Application Repartitioning in the Cloud

Lei Yang and Jiannong Cao, The Hong Kong Polytechnic University

This installment of Computer’s series highlighting the work published in IEEE Computer Society journals comes from IEEE Transactions on Cloud Computing.

Mobile cloud computing aims to enhance mobile application performance by tapping into cloud resources. A challenge of this emerging paradigm is application partitioning: deciding which parts of an application should be offloaded to the cloud and which should be executed on mobile devices. This problem is easily solved when the device status and network connections are static while the application runs. Dynamic mobile cloud environments, however, require a different approach. In “Run Time Application Repartitioning in Dynamic Mobile Cloud Environments,” Lei Yang and his colleagues propose a novel solution: application repartitioning (IEEE Trans. Cloud Computing, vol. 4, no. 3, 2016, pp. 336–348).

An initial partition is generated when the application starts (see Figure 1). As the application runs, the framework periodically predicts the device and network status and continuously updates the partition. To demonstrate the framework’s real-world effectiveness, Yang and colleagues present a case study in which they take the dynamic network connections to the cloud. Their first task was to accurately predict mobile users’ real-time network bandwidth. They applied a crowdsourcing technique in which they collected users’ network fingerprints, including locations and the corresponding bandwidths. Users’ bandwidth could then be projected by predicting where they would go in the future. The authors used this approach because an application’s lifecycle is only a few seconds to several minutes. In this short period, the bandwidth fluctuation for one user is dominated by the user’s mobility.

Using these network bandwidth predictions, the authors developed an online algorithm that periodically updates the application’s partitions. At each repartitioning, the algorithm compares the cost of switching to an optimal partition under the predicted network bandwidth with the cost of sticking to the previous partition, and then chooses the less costly option. Through simulations on real-world Wi-Fi traces, Yang and his colleagues show that their proposed solution reduces application completion time by 35 percent compared with previous efforts.

As the first to study application repartitioning, this work provides solutions to the performance degradation issues arising from dynamic network and device statuses. Although focused on dynamic networks, the proposed framework could apply to any environment in which device workloads change dynamically during an application’s execution.

LEI YANG is an associate professor in the School of Software at South China University of Technology, and a postdoctoral fellow in the Department of Computing at The Hong Kong Polytechnic University. Contact him at sely@scut.edu.cn.

JIANNONG CAO is a chair professor and head of the Department of Computing at The Hong Kong Polytechnic University. Contact him at csjcao@comp.polyu.edu.hk.