ericsson.com/
mobility-report

Enabling demanding use cases with CSP edge computing

Extract from the Ericsson Mobility Report
June 2022

Enabling demanding use cases with CSP edge computing

Edge computing is key to enabling latency-critical and bandwidth-hungry 5G use cases, representing significant growth potential for communications service providers (CSPs).

Key insights

- Emerging low-latency, high-bandwidth 5G use cases require new capabilities that can be addressed by CSP edge computing.
- The cost of building out and operating CSP edge compute resources is marginally more than at large-scale data centers, but significantly less than enterprise on-premise compute solutions.
- Deploying the compute resources at the edge of a CSP mobile network brings additional advantages, including reduced latency, higher performance, and improved data security and privacy.

Demand for immersive use cases has been held back by factors in the development of a new ecosystem, including networks, devices and applications. As this ecosystem matures, we expect the value brought about by edge computing will overcome the cost advantages held by large-scale data centers. Our analysis indicates that it is clearly possible for a CSP to build-out edge computing with an annual cost base not materially higher than a data center.

Historically, enterprises could either run their application workloads on-premise, based on the company's own IT infrastructure, or hosted in centralized data centers. There are several fundamental differences between these deployment options, including cost, control, security and regulatory compliance.

With the rollout of 5G, CSP mobile networks present an attractive proposition for running demanding enterprise applications close to target customers. A cost analysis of deployment shows that the cost to CSPs to deliver edge compute resources to enterprise customers is nearly half of what it would cost for an enterprise to build its own on-premise infrastructure with similar performance, reliability and data security.

Enter edge

Edge refers to the distribution of compute resource and applications to geographically distributed sites on the premises of an enterprise or in a CSP network. It provides compute resource and storage closer to where the data is generated and consumed. It offers significant advantages by enabling advanced data processing capabilities located close to where they are needed, reducing the latency inherent in centralized data centers. Deploying software at the edge comes with an increased cost compared with centralized deployment, but also enables a range of enhanced capabilities, including increased performance, reliability, data security and privacy, as well as reduced cost/bandwidth for the transport network.

Since data does not need to travel to remote locations for processing, analysis and rendering, enterprises can save precious milliseconds on round-trip times (RTT) while benefiting from more reliable data throughput. Enterprise on-premise edge computing can help insulate their networks from cyberattacks and distributed denial-of-service (DDOS) attacks on more centralized locations. There is also reduced risk of data being intercepted in transit, further adding to the security and privacy features of edge computing.

Edge computing can help organizations to fully comply with jurisdictional data regulations and sovereignty laws by allowing data to be processed close to its source.

CSPs can leverage the proximity of their existing sites to end users to set up edge compute, providing low-latency and high-performance IT capabilities for enterprise workloads as a service. For example, one way enterprises can reduce on-premise IT infrastructure is by deploying "infrastructure-less" branch offices; all IT on-premise applications, from communication, image processing and analytics to specialized enterprise services, can be hosted on the network edge.

A number of considerations must be addressed while rolling out compute capabilities alongside connectivity. There can be limitations to adding resources to some sites due to constraints on space, power and/or network capacity. Another challenge could arise from low fault tolerance of the commercial off-the-shelf (COTS) hardware used at the edge sites. CSPs may also require new sites to provide both continuous coverage and compute capabilities at critical locations to enable particularly demanding use cases.

The cost of the edge

To compare the cost of deploying compute resources at different scales, we convert capital expenditure into depreciation by dividing each asset category by the number of years it will be written down, and then add the resulting depreciation to the annual opex, providing a snapshot of the yearly cost structure. For example, power and cooling systems are written down over 14 years, whereas COTS servers are typically written down over 3 years.

Capex includes:

- Server capex is mainly the cost of COTS servers and virtualization software.
- Other capex consists of the cost of components such as power distribution and cooling systems.

Opex includes:

- The electrical power required to run and cool the servers.
- Other opex, mainly the cost of operations and maintenance (O&M).

