Session 1  Tuesday February 15, 2022

3:15 PM  Serverless Computing

- **Opening Remarks — Christos Kozyrakis**
- **Apiary: Tightly Integrating Compute and Data for Efficient, Transactional, and Observable Functions-as-a-Service — Qian Li & Peter Kraft**
  Apiary is a novel FaaS platform for data-centric applications. Apiary guarantees functions run as ACID transactions with end-to-end exactly-once semantics and provides advanced observability capabilities like automatic data provenance capture. Despite offering more features and stronger guarantees than existing FaaS platforms, Apiary outperforms them by 7–68× on realistic microservice applications by greatly reducing communication overhead.
- **Fixpoint — Yuhan Deng**
  Fixpoint is a functional operating system that is based on a data model that represents computational relations between content-addressed objects. By utilizing WebAssembly, Fixpoint provides functional computation with a low cost of transition between the system and user programs. In addition, any object is associated with the computation history that outputs it, computations in Fixpoint are reproducible. We think it is easier for users to program and debug in Fixpoint as programs expand in size and complexity, and it also unlocks more potential for cloud-computing providers to offer "computation as a service".

4:30 PM  Time-aware Systems

- **Self-Programming Networks (SPNs) Overview: CloudEx, On-Ramp, and Nezha — Balaji Prabhakar**
  The SPN research group in the Platform Lab has been working on several key systems that can be built using their scalable and accurate clock sync algorithm — Huygens. In this talk, Balaji Prabhakar will overview the status of SPN projects CloudEx, On-Ramp and Nezha.

- **4:45 — 5:15 pm  Nezha: A high-performance consensus protocol using accurately synchronized clock — Jinkun Geng**
  We present Nezha, a high-performance and deployable consensus protocol that exploits accurate clock synchronization. Nezha only relies on clock synchronization for high performance but not for correctness, and its deployment does not require special hardware or physical network access. We compare Nezha to 4 baselines in
public cloud: Multi-Paxos, Fast Paxos, NOPaxos and Raft. Evaluations show Nezha outperforms the baselines by a median of 5.3x (range: 1.8x-19.9x) in throughput, and by a median of 1.5x (range: 1.2x-2.2x) in latency.

5:15  Break

5:30  Panel: Future of Cloud Panel

- Panelists: Neeraja Yadwakar (VMware / UT Austin), Rodrigo Fonseca (Microsoft Research), Keith Winstein (Stanford)

Session 2  Wednesday, February 16th

2:00 PM  ML Optimization –

- Sparse GPU Kernels for Deep Learning — Trevor Gate
  Scientific workloads have traditionally exploited high levels of sparsity to accelerate computation and reduce memory requirements. While deep neural networks can be made sparse, achieving practical speedups on GPUs is difficult because these applications have relatively moderate levels of sparsity that are not sufficient for existing sparse kernels to outperform their dense counterparts. In this work, we study sparse matrices from deep learning applications and identify favorable properties that can be exploited to accelerate computation. Based on these insights, we develop high-performance GPU kernels for two sparse matrix operations widely applicable in neural networks: sparse matrix-dense matrix multiplication and sampled dense-dense matrix multiplication.

- RecShard: Statistical Feature-Based Memory Optimization for Industry-Scale Neural Recommendation — Geet Sethi
  We propose RecShard, a fine-grained embedding table (EMB) partitioning and placement technique for deep learning recommendation models (DLRMs). RecShard is designed based on two key observations. First, not all EMBs are equal, nor all rows within an EMB are equal in terms of access patterns. EMBs exhibit distinct memory characteristics, providing performance optimization opportunities for intelligent EMB partitioning and placement across a tiered memory hierarchy. Second, in modern DLRMs, EMBs function as hash tables. As a result, EMBs display interesting phenomena, such as the birthday paradox, leaving EMBs severely under-utilized. RecShard determines an optimal EMB sharding strategy for a set of EMBs based on training data distributions and model characteristics, along with the bandwidth characteristics of the underlying tiered memory hierarchy. In doing so, RecShard achieves over 6 times higher EMB training throughput on average for capacity constrained DLRMs. The throughput increase comes from improved EMB load
balance by over 12 times and from the reduced access to the slower memory by over 87 times.

3:00 PM  Systems for ML

- **ML-EXray: Visibility into ML Deployment on the Edge — Hang Qiu**
  When ML models get deployed on edge devices, very often, the performance can abruptly drop without obvious reasons. The key challenge is that there is not much visibility into ML inference execution on edge devices, and very little awareness of potential issues during the edge deployment process. We present ML-EXray, an end-to-end framework that provides visibility into layer-level details of the ML execution, and helps developers analyze and debug cloud-to-edge deployment issues.

- **VIVA: An End-to-End System for Interactive Video Analytics — Franky Romero & Daniel Kang**
  The growth of video volumes and increased DNN capabilities has led to a growing desire for video analytics. In response, the data analytics community has proposed multiple systems that optimize specific query types (e.g., selection queries) or a particular step in query execution (e.g., video retrieval from storage). However, none of these systems provide end-to-end, practical video analytics for users to iteratively and interactively engage with queries, as is the case with analytics systems for structured data. In response, we are building VIVA: an end-to-end system for interactive video analytics. In our talk, we will describe the challenges and design requirements for VIVA’s development and outline ongoing and future work for realizing VIVA.

3:45 pm  BREAK

4:30 PM  ML Systems Panel

- Panelists: Kim Hazelwood (Meta), Peter Mattson (Google), Luis Ceze (UW / OctoML), Michael Carbin (MIT), and Kunle Olukotun (Stanford).

Session 3  Thursday, February 17, 2022

2:00 PM  Efficient communication

- **Zero-Copy Serialization for Microsecond-Scale Networking — Deepti Raghavan**
  Data serialization is a critical but costly component of datacenter applications. As datacenter NICs and kernel-bypass networking stacks increasingly operate at
microsecond latencies, serialization has become unaffordable. There is a tension between microsecond-scale networking, which commonly minimizes memory copies to reduce overhead, and software serialization libraries, which rely on in-memory copies to transform data from application-friendly formats to a single contiguous buffer for I/O. To resolve this tension without requiring new datacenter hardware, