



Monetize the Edge

Drive business success through distributed edge computing



The business impact of edge computing

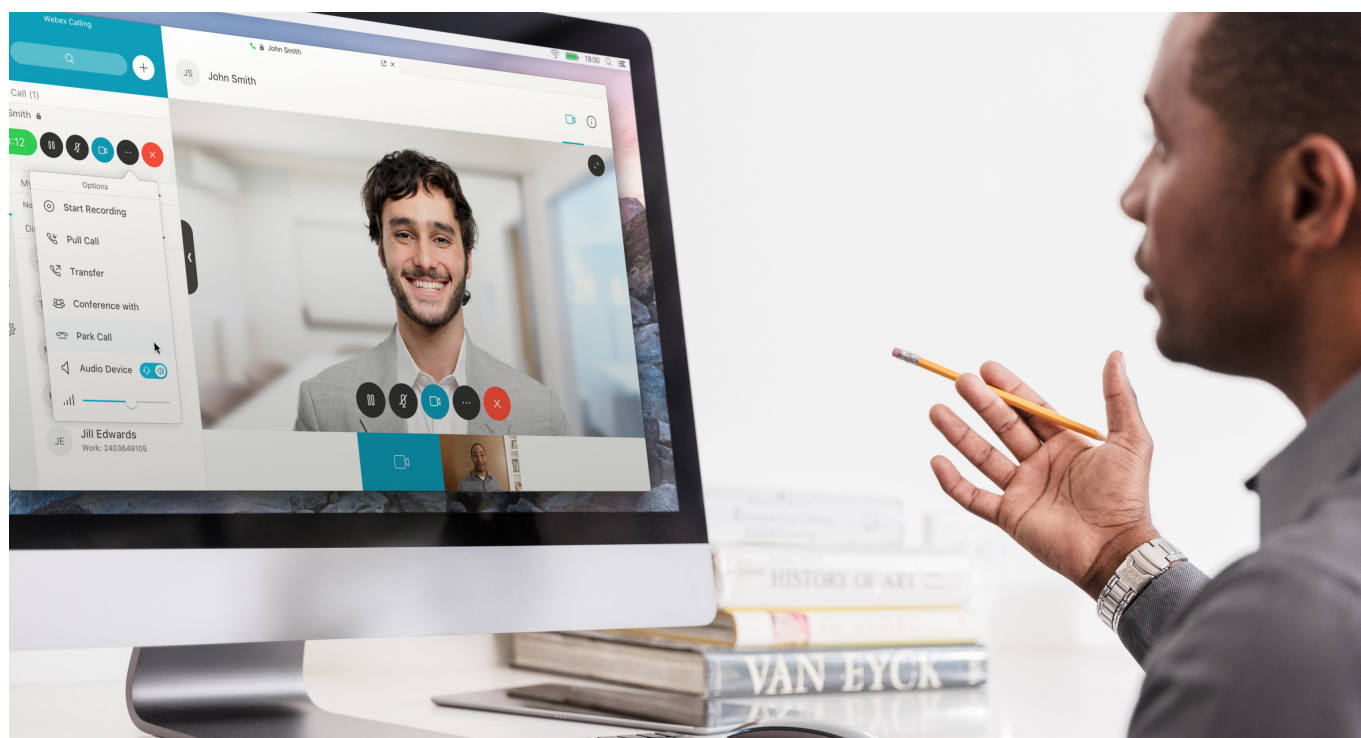
Edge computing is becoming mainstream. For future business success, service providers need to deliver high-quality customer experiences and monetize their network infrastructures through simplified access for partners and application and content developers. They also must make edge resources such as computing and storage capacity easily consumable for new vertically targeted services. Creating a platform for low latency can help achieve three important goals:

- Grow top-line revenue by creating unprecedented user experiences.
- Cut costs by reducing operational complexity in delivering high value services.
- Reduce risks by using methods, technologies, and partners you can trust.