As an example, we estimate the cost of compute resources for a CSP in Sweden. Initially, edge compute rollout is expected to be on aggregation sites having power capacity installed up to 10kW, hosting an average of 8 server units, each with 4 cores. With approximately 8,000 access sites and 1 aggregation site per 10 access sites, there is a virtual processor (vCPUs) capacity of 25,600 (800 sites x 8 servers per site x 4 cores per server) for enterprise applications at CSP-owned edge sites. Capex depends on the required capacity plus redundancy in the edge hardware components to meet the reliability requirements for edge services or applications. The geographic distribution can also be leveraged to improve the system availability by avoiding a single point of failure. We categorize the capex into server capex and other capex due to the faster cycle of server performance improvement compared to others. Servers are typically depreciated over 3 years while investments in power and cooling systems are depreciated over 14 years. Upgrading aggregation sites with edge compute capability, with an average of 8 units of servers, can draw up to 1.6MW (800 sites x 8 servers per site x 250W per server) for running the servers. With an assumed power efficiency factor of 2, 3.2MW power is needed on average to power all the aggregation sites. The cost of compute resource at each aggregation site is estimated be around USD 20,000. Hence the USD per critical watt for an edge site is USD 20,000/(8 servers x 250W/server) = USD 10/W. This cost is very similar to USD per critical watt for building a large-scale data center.

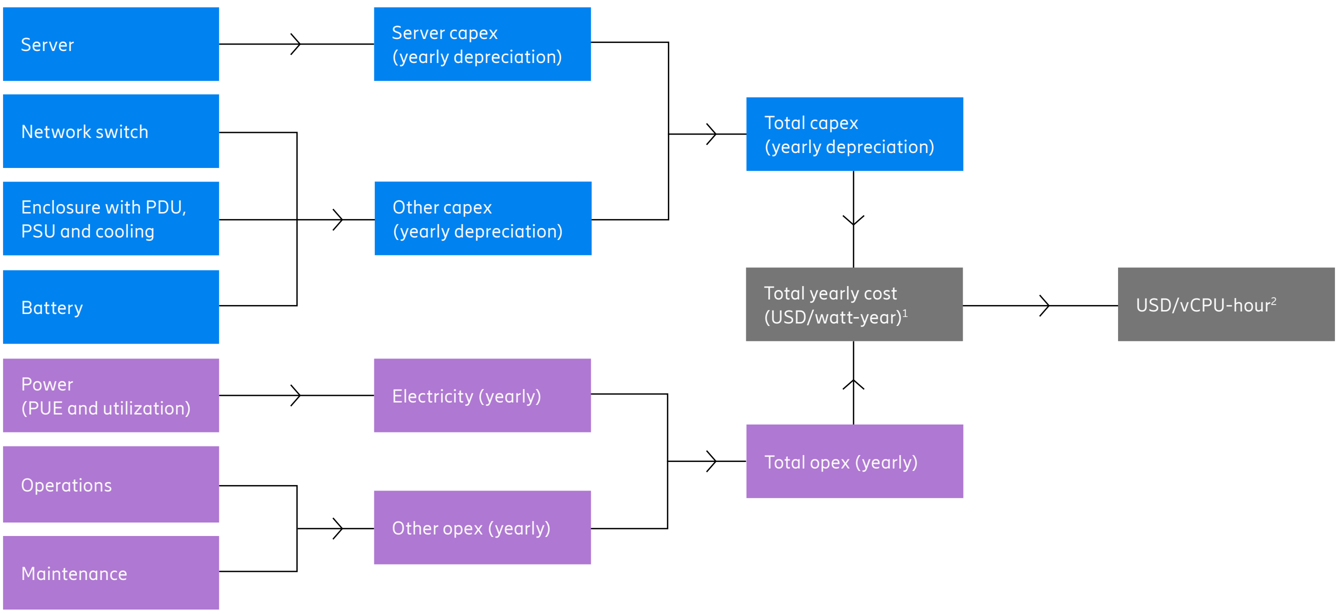
Opex is the sum of electricity cost and O&M. For the current study, we assume it to vary in the range of USD 0.10–0.15/kWh. For O&M, the cost of full-time employees required to manage and maintain the distributed edge servers is projected.

