure such as digital signage, and intelligent traffic flow systems, cars can be metered onto highways and instructed to slow down as they approach congestion or the scene of an accident, increasing safety, maximizing throughput, maintaining maximal safe speed, and reducing carbon footprint.

TRUST, PRIVACY, ACCURACY, RELIABILITY, AND SECURITY

A user, customer, thing, or application needs to trust that a computation node accurately and reliably executes its tasks, e.g., processing a payment transaction, authenticating a user, telling a car that it can speed up, etc. Similarly, a user, customer, thing, or application needs to trust that its data remains private. In fact, data sovereignty laws and data privacy regulations often apply. And, computation nodes need to be secure against various types of attacks, such as distributed denial of service. These are all true whether the computations and data reside in the cloud, at the edge, or somewhere in between.

ENERGY USE AND SUSTAINABILITY

The growing realization that climate change is a clear and present danger deeply underscores the importance of minimizing energy use and enhancing sustainability. Multiple approaches can be used: lower power, using idling technologies and new generation technologies, such as 5G cellular wireless networking; shared infrastructure, which runs multiple applications and services on the same physical platform but only uses as much power as a single application; and ambient cooling, dissipating heat in the nearby environment rather than having to dedicate a substantial amount of energy to cool concentrated computing resources in a hyperscale facility. Additional strategies include energy scavenging, i.e., harvesting sufficient energy through things like microscopic vibrations or temperature fluctuations in the environment, and the use of renewable energy sources, such as solar and wind.

DIVERSITY AND HETEROGENEITY

Devices and things are diverse in terms of their users, uses, manufacturers, operating systems, applications, sensor and actuator types, mobility, etc. However, they may all need to connect together. For example, collision avoidance must work not just between Ford trucks and other Ford trucks, but Ford trucks and Ford cars, Fords and Chevys, Fords and Trek bicycles, and Fords and users carrying Android devices. Cars must interoperate not only with each other, but with roadside infrastructure such as sensors and video cameras, parking space allocation systems, congestion pricing toll software, global traffic flow optimizers, entertainment systems, accident reporting, emergency alerting, and so forth.

SHARED INFRASTRUCTURE, DYNAMIC ALLOCATION

It is economically infeasible and physically unrealistic for each user, business, or application to have globally dispersed private, dedicated infrastructure built to peak to support potential peak workload levels. Instead, a secure, reliable, shared infrastructure is required that can nearly instantaneously allocate processing and storage capacity and execute software applications and services together with the data that they need on any node in the architecture, as needed.

GLOBAL REACH

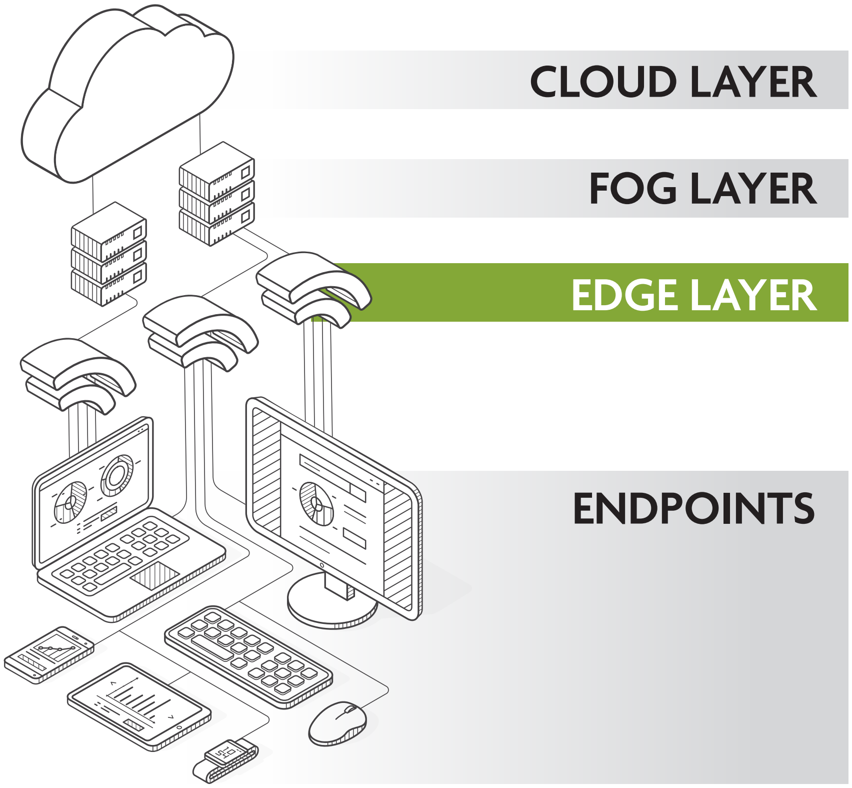
Some edge applications can be relatively constrained. For example, a farm with irrigation sensors and actuators can be self-contained, with an edge node acquiring data, running pre-installed software, and watering when necessary. For many applications, however, a global (shared) infrastructure is required, because users—using smartphones, tablets, wearables, or implantables—and their devices and things, may need that infrastructure as they travel anywhere in the world. As a result, supporting infrastructure for these users and applications must match their global reach.