Preparing for 5G

To prepare for 5G, service providers need to update their infrastructure and embrace a distributed cloud model that includes multi-access edge computing (MEC). These improvements need to support:

- Infrastructure such as host virtualized network functions such as 5G CUPS, virtualized central unit (CU), and distributed unit (DU) in the context of a cloud-based radio access network (RAN).
- Operator-branded services such as consumer-oriented services that are provided under the operator's brand.
- Business services that include services offered to enterprises, the public sector, public cloud providers, and IoT service providers.
- Private radio for enterprise that supports low latency and security.



Changing dynamics

The changing dynamics of edge computing lead to:

- **User experience (UX) transformation.** Edge computing reduces latency from hundreds of milliseconds to tens of milliseconds. This low latency will vastly improve the customer experience and make new services possible.
- **Simplification.** Simplifying operations can reduce the time of site delivery to first service availability and ongoing operational costs. Simplification also accelerates adoption and lowers deployment risks.
- **Monetization.** Edge APIs expedite the time to service and edge resource monetization by tapping into the innovation and reach of partners, application, and content developers.

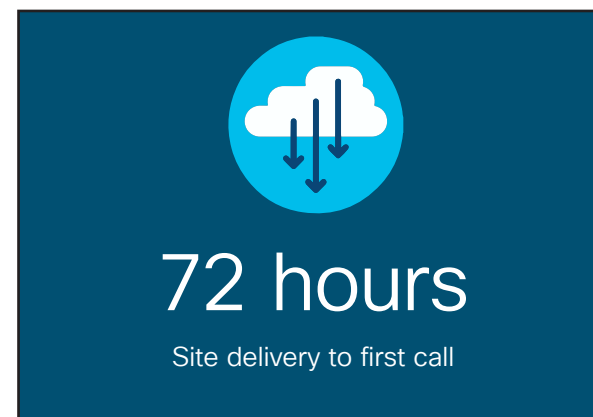
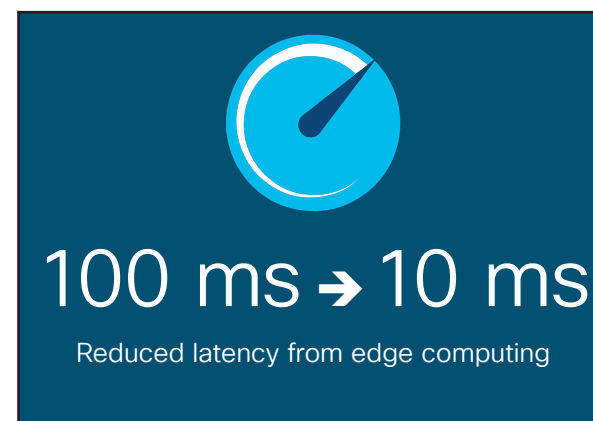
Improving your time to service can lead to more revenue opportunities for you.

User experience transformation

A significant 5G differentiator will be a better user experience from the reduced latency that results from edge offload. Consider this question: would you pay more for a faster data plane or for a high-fidelity, real-time gaming experience that has advanced augmented reality (AR) functionality?

Simplification

Network complexity leads to high costs and a slow time-to-market. Simplifying your operations through full-lifecycle automation and a reduced number of touch points becomes critical. The problem is compounded by highly distributed cloud workloads, which can have a direct impact on total cost of ownership. A consistent automation tooling for integration, validation, and operation ensures a faster service rollout.



Real examples

Rakuten

A 'Push button' deployment of 4000 edge sites co-located at "Group Centers" (owned by Nippon Telegraph and Telephone - NTT) and 50,000 Radio sites in Rakuten.

Vodafone

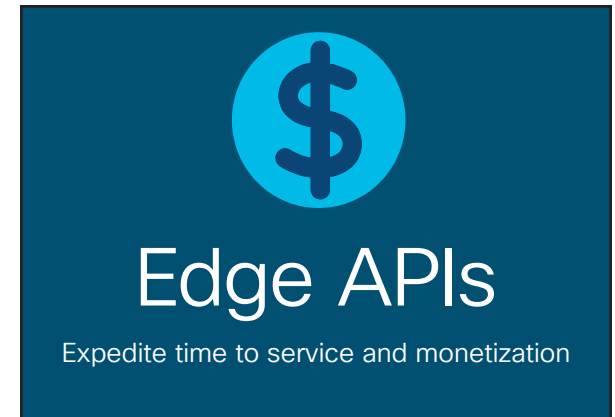
72 hours from 'Site delivery to first call' for virtualized EPC rollout.

