WHITE PAPER

The Adaptive Network: A Framework for Understanding the Networking Implications of the Edge Cloud

Executive summary

Cloud services are pervasive. From individual users binge-watching Over-The-Top (OTT) video services to enterprises deploying Software as a Service (SaaS), cloud services are how people and organizations consume content and data. For years, large, centralized data center and cloud architectures have provided access to these services.

Now, a new generation of cloud-native applications is emerging in categories such as entertainment, retail, manufacturing, and automotive, which, in many cases, will be more compute-intensive and latency-sensitive. Traditional centralized cloud architectures will not meet the Quality of Experience (QoE) expectations for these applications and will require a more dynamic and distributed cloud model. As a result, compute and storage cloud resources need to move closer to the edge of the network, where content is generated or consumed, to meet the expected QoE. This new approach is referred to as Edge Cloud.

This shift to a distributed Edge Cloud model will result in an estimated three times as many data centers at the network edge as there are today, and will require the entire cloud ecosystem to think differently about the role of network connectivity[1].

This paper examines the drivers and implications of edge computing and explores how Ciena's Adaptive Network™ vision can provide an effective framework for the evolution to a distributed Edge Cloud architecture.

The world is changing, with data moving closer to the network edge

Imagine someone who is about to leave for a business trip making a simple request: *"Hey Siri, give me directions to the airport."* Within seconds, Apple Maps determines the shortest route and provides turn-by-turn directions. En route, a pop-up notification provides new directions to avoid traffic congestion ahead, adding just a few minutes to the trip. In each case, the Apple Maps request was likely served by a centralized data center that could be thousands of kilometers away, resulting in additional latency (delay) in processing the request. For a non-critical application like Apple Maps, such a response time is acceptable, and usually does not affect a user's ability to correctly navigate to their destination.

In another scenario, a shopper walks into their favorite grocery store, checking in via their smartphone app. Their movements—picking up and putting down items from the shelves—are captured by cameras embedded in the store ceiling. Computer vision AI analyzes these images to determine what the shopper has purchased and bills their credit card directly, eliminating the need for them to go through a cashier at the check-out counter. Significant compute resources will be needed either in the retail stores or at the edge to perform this near-real-time image processing to deliver a seamless customer experience.

An additional driver for computing at the edge is the value of processing the vast amounts of data generated locally by these applications and reducing backhaul traffic back to the central cloud. The goal is to reduce the latency and amount of backhaul traffic to the central cloud and more effectively serve the large-scale analytics required to push inferences and predictions to the devices at the edge, improving application performance.

Numerous emerging applications require lower latency than can be serviced via centralized data centers, as shown in Figure 1. Revenue for these applications is expected to see a 42 percent CAGR, from US$1.2B in 2020 to over US$5B in 2024, with the largest revenue drivers coming from video/content delivery networks, cloud gaming, and automotive applications[1].

1 Mobile Experts: "Edge Computing for Enterprises 2019", July 2019.