ARCHITECTURE OPTIONS

Given these trends in and requirements for digital technology, there are numerous implications on the digital architecture that businesses, governments, and militaries must deploy to serve their customers, employees, partners, and citizens.

There are four generic layers where functionality—i.e, computing, storage, security, and applications—can reside: the endpoint, the edge, the fog, or the cloud.

Endpoints are devices or systems that interact with the real world, containing sensors and/or actuators, some local processing and storage, and a connection to other elements, either on a peer-to-peer or hierarchical basis. A smartphone, e.g., has sensors such as accelerometers to detect gravity and movement, microphones to detect and transduce sound waves, still/video cameras to detect and transduce light waves, and touchscreens for touch input. It also has actuators such as speakers to create sound waves, a display to create light patterns, and an element to create vibrations. Other endpoints are as diverse as tablets, smart TVs, light bulbs, fire sprinklers, lawn sprinklers, automobiles, tanks, jet engines, drones, 3D printers, ingestible pills, connected pacemakers, assembly robots, etc.



The **edge** is the layer that is not an endpoint itself, but connects directly to endpoints. Edge nodes can be primarily about computing and storage, or primarily for networking, as with 5G base stations and residential routers and gateways, or perform all these tasks.

At the other extreme, there is the **cloud**, which may be a private cloud in an enterprise data center or colocation facility, or the hyperscale data centers that are the home for large public cloud providers. Endpoints are ubiquitous, the edge is globally dispersed to be near endpoints, and the cloud is concentrated and consolidated. While there are tens of billions of endpoints, and there will soon be hundreds of billions of them, there are only dozens, or perhaps hundreds of hyperscale cloud data centers for any given cloud provider.

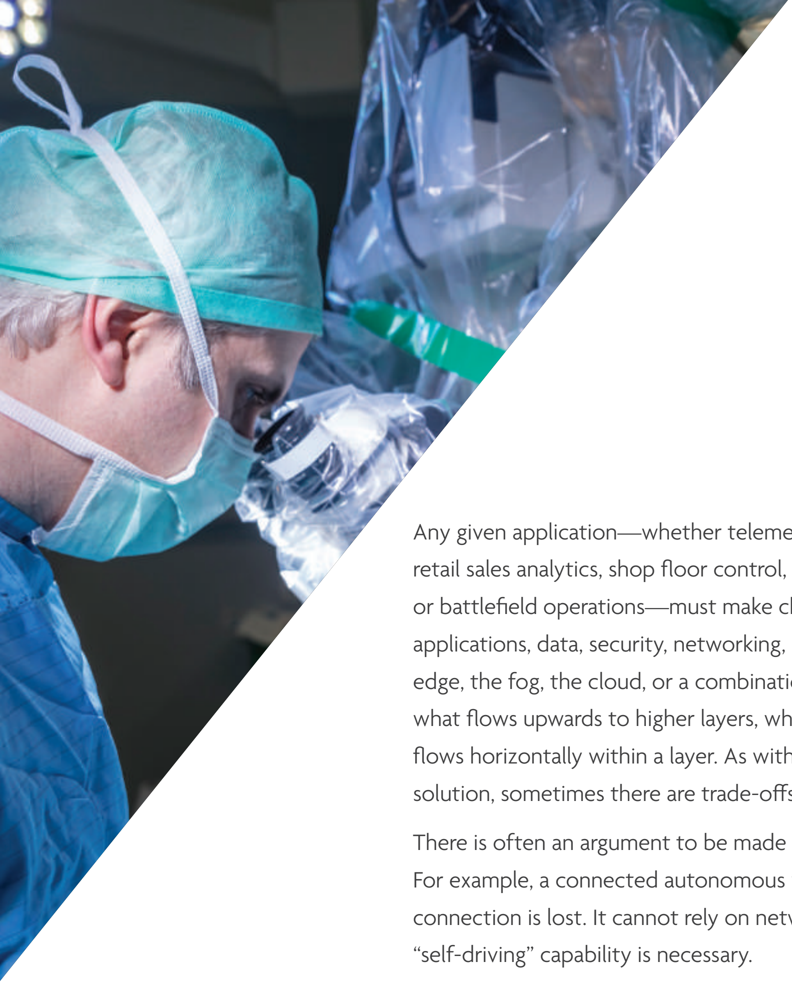
The **fog** sits between the cloud and the edge, and comprises a gradient of intermediate facilities that may be used for regional monitoring, processing, and control. These facilities are more dispersed than the hyperscale data centers of the cloud, but less dispersed than the edge.

THREE TRAFFIC PATTERNS

The first path occurs when data from lower layers travels to higher layers for fusion and aggregation, the same way that lower layer nerves that help visual and auditory inputs are fused and aggregated in the various areas of the brain. In the same way, data from temperature sensors, humidity sensors, barometric pressure sensors, and wind direction sensors may be aggregated and fused to determine that a tropical storm is approaching.

The second path occurs from higher to lower. It may be data that travels this path, or control impulses. A content delivery network distributes data, such as a movie, from an origin server to edge nodes. A closed-loop control network can distribute control impulses to multiple actuators to take action. For example, a vehicle collision system can avoid an accident by controlling steering and braking appropriately.

The third path involves peer-to-peer data transfer, coordination, resource sharing, and control. A school of fish evading a predator is engaging in peer-to-peer communication and coordination without a central authority. Peer-to-peer approaches for resource sharing enable overloaded resources to transfer work to more lightly loaded ones. Peer-to-peer approaches for data include various types of replication, including distributed ledger mechanisms such as blockchain. And peer-to-peer approaches for control include mechanisms such as voting and alternate route determination.



Any given application—whether telemedicine and remote robotic surgery, retail sales analytics, shop floor control, synchronous mirroring and data backup, or battlefield operations—must make choices as to where to locate services, applications, data, security, networking, and other functions: the endpoint, the edge, the fog, the cloud, or a combination of all of these. It must also determine what flows upwards to higher layers, what flows down to lower layers, and what flows horizontally within a layer. As with most things, sometimes there is an obvious solution, sometimes there are trade-offs to be balanced.

There is often an argument to be made for pure standalone endpoint functionality. For example, a connected autonomous vehicle must still operate safely if a connection is lost. It cannot rely on network availability or reliability; literal "self-driving" capability is necessary.