Deutsche Telekom

Zero-touch OpenStack software release upgrade and patching in less than 24 hours.

Monetization using edge APIs

It's important to expose network resources to global reach partners and application and content developers using the same set of edge APIs. These APIs use standard cloud consumption models, which industrialize the underlying infrastructure.



Edge APIs

Expedite time to service and monetization



Why latency matters

Today, the typical round-trip time (RTT) latency between an end-user and cloud or content distribution center (CDN) services is around 150 ms. This time includes:

- Local access over-the-air latency between 20–30 ms (real network LTE scheduler delay).
- IP transport contributing 10 ms.
- Unpredictable Internet latency ranging between 20 and greater 200 ms (worst case).

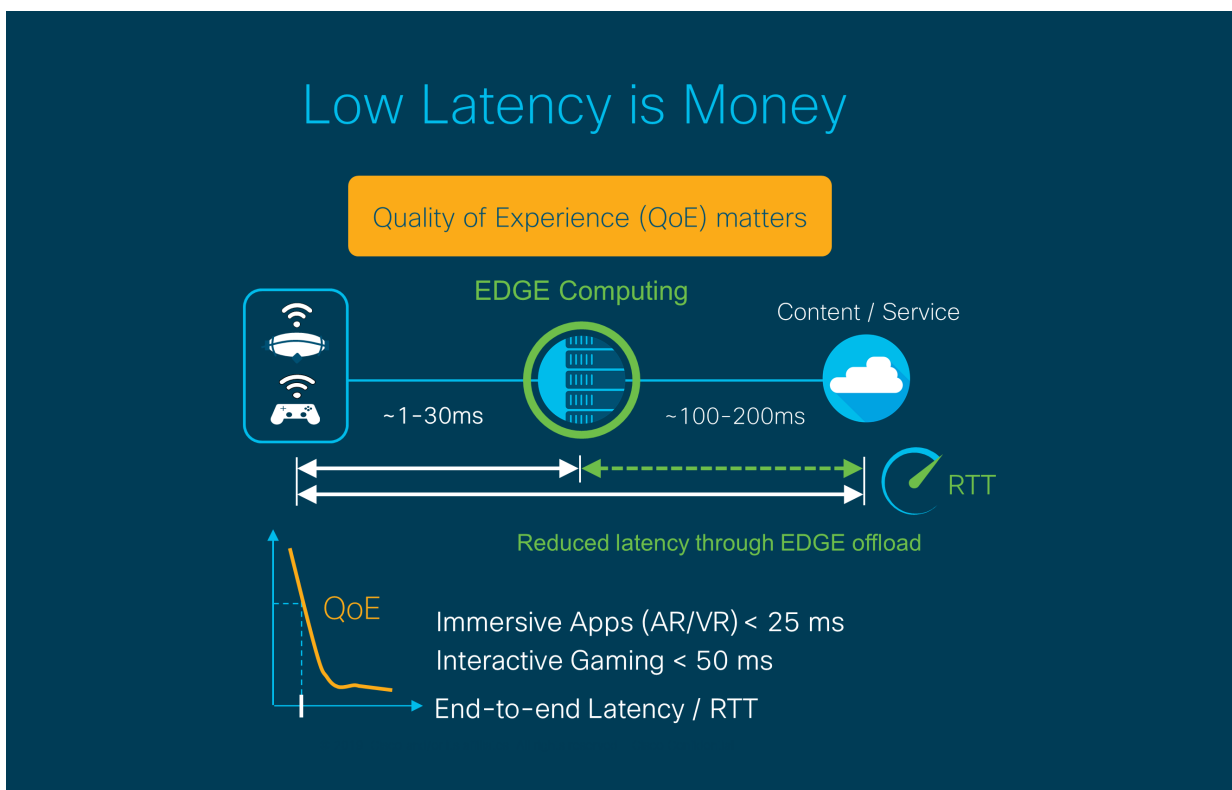
New cloud applications such as interactive gaming and immersive augmented and virtual reality suffer from higher latency times.

