Why Do We Need Data-driven Datacenters?

Designing datacenters that are reliable, energy-efficient, and capable of delivering high performance and high utilization is a nontrivial mission. In this talk, I argue that we need more data-driven operations in datacenters to better optimize these design goals. I will present three examples that demonstrate how data-driven predictive modeling can help us anticipate and mitigate critical events in datacenters such as job failures [ICDCS’17], database server load changes [SIGMOD’18], and even changes in the demands of co-located jobs [HPCA’18]. I will discuss how these different predictions can improve the way we allocate and manage resources dynamically in datacenters. Finally, I will present a recent, open-sourced tool that leverages clustering techniques to unlock significant performance on modern datacenter hardware through lightweight cache profiling and partitioning.