We constructed four different scenarios to estimate and compare the compute resource cost, based on USD per vCPU-hour.

- Scenario 1 is a base case with costs assumed for a small- or medium-sized enterprise handling its compute needs with its own IT infrastructure.
- Scenario 2 is an estimation of cost for a large-scale data center to provision the same capacity as the first case.
- Scenario 3 is built around provisioning the capacity used in the first two cases by deploying edge computing on the CSP network.
- Scenario 4 is an extension of the third case, with the addition of the cost to implement a set of measures to reduce power consumption. These include using renewable energy, dynamic usage of battery/power storage at peak times and advanced cooling technologies, including a heat exchanger for the server cabinets.

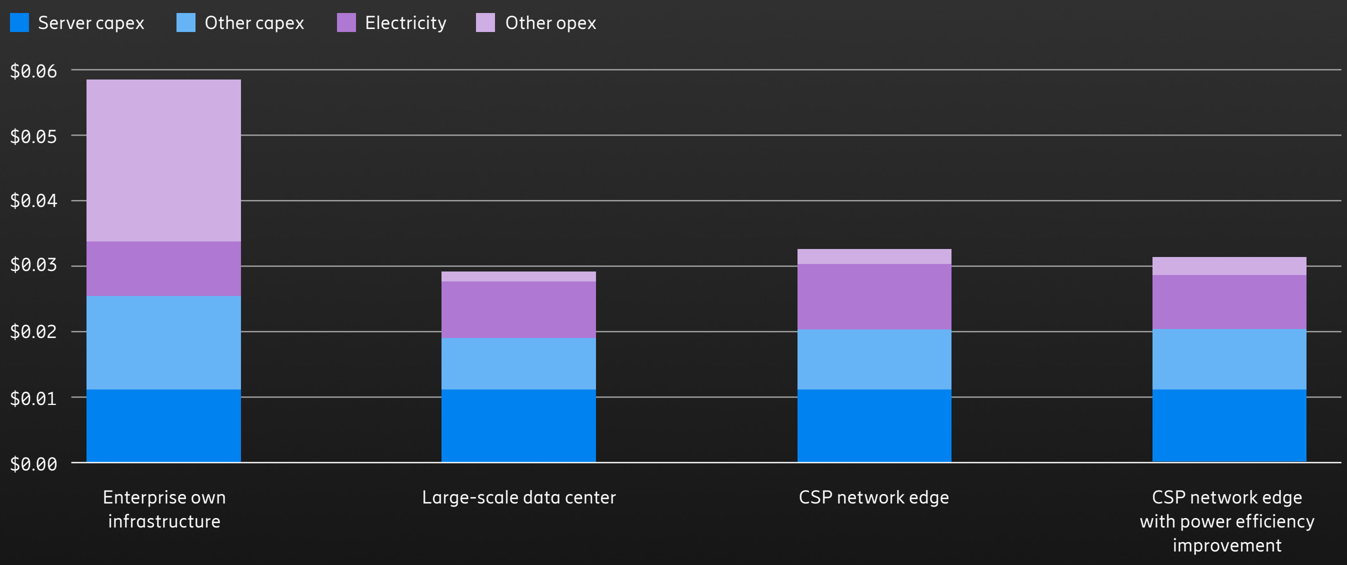
Server capex is the most significant parameter for all the scenarios except the base case where O&M (other opex) dominates due to the lack of scale. Electricity cost is the second largest factor in USD/CPU-hour for scenario 3. This leads to the significance of additional power efficiency elements in scenario 4.
With an estimate of expenditure in use cases suitable for edge deployment, the cost of edge compute resources can be just 10 percent more than that of a large-scale centralized one. Capacity utilization is the most important parameter for increasing the cost efficiency of the edge resources.

