Rethinking the Cloud Architecture for a Mobile Edge Cloud

Prashant Shenoy, UMass Amherst
shenoy@cs.umass.edu

The era of cloud computing is upon us. Cloud computing allows Internet applications to be hosted on remote server and storage resources that are leased on-demand from the cloud providers. Today's cloud platforms host a plethora of popular Internet applications in domains ranging from news, e-commerce, entertainment, banking, games and social media. A concurrent trend is the increasing ubiquity of mobile devices and their use as the user's primary Internet device. According to a 2015 study, 51% of total time spent consuming digital media in the USA in 2015 was on smartphones vs 2008 where smartphones were only 12% of traffic. In developing countries, where the smartphone is often the *only* Internet device owned by the user, wireless hosts dominate the traffic seen by cloud applications. The intersection of these technology trends —the rise of *cloud computing* and the ubiquity of *mobile computing*—raises a number of interesting research challenges. First, traditional cloud platforms are agnostic to the type of clients, whether wired or wireless, served by end-users. Second, mobile devices significantly differ from wired hosts in their characteristics—they tend to be more resource-constrained than their wired counterparts, are nomadic and battery-powered, and often access cloud services in a location-aware manner. The interplay between these challenges raises a key research question: how can we rethink cloud computing platforms to better support the needs of mobile users and applications? Our approach to addressing this challenge is a new architecture for a *mobile-aware edge cloud platform* that can tailors it mechanisms to serve a increasingly mobile user base.

To rethink cloud computing in the mobile era, we must first understand the key differentiating characteristic of mobile workloads and users. First, unlike traditional web workloads, mobile workloads seen by cloud services exhibit both *temporal* and *spatial* dynamics. Temporal dynamics include variations seen over long and short time scales such as time-of-day effects and seasonal variations. Spatial dynamics arise due to mobile users who move from one location to another over a period of time, resulting in variations in the volume of requests coming from different locations—unlike wired hosts where the locations do not change. Second, cloud applications needs to be aware of the location of their end-users, in order to tailor the content to the user's current location or to serve the user from a nearby server. Until now, location-aware optimizations were left to the application-level without any direct support from the cloud platform. Third, the cloud application needs to be aware of the resource- and energy-constraints on end-devices. For instance, video content that is streamed to a device can be tailored in terms of bandwidth and resolution based on the screen size and available wireless bandwidth. Energy demands of a mobile application accessing a cloud service can be optimized by performing more "in-network" or cloud-side processing rather than local processing on the device. Finally, since users accessing a cloud application may be distributed across different regions, the cloud service needs to be geographic-aware, in order to implement optimizations such as geographic replication ("geo-replication") or geographic provisioning ("geo-elasticity") . Thus, the need to be workload-aware, location-aware, resource-aware and geographic-aware results in a fundamentally different cloud architecture from today's cloud platforms, and are all features that are central to a mobile-aware edge cloud platform.