October 2020

Recommending Optimal Cloud Location For Workload Deployment

Mehmet Ozkan

Recommending Optimal Cloud Location For Workload Deployment

ABSTRACT

Edge computing is a technique by which cloud service providers (CSP) deploy their services to the edge of the network, closer to end users. Edge computing offloads substantial amounts of traffic from centralized cloud data centers and reduces latency. Unlike traditional cloud computing, which has only a handful of data centers that a CSP customer has to select from, edge computing entails tens of thousands of data centers and therefore, it is nontrivial to select the cloud data centers for an application to be deployed at. This disclosure describes techniques to predict optimal locations of a cloud resource based on keywords and/or network performance targets specified by the customer that is deploying an application. The techniques provide improved application performance for end users and an optimal cost of deployment for the customer of the cloud service provider.

KEYWORDS

- Cloud computing
- Edge computing
- Network virtualization
- Virtual server placement
- Search trends
- Demand signals
- Demand geography
- App scaling
- Geolocation
- Application programming interface (API)

BACKGROUND

Cloud computing is today an essential platform to exchange and deliver content for mobile, online video, social media, and other bandwidth hungry applications for billions of connected devices. The shift from traditional, on-premise computing to the cloud has created a substantial workload in centralized cloud data centers. Cloud service providers (CSP) have responded by adding hundreds of data centers in virtually every geographic region.

However, even with exponential growth in cloud computing capacities and global geographic expansion, substantial demand remains unmet. Besides, the centralization of services in data centers has resulted in increased latency, which is unacceptable for certain applications, e.g., artificial intelligence, machine learning, advanced analytics, virtual/augmented reality, Internet-of-Things (IoT), autonomous vehicles, dark factories, etc.

Edge computing is a technique by which CSPs deploy their cloud services to the edge of the network, closer to end users. Edge computing offloads substantial amounts of traffic from centralized cloud data centers and core network backbones and enables the distribution of traffic to thousands of smaller edge cloud locations closer to users. The proximity of edge servers to users is further attractive as it reduces latency, e.g., to under 10 ms. Indeed, the latest wireless 5G standard uses cloud native technologies, creating an opportunity for both telecom operators and incumbent CSPs to deploy thousands of telecom-edge (e.g., co-located with base station) locations for edge cloud services. A CSP can thus be expected to deploy publicly available compute resources globally at tens of thousands of edge-of-network locations.

Cloud computing enables enterprise users to deploy an application without making a detailed estimate of its potential demand and its geographic appeal. It enables enterprises to start small, test particular markets (geographies), and increase capacity and market footprint as

needed. Nevertheless, application developers are still required to select the geographic regions or zones of the CSP that the application will run in. Thus far, selecting a region during cloud compute setup is a relatively simple task, given the limited number and low granularity of options, which typically cover regions as large as continents, subcontinents, or countries.

As mentioned earlier, with edge computing, the number of data centers can run into the tens of thousands. An application requiring ultra-low latency may need to run at hundreds of locations. It is no longer a trivial task to select the cloud data centers or locations that the application should be deployed in. Incorrect selection of a cloud data center in edge computing can increase latency to the point that a deployed application is ineffective for end users. On the other hand, deploying an application at thousands of edge locations, although reducing the chance of application failure, can be prohibitively expensive.

Lacking historical data such as areas of user base concentration, usage statistics, etc., an application developer cannot optimally select edge locations for deployment of the application. For example, a gaming startup that is deploying a game for the first time will not have any data to choose the right edge locations. On the other hand, the mere collection and evaluation of edge-network performance metrics (e.g., latency), although important, is also unhelpful to the problem at hand, since the measured latencies indicate the performance of the network in the same edge zone, while the majority of actual users may be in other regions.

DESCRIPTION

This disclosure describes techniques to predict optimal cloud or edge locations and sizes for the deployment of a cloud resource based on the keywords, e.g., category, brand name, type, target customer segment, etc., and other factors provided by the application developer or

deploying party. Deployment of cloud resources based on the prediction can enable improved application performance for end users with an optimal cost of deployment.

