Project Proposal for CSC 2228 : Scheduling in Multi-Tier Cloud Architecture

Cloud computing resources scale elastically, and hence reduce the risk of under provisioning and wastage due to over provisioning during non-peak hours. Further, with the widespread use of mobile devices and their increasing resources, Mobile Edge Computing is gaining importance. By offloading computation to edge servers, device energy consumption and execution delays can be greatly improved. This paradigm introduces new scheduling and provisioning challenges that are yet to be solved. We seek to address the problem of dynamic resource provisioning in multi-tier cloud data centres.

This problem is even more difficult for edge computing because depending on the tier used, the latency changes whereas the latency is known before hand in cloud computing. Moreover, the machines in different tiers have different resource capacities, initialization overheads and expenses and hence load balancing is not straightforward. Another issue is where scheduling should be done: it must be done on a trust worthy tier to avoid malicious users from abusing the system[1]. Few scheduling schemes exist for edge computing.

Several research efforts have been made to efficiently solve the dynamic resource allocation problem. Some algorithms use queuing models that determine which tier to re-provision[2], some algorithms also factor in heterogeneity of resources in different tiers and use profiling of machine performance to select a tier[3] and some algorithms are designed based on low level hardware based performance models[4]. Finally, some methods use control theory to allocate resources rather than using predictive methods[5].

We propose to start with a simple model for a two-tier structure having one parent and two children. We plan to have CPU, memory, storage and access latencies as some of the parameters used to determine the cost of provisioning decisions. We will analyse the a greedy strategy: the tasks with minimum resource usage and low latency requirement will be scheduled on edge and others will be scheduled higher up in the tier hierarchy. Further, we plan to extend our analysis to more than two tiers.

[1]. Hao,Z., Yi, S., Novak,E., Li, Q., Challenges and Software Architecture for Fog Computing

[2]. Urgaonkar, B., Pacific,G., Shenoy,P. et.al, An Analytical Model for multi-tier internet services and its applications.

[3]. Dejun,J., Pierre, G., Chi, C., Resource Provisioning of Web Apps in Heterogenous Clouds

[4]. Marin, G., and Mellor-Crummey, J. Cross architecture performance predictions for scientific applications using parametrized models.

[5]. Kalyvianaki, E., Charalambous, T., and Hand, S. Self adaptive and self-configured cpu resource provisioning for virtualised servers using Kalman filters.