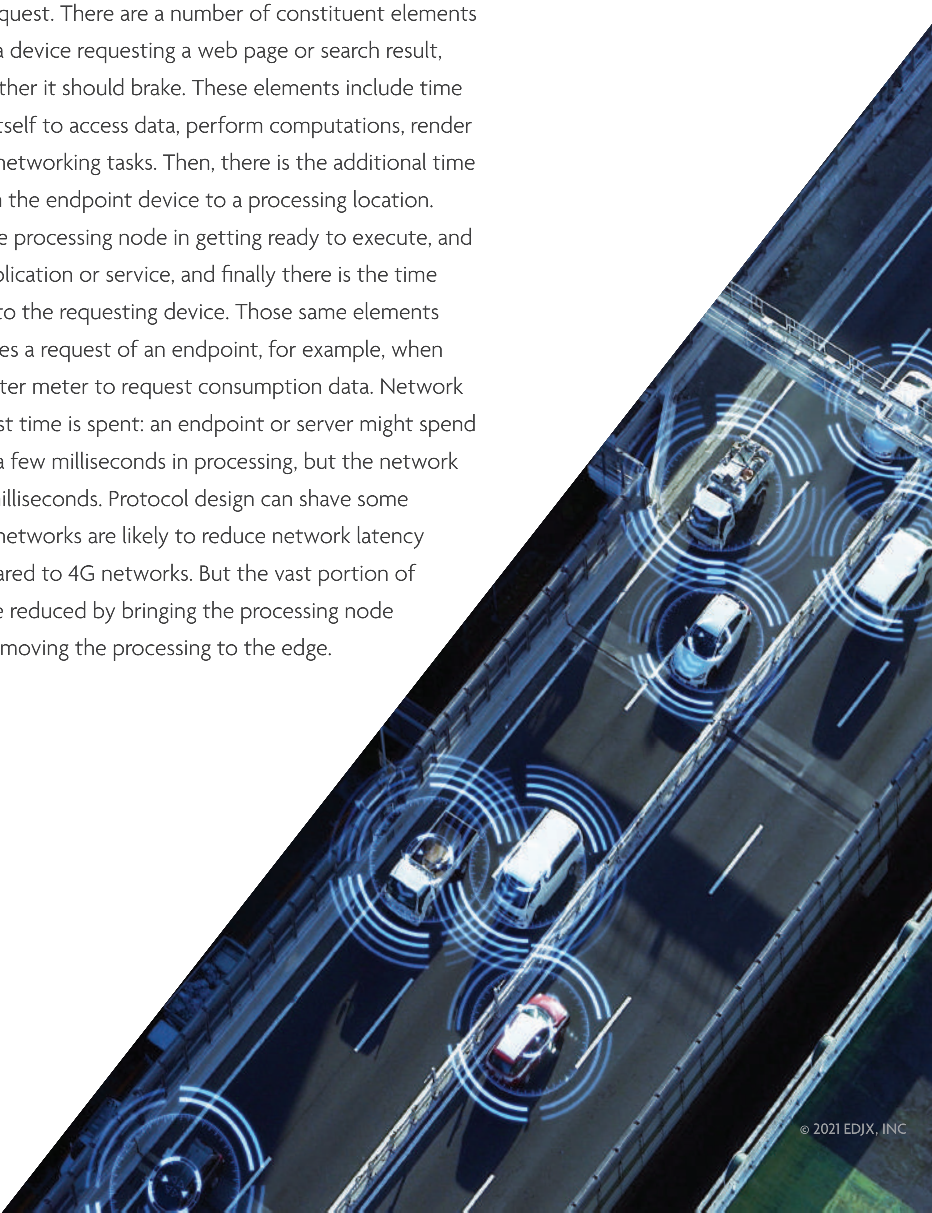
There are also arguments to be made for distributing functionality across multiple layers. For example, if the endpoint experiences a failure or complete loss, it may be important to preserve data from that endpoint. Whether it is cloud back-up of contacts, or remote data back-up of enterprise storage via mirroring and replication, there are benefits to maintaining a copy of data remotely. The necessary distance depends on the type of potential failure. To protect against dropping a phone on the floor, a backup to a storage device in the same room will work just fine. To protect against a flood or hurricane, backing the data up to a nearby metro in the region is likely to work. To protect against a major regional or national disaster, remote backup to another region or continent may be necessary.

EDGE BENEFITS AND APPLICATIONS

Given the trends and requirements identified earlier, however, one of the most typical scenarios is to provide most processing and storage at the edge. The Return on Edge includes many financial and non-financial benefits.