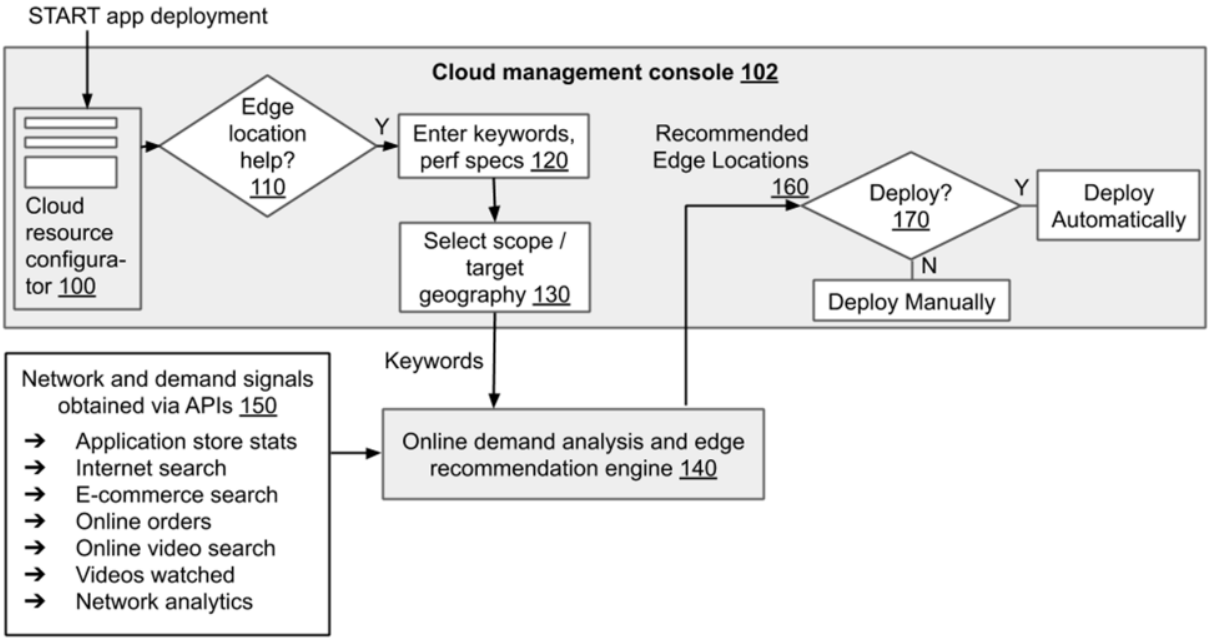

Fig. 1: Recommending an optimal cloud location for workload deployment

Fig. 1 illustrates a process to generate recommendations of optimal cloud location(s) for workload deployment, per the techniques of this disclosure. An authorized user, e.g., a person, entity, automation script, computer application, etc. representing the app developer or deploying party, accesses the management console (102) of a CSP (possibly via remote access terminal, e.g., command line interface or cloud shell) to create a cloud resource. During resource creation, the user enters configuration parameters (100) for the cloud resource, such as a virtual server. The parameters can include, e.g., CPU, memory, storage, network, operating system, etc. used to specify the cloud resource, as well as the location of the cloud resource to be created. The location represents the geographical regions or zones of the physical data centers where the compute resources (virtual and/or physical) are allocated.

The CSP management console offers to estimate optimal locations for the workloads (110). If the user chooses to leverage this capability, the user can enter keywords, target network performance metrics (e.g., latency, bandwidth, etc.), and other parameters (120) for the application. The keywords can include the brand name of the deploying party or the application, the product or service category that the application is associated with, any word or phrase that represents the application or its associated demand, etc. Examples of product categories or types can be, e.g., online games, electric scooters, online food delivery, etc.

The user enters the target market segment, the target geography, and the scope of the application (130). The target geography of the application can be specified at various granularities, e.g., global, regional, country, state, city, neighborhood, zip code, etc. The scope of the application can include, for example, the target language, e.g., English; the target time frame, e.g., the previous month, 2019, 2016-2020; etc.

The parameters entered by the user are passed to an online demand analysis and edge location recommendation engine (140). The engine may leverage other platforms via application programming interfaces (APIs, 150). For example:

- The engine can leverage trends information available via a trends platform to determine the number of internet searches related to the given keywords within the given geographic scope, language, and time frame.
- The engine can leverage a network monitoring and analytics platform to determine network performance metrics such as latency, available bandwidth, usage patterns of similar apps, etc.
- The engine can leverage statistics from an application store (e.g., a mobile app store, a desktop app store, a wearables app store, etc.) to determine locations from which similar applications are downloaded or used.

- The engine can leverage aggregated information content platforms such as video hosting websites or video providers, social media websites, e-commerce websites or apps, etc. to gain insight over viewership, product popularity, market trends, etc.

The demand analysis and edge location recommendation engine analyzes data acquired from various platforms using network analytics, correlation, prediction, artificial intelligence, and/or machine learning techniques to generate a sorted list of recommended geographical areas where the given keywords are associated with high demand. The engine returns the list of recommendations (160) to the cloud management console. For each listed geographic area, the most appropriate, e.g., geographically closest, cloud zones are deployed (170), either automatically or manually, based on user preference.

Example use case

A maps application company develops a mobile application that includes augmented reality (AR) features. The application runs on a mobile device that a pedestrian can use to identify surrounding buildings, streets, objects, e.g., trees, offering real time gamification capabilities to improve overall user experience. A successful marketing campaign for the app creates a buzz, which the company wants to leverage to expand into new international markets. The AR features of the application require ultra-low latency for a seamless experience. The company fears negative reviews due to an initial launch with high latency, subpar, user experience.

The company therefore plans to deploy their app at the edge cloud locations; however, there are tens of thousands of edge locations available. The company cannot afford to deploy the application in all the available locations. On the other hand, choosing the wrong locations can mean unacceptably high latencies for users accessing the app from geographic areas not covered

by the deployed edge locations. Since the company has no market data for similar products or games, choosing the right edge locations is difficult.

Per the techniques of this disclosure, the company starts app deployment on a cloud platform by entering as parameters its brand name, the product categories ("maps," "mobility," "exercise application," etc.), the target international geography ("Europe"), the target network performance (latency less than 6 ms; bandwidth at least 200 Mbps); etc.

Based on the above inputs, the online demand analysis and edge location recommendation engine as described herein analyzes available network traffic data to determine the highest previous usage of similar applications. Using the given keywords, the engine leverages cloud demand analytics, search trends, demand signals, etc. from search engines, social media, e-commerce websites, content providers, etc. to determine the recommended locations. The recommendation engine determines optimum edge cloud locations for the deployment and recommends a list of initial deployment locations.

In this manner, the techniques described herein enable application developers to select with surgical precision target markets within large, e.g., continental or subcontinental, geographies and obtain recommendations for deployment of cloud infrastructure at suitable edge locations that best serve (in terms of latency or other performance criteria) the target markets, selected out of a large number of potential edge cloud locations for app deployment.

CONCLUSION

This disclosure describes techniques to predict optimal locations of a cloud resource based on keywords and/or network performance targets specified by the customer that is deploying an application. The techniques provide improved application performance for end users and an optimal cost of deployment for the customer of the cloud service provider.