The University of Chicago Computer Science Department
PRESENTS:

Resource Management and Scheduling for Emerging AI Applications

Alexey Tumanov
University of California, Berkeley

Abstract:
A new class of Artificial Intelligence applications is emerging that imposes a challenging set of requirements on how we program the cloud and how we manage cloud resources efficiently. With the end of Moore's Law and Dennard scaling, coupled with simultaneous increase in the heterogeneity of increasingly interactive AI applications with end-to-end latency constraints, the future of AI Systems depends on advances in resource management and scheduling for these applications. First, these applications generate an increasingly heterogeneous set of tasks, both in the resources optimal for their performance and in the time scale of individual tasks. Second, they are increasingly user-facing, imposing a set of soft real-time constraints on the frameworks serving these workloads. Third, they individually expect or benefit from heterogeneous and often conflicting resource allocation policies --- a challenge for unifying frameworks that aim to support them. Thus, a set of three emergent requirements must be efficiently addressed: (1) heterogeneity awareness in space and time, (2) soft real-time end-to-end latency constraints, and (3) scheduling policy heterogeneity at the application level. To address these requirements, I will present (1) **TetriSched** --- a mathematical framework to capture the performance as a function of resource space and timeliness requirements of these applications for cost-efficient and heterogeneity-aware resource allocation, (2) **Inferline** --- a soft real-time system for achieving these requirements under unpredictable bursty workloads when multiple ML models are composed for inference; (3) **Ray** --- an active open source project that brings some of these ideas together and serves as the unifying framework for distributed ML, addressing the challenge of scheduling policy heterogeneity.

Bio:
Alexey Tumanov is a Postdoctoral Researcher at the University of California Berkeley, working with Ion Stoica and collaborating closely with Joseph Gonzalez in RISELab, Department of Computer Science. Alexey completed his Ph.D. at Carnegie Mellon University, advised by Gregory Ganger. At Carnegie Mellon, Alexey was awarded the prestigious Canadian government fellowship, NSERC Alexander Graham Bell Canada Graduate Scholarship (NSERC CGS-D3) and was a member of the Intel Science and Technology Center for Cloud Computing and the Parallel Data Lab. Alexey's Systems research spanned the entire stack, starting with agile stateful VM replication with para-virtualization at the University of Toronto (working with Eyal de Lara) and most recently involving resource management for emerging AI applications. Alexey is the recipient of several awards, including the Best Graduate Student Teaching Assistant at Carnegie Mellon and the ***Best Student Paper*** *award for his thesis work on TetriSched at EuroSys 2016.*

Friday, April 12, 2019
11:00 am
JCL 390
Host: Sanjay Krishnan