LOWER LATENCY AND RESPONSE TIME

The edge benefits that many are familiar with are reduced latency, i.e., one-way lag time, and faster response time, i.e., round-trip delay between making a request and receiving a response to that request. There are a number of constituent elements driving response time, say for a device requesting a web page or search result, or a thing like a car asking whether it should brake. These elements include time spent in the endpoint device itself to access data, perform computations, render display graphics, and conduct networking tasks. Then, there is the additional time needed to transport data from the endpoint device to a processing location. Then, there is time spent at the processing node in getting ready to execute, and then actually executing, an application or service, and finally there is the time spent to carry the result back to the requesting device. Those same elements are present when a server makes a request of an endpoint, for example, when the water company pings a water meter to request consumption data. Network latency is often where the most time is spent: an endpoint or server might spend hundreds of microseconds or a few milliseconds in processing, but the network might consume hundreds of milliseconds. Protocol design can shave some milliseconds. For example, 5G networks are likely to reduce network latency by 10 to 20 milliseconds compared to 4G networks. But the vast portion of the remaining time can only be reduced by bringing the processing node closer to the endpoint, i.e., by moving the processing to the edge.



DECISION-MAKING AND CONTROL

A fundamental framework for decision-making, management, control, combat, and competition is the OODA loop: Observe-Orient-Decide-Act. The steps in this loop are to collect data from the environment using a variety of sensors; orient, i.e., filter and process the data; make a decision as to a course of action based on the filtered data; act in accordance with that data, using actuators; and repeat the cycle (or smaller cycles, such as collecting more data after initial filtering) in perpetuity. The idea is that—all other things being equal—the participant who is able to execute the loop more quickly will have an advantage, whether in flying warplanes in combat, which was the origin of the idea; marshalling business resources in marketplace competition; or taking corrective action in a self-driving vehicle. The reduced latency that edge processing offers is not just an abstract performance metric; it provides real benefits in improved decision-making and execution.

FASTER LEARNING

Even if insights from the “Observe” and “Orient” phases are not immediately used to make and act upon decisions, faster learning can occur in the observe phase or in an observe and orient mini-loop. For example, a surveillance camera can capture an image, detect an anomaly, and zoom in to capture a higher resolution image. Or machine learning can occur more quickly. Or machine learning algorithms can use higher volumes of more fine-grained raw data, because this data does not need to be compressed or sampled to be sent over a network.

ENHANCED USER / CUSTOMER EXPERIENCE

Faster response times also mean a better experience for users and customers. A variety of studies in a variety of contexts and industries illustrate the range of usability benefits. For example, search portals experience more click-throughs when results are presented more quickly. The usability of a function such as "autocomplete" for partially spelled words or partially entered phrases depends on it being faster than users can type to completion. Controlling a car, remote mining equipment, or a drone—whether in real life or in a video game—requires rapid response times. Augmented reality for equipment maintenance or surgery requires an accurate overlay based on the conditions now, not hundreds of milliseconds ago.

ENHANCED REVENUES AND PRODUCTIVITY

When customers' experiences are better, they buy more. For example, in ecommerce, faster response times mean that customers can view more items, find items that meet their needs, and thus are more likely to have increased purchase variety and purchase sizes and are more likely to conclude their purchase. Conversely, slower response times may cause the customer to "balk or renege"—in other words, either never start the purchase process because they may believe the site or app is down, or stop their efforts because the experience is painful. Even employees and partners can be more productive, because they spend more time working and less time waiting. A few hundred milliseconds may not seem like much, but when you consider how frequently a delay is incurred and across how many employees, it quickly becomes apparent that benefits of the edge can run into millions of dollars.

