SeamlessClouds for Real-Time, Distributed Edge Applications

Yanyong Zhang, Winlab, Rutgers University

The confluence of the three maturing areas of ubiquity, cloud computing and big data will open doors to a new class of "edge" applications with a transformative impact on society. Ubiquitous deployment of computing devices, ranging from tablets and smart phones to miniature sensing devices, makes it possible to do real-time data capture and information dissemination in the cyber, physical and/or social worlds. These devices, are not necessarily constrained by what they have on board, but can access the vast computational and storage resources of a large datacenter that hosts cloud services for these applications over the network. Big data analytics could leverage these computational resources to perform sophisticated knowledge extraction and serve very complex queries. However, moving geo-distributed data across the Internet to avail of these computational resources may negate the real-time capture, processing and responsiveness mandates of these applications. Instead, there are recent proposals to move the computational resources close to the edge, in the form of edge-clouds. Prior proposals on these edge-clouds keep them relatively compartmentalized, making it difficult to compose and structure computations that can span many such clouds, map and orchestrate these computations efficiently on the underlying infrastructure, enhance resource utilization, and seamlessly move computations/data across these edge clouds. To address these pitfalls, this project proposes SeamlessClouds, a seamless platform across these distributed edge-clouds and the back-end cloud for deploying next generation edge applications.

This project will design, prototype and evaluate important parts of the underlying infrastructure needed to seamlessly integrate the disparate edge-clouds and the back-end cloud. With the Map-Reduce framework naturally suited to the parallelism and pipelined stages in these applications, a task graph representation of these Map-Reduce tasks in the computation will be used as the starting point for their structuring, placement and movement across the entire infrastructure. Through rigorous experimentation and analysis, this research will provide key insights for the following questions: how can these tasks be mapped on the geo-distributed Infrastructure for proximity of the data to computations and/or reducing the communication overheads while maximizing the parallelism, with even individual map/reduce phases spanning different sites? Can the computation itself be restructured based on the resource availability on a holistic scale for better computation-communication trade-offs? Would approximations that can be made in many of these applications be used to reduce the overheads of communication and data related stalls? Can the resources be elastically scaled, and consequent utilization enhanced, beyond the boundaries of a single cloud? What does it take to build a virtually shared infrastructure, especially from the networking angle, across multiple edge-clouds towards making mobility and data migration seamless? In the process, this exercise will demonstrate the benefits of the seamless infrastructure spanning multiple clouds, compared to the centralized and compartmentalized alternatives.