Meeting new requirements

To meet customer application demands, you need to determine where the service edge needs to be and where the low-latency services edge exists in your network. You should start with the initial service requirements, which are typically 10–30 ms. Examining service needs is better than a bottom-up-approach which could inadvertently result in thousands of difficult to manage edge cloudlets put as close as possible to cell sites, whether or not there's a real need for them.

Note that proximity does not equal low latency.

A properly designed transport network supports low latency through well-known quality of service mechanisms, which reduce queuing delay. It also supports new technologies such as low-latency segment routing, which allows the concentration of functions at central office locations. This approach provides an optimal balance between app requirements and economic efficiency.



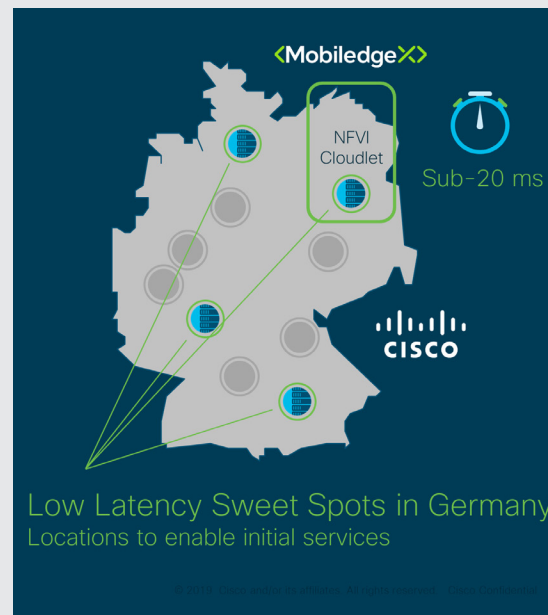
Real example

Together with MobileEdgeX, Cisco has built the first public mobile edge network. It provides sub-30 ms latency anywhere in Germany. The network is already live in six sweet spot locations and will be extended over time. It will open up lower latency windows down to 1 ms as the new 5G capabilities are deployed into the network.

The network functions virtualization infrastructure (NFVI) cloudlets in this foundation are enabled by Cisco NFVI solution that is powered with Cisco Virtual Infrastructure Manager (CVIM). The embedded function-as-a-service (FaaS) system, dynamically places application backends as close to the requesting mobile applications as possible. It uses device- and platform-independent software development kits that connect users automatically to the nearest edge location.

Application developers can use edge APIs to provide a better user experience:

- The trusted location API finds the closest cloudlet.
- The pose/face detection API offloads backend functions to the nearest graphics processing unit (GPU) location.
- The dynamic grouping API provides lightweight gaming overlay meshes.
- The predictive quality of experience API provides known latencies.



Creating a successful low-latency setup

The ingredients for a successful low-latency setup start with a distributed telco cloud that is fully automated. It should feature a loosely coupled, modular architecture where the integration between layers is driven by open APIs and data models. The modularity of this architecture allows compliance to any disaggregation requirement and the ability to scale from hundreds to thousands of locations.

The distributed telco cloud is composed of two components:

- A carrier-class cloud platform that is designed for horizontal scale allowing to put any workload anywhere.
- An underlying transport network that is fully programmable to provide predictable low-latency paths.