REDUCED TRANSPORT COSTS FROM ENDPOINTS

Frozen concentrated orange juice is a real-world example of the benefits of processing at the edge to reduce the cost of transport. A little bit of processing cost is incurred to extract the orange juice concentrate at the edge, i.e., near the orange grove, so as to avoid the cost of transporting the low value water, which is based on volume and weight. The total reduction in weight and volume can be three- or four-fold. Similarly, processing raw data at the edge to extract the juiciest high value insights rather than incurring the cost of transporting watered-down, low-value data is almost always worthwhile. Even though we like to think of network connections as being distance-insensitive, the truth is that as network usage becomes nontrivial, network connection costs do as well. Now, however, compression can reduce the cost to one-millionth or one-billionth as much. For example, consider the difference between streaming video surveillance data, at several megapixels per frame (which, for a 4K image is 4,096 x 2,160 = 8,847,360 pixels x 24 bits / pixel = 212,336,640 bits), vs. using machine learning and image processing at the edge to reduce that down to 1 bit: namely, is there a robbery, fire, or other emergency at the site or not? If we send 24 or 30 frames per second, with multiple cameras, over a long period of time, the contrast becomes even more dramatic. The exact savings depend on the application and infrastructure: for example, networks can drive economic impacts based on depreciating capital investments or cloud data ingress / egress operating expenses.

REDUCED DISTRIBUTION COSTS TO ENDPOINTS

The edge not only reduces the cost of transport from endpoints to central aggregation points but also reduces the cost of outbound distribution. A content delivery network (CDN) is an example of this. By streaming or distributing content or any other kind of data from a nearby edge to endpoints rather than from a distant origin to those endpoints, substantial reductions in data transport needs can be achieved.

REDUCED CAPACITY REQUIREMENTS, HIGHER UTILIZATION, LOWER RESOURCE COST

Performing processing in your personal, dedicated endpoint vs. executing in a shared, multi-tenant edge (or cloud) is like the difference between building a guest room in your house and getting a hotel room for the night. The guest room is rarely, if ever, used, and is a stranded asset at worst or an underperforming capital investment at best. Conversely, hotels typically get high utilization out of each room (pandemics excluded) by dynamically allocating room resources to the travelers who need them. In an edge computing context, consider an application such as visual license plate recognition for toll payment and stolen car tracking, or image analysis from a medical diagnostic lab. In each case multiple image sensors generate video or imaging. There are enormous amounts of data generated at high speeds, requiring high bandwidth. The cameras could each have a massive amount of processing power, but since they are intermittently used, there are resource utilization benefits and thus preferred economics by sharing that processing power. Such statistical multiplexing increases utilization of resources and thereby reduces the cost of any usage, in the same way that a hotel room night costs much less than the building costs of a guest room. In short, reduced capacity requirements for a given amount of work mean higher utilization of the resources that are employed. In turn, because a customer or user pays not only for resources used but also ends up paying for resources that are wasted but not used that get incorporated into the cost structure, this results in a lower cost for the resources that are used.

REDUCED ENERGY AND RESOURCE USE

Whether we consider the energy used to transport data across long distances, the energy used to power underutilized compute resources, or the energy used to manufacture data networking and data processing equipment, utilizing multi-tenant resources at the edge to process data locally can dramatically reduce energy use. If recycled / renewed computing resources are used at the edge, resource use is decreased in two ways—through statistical multiplexing of resources and through reductions in manufacturing and raw materials. Moreover, energy use is further reduced in terms of the energy needed to manufacture those saved resources. And, by using virtual power, i.e., "software-defined power," energy efficiency can be even greater.

INCREASED RELIABILITY

Concentrated resources are more likely to be simultaneously impacted if they are affected by a natural disaster such as a flood, earthquake, hurricane, or tornado. For example, lightning strikes at major infrastructure and software-as-a-service cloud providers have disrupted services to entire regions, because the entire data center went down. Conversely, dispersed resources are more likely to have at least some resources that survive.

FINER LOCATION GRANULARITY

Although GPS is the standard for location-based services, it is notoriously inaccurate. Its degree of error of several meters is insufficient for, say, ensuring that a connected autonomous vehicle stays in its own lane. It's not just for transportation, because such fine-grained location data can be used to determine exact locations of people for augmented reality, determine the exact location of a gunshot in a city, etc.

