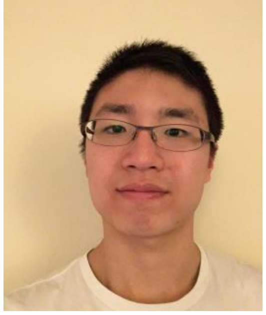
Carnegie Mellon University
Computer Science Department

THESIS PROPOSAL



Andrew Chung

Realizing value in shared compute infrastructures

Thursday, February 6, 2020 – 2:00 p.m. – GHC 6501

As operations become increasingly digitized and as data processing tasks become more and more specialized with the proliferation of various types of data applications, companies are moving their workloads off of dedicated, siloed clusters in favor of more cost-efficient shared data infrastructures. These shared data infrastructures are often deployed on highly heterogeneous servers, are multi-tenant with server resources shared across multiple organizations, and serve widely diverse workloads.

Both operators and users of such shared data infrastructures strive to optimize for value. Operators seek to satisfy the demands of their customers (i.e., help users maximize their value) to increase adoption and lower turnover, all the while without sacrificing cluster operation costs and overhead. At the same time, users look to complete their tasks in an efficient and timely manner without having to pay large amounts of money.

But, the highly heterogeneous nature of these shared environments imposes a high barrier to value attainment for both operators and users: Operators face difficult challenges in knowing how to assign compute resources to customers when heavily loaded. Users, on the other hand, have a wide variety of different types of compute resources available for rent, making it difficult for them to make value-efficient resource acquisition decisions for their applications, given application constraints. Indeed, maximizing value in shared data infrastructures necessarily requires effort from both operators and users.

In my work, I explore the problem of value attainment in shared data infrastructures from both the perspectives of operators and users. On the operator front, I explore using the notions of historic inter-job dependencies and expected job utility to inform cluster resource managers about upcoming jobs, their resource requirements, and the potential value they generate to users. Cluster resource managers can in turn use the information to effectively allocate cluster resources to jobs to achieve high user value attainment. On the user front, my work proposes and evaluates two resource acquisition strategies and systems for renting virtual machine (VM) instances in the public cloud, one for running online services and the other for general batch analytics jobs, with each demonstrating significant cost savings for users.

Thesis Committee:
Greg Ganger, Chair
Phil Gibbons
George Amvrosiadis
Carlo Curino, Microsoft Research

Thesis Summary: http://cs.cmu.edu/~afchung/resources/docs/thesis-proposal.pdf