The carrier-class cloud platform

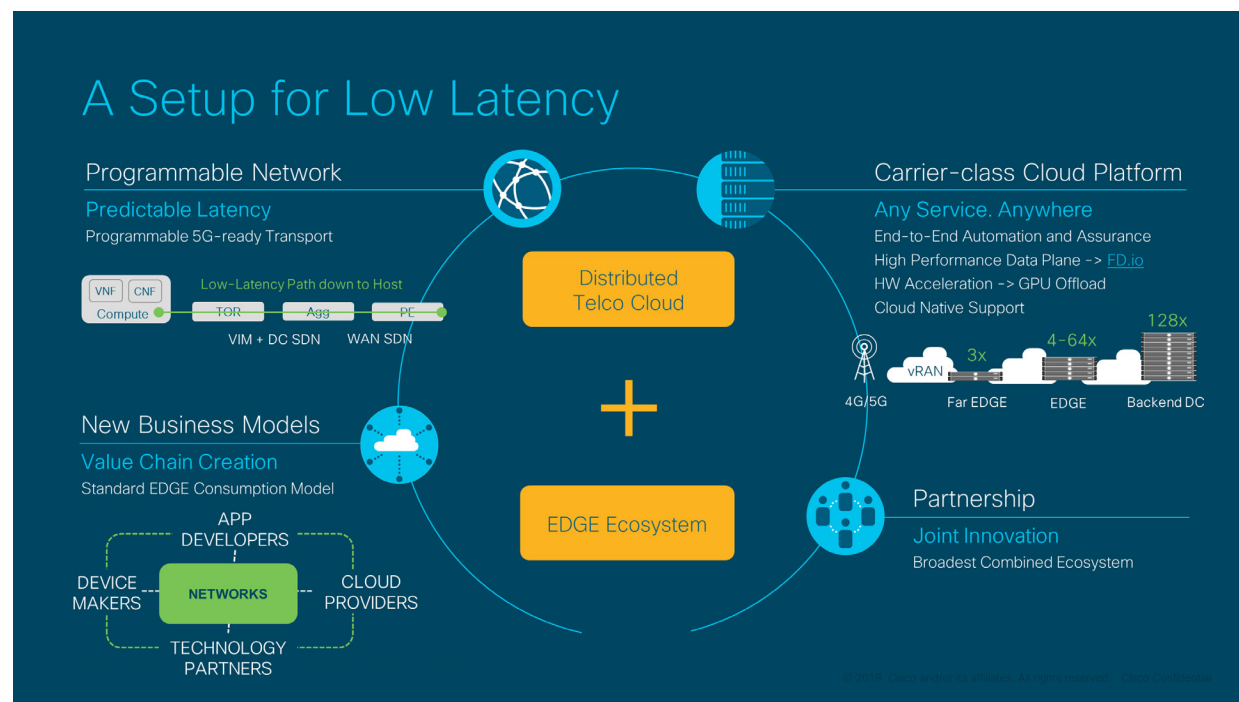
To provide a fast vSwitch data-plane, the carrier-class cloud solution is built on a standards-based architecture with various open source components such as OpenStack, KVM, Linux, Docker, Kubernetes, and fd.io. The platform is engineered to enable optimal footprints that range from more than 100 servers down to three servers in physically constrained locations like cell-site cloudlets.

Key capabilities include vGPU support for offloading AR/VR processing for rendering and encoding and future cloud-native support for hosting container workload on bare metal servers.

The transport network

The 5G-ready transport infrastructure foundation is a lightweight software defined networking (SDN) implementation that is based on innovative segment routing technology (SRv6, FlexAlgo). It provides deterministic end-to-end latency paths from Internet peering down to host-level at the farthest edge.

All network infrastructure and cloud components are managed and controlled by a common end-to-end automation and orchestration layer that provides a single view for service and function management.



Working with partners

With the proper setup, you can reach partners and provide access to a global application developer audience and community. The main objective is to establish a normalization layer with the same set of APIs across different infrastructures. A global edge cloud SaaS portal allows operators to visualize application delivery performance and developers to deploy their application container. This approach gives application and device makers a good experience so they can get applications into production quickly.

Customer experience matters

Low-latency offers a better customer experience and is becoming the new growth engine for network service providers. With expertise, innovation and service excellence from Cisco and our partners, you can monetize your network, save money through innovation, and lower network transformation risks with proven methodologies and expertise you can trust.

To learn more, visit:

- [cisco.com/go/mobile](https://www.cisco.com/go/mobile)
- [cisco.com/go/edge](https://www.cisco.com/go/edge)