These different benefit categories will apply to a greater or lesser extent depending on the exact application, industry segment, and context. For example, consider the variety of ways in which just response time improvements drive business results:

- In ecommerce, better response time means that shoppers can view more web pages, are more likely to find what they are looking for, and thus are more likely to complete an order. This results in more frequent purchases, larger purchase sizes, and therefore higher revenues.
- In healthcare, better response time means that a remote-controlled robotic surgeon is more likely to remove a tumor and less likely to accidentally slice through an artery.
- In manufacturing and processing industries, the shorter the cycle time needed to become aware of a potential safety issue such as a worker in the path of a robot arm, the more likely it is that the robot arm can be stopped.
- In warehousing and logistics, inventory control and shrinkage prevention require near-real-time analysis and validation of material movement, not after-the-fact analysis.
- In transportation, better response time means that a vehicle is more likely to avoid a collision, that traffic light timing can reduce carbon footprint, or that emergency vehicles can be given priority.
- Perhaps the most extreme example of the criticality of response time is under battlefield conditions. Imagine a self-configuring array of sensors that are parachuted onto a battlefield. The enemy begins firing, and the sensors are used to triangulate the enemy position. There is the need to fuse the information from the sensors—no sensor can triangulate by itself. There probably isn't much of a wide-area network connection, but the forces must be able to identify the enemy's location and counter-attack as quickly as possible.
- A complementary example is emergency relief. The battlefield may now be an earthquake-ravaged or storm-ravaged area, and instead of tracking enemy fire, one needs to identify the location of injured people, aftershocks, or fires.
- Regardless of industry, horizontal applications such as augmented reality (AR) require extremely short response times, not just to enhance the user experience, but to make the application feasible at all. Healthcare is using AR to help guide surgeons through complex spaces based on prior MRI or CT scans. Manufacturing uses AR for equipment repair. Commerce uses AR for enhanced experiences and virtual pop-up stores.

RETURN ON EDGE SUMMARY

There is a powerful Return on Edge, making the edge a critical element of virtually any information technology and digital transformation strategy—helping to meet customer needs, fulfill missions, and achieve business results.

ABOUT THE AUTHOR

Joe Weinman is a global keynote speaker and author or editor of books on cloud economics, mobile communications, and digital strategy, including *Cloudonomics: The Business Value of Cloud Computing* (Wiley, 2012); *Digital Disciplines: Attaining Market Leadership via the Cloud, Big Data, Social, Mobile and the Internet of Things* (Wiley CIO, 2015); *Fog and Fogonomics: Challenges and Practices of Fog Computing, Communication, Networking, Strategy, and Economics* (Wiley Information and Communication Technology Series, 2020); and has contributed chapters to other books published by Springer and MIT Press. He was the contributing editor of the Cloud Economics column of *IEEE Cloud Computing* magazine for its entire publication life of 2014-2108, and has been published in the print and/or online editions of publications such as *Harvard Business Review*, *CIO*, *Forbes*, *InformationWeek*, *The New York Times*, and various academic journals and conference proceedings.

He has held a variety of executive positions at AT&T Bell Labs, AT&T Business, HP, and was most recently SVP, Cloud Services and Strategy at Telx. He has a BS in Computer Science from Cornell University; an MS in Computer Science from UW-Madison; and has completed in-person and online executive education at the International Institute for Management Development (Lausanne), MIT Sloan School of Management, and Harvard Business School. He has been awarded 26 U.S. and international patents in a variety of fields, including voice and data communications and cloud computing. He sits on the advisory boards of multiple companies, including EDJX.

Visit our website at:

EDJX.IO

EDJX is an edge computing platform that makes it easy to write edge and IoT applications using serverless computing, accelerate content delivery, increase the responsiveness of edge applications, and secure edge data at the source. EDJX helps businesses handle the explosive demand for data processing to serve real-world edge computing applications, including industrial IoT, artificial intelligence, augmented reality, and robotics. Led by cloud industry veterans John Cowan and James Thomason, EDJX is a privately held company based in Raleigh, NC. Visit EDJX and follow EDJX on LinkedIn and Twitter.