Figure 26: Annual cost estimation framework for compute resources



[1] USD/critical watt-year is equal to the total annual cost (depreciation) divided by the power (watts) required by the servers dimensioned for a defined compute capacity.
[2] USD/vCPU-hour: A unit of cost for compute resource that is equal to the total number of virtual processors divided into total annual cost of compute resource. This is an accepted metric for comparing the cost for alternative implementations.

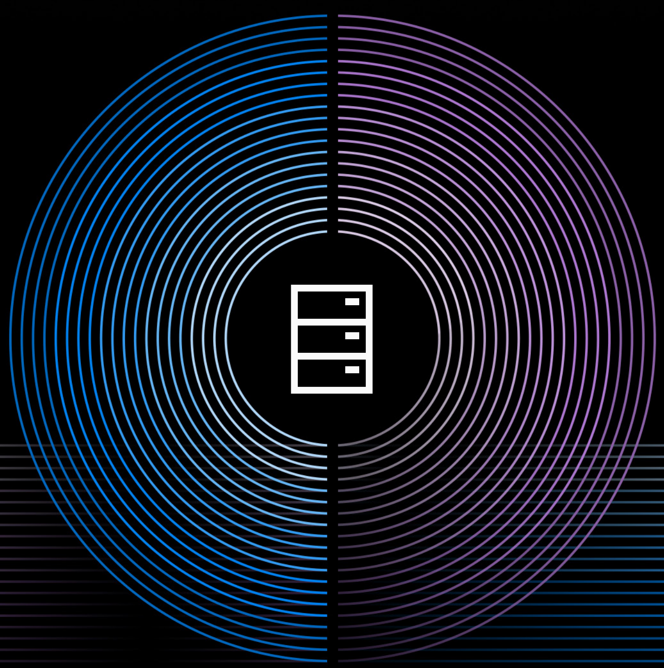
Figure 27: Cost breakdown of compute resources for different scenarios (USD/vCPU-hour)



CSP edge infrastructure resources are marginally more expensive than those at a large-scale data center but much less than those at an enterprise on-premise compute solution. CSP edge infrastructure also provides better latency and proximity to enterprise applications.

When comparing the costs of a large-scale traditional data center and a CSP network edge, we need to consider that those alternatives enable different use cases. Positive features for a CSP-operated edge include high location sensitivity, reduced latency (in the millisecond range), and guaranteed connectivity. However, edge compute infrastructure will have limited scalability compared to large data centers.

The short- to mid-term edge opportunity for CSPs should be seen in the wider context of the enterprise opportunity, where edge computing will be an enabler for a broad range of use cases, for example offerings such as private 5G networks, IoT platforms, cloud gaming and immersive experiences with XR. In the long term, when compute is deeply integrated in mobile networks, the most demanding use cases, including closed-loop industrial control systems, industrial robotics, extended reality with real-time synchronous haptic feedback (the Internet of Senses) and negotiated automatic cooperative driving for autonomous vehicles, will open up an expanding set of opportunities.



Building out edge computing on the CSP network unlocks significant business opportunities.

About Ericsson

Ericsson enables communications service providers to capture the full value of connectivity. The company's portfolio spans the business areas Networks, Cloud Software and Services, Enterprise Wireless Solutions, and Technologies and New Businesses. It is designed to help our customers go digital, increase efficiency and find new revenue streams. Ericsson's innovation investments have delivered the benefits of mobility and mobile broadband to billions of people globally. Ericsson stock is listed on Nasdaq Stockholm and on Nasdaq New York.

www.ericsson.com

Ericsson
SE-164 80 Stockholm, Sweden
Telephone +46 10 719 0000
www.ericsson.com

EAB-22:005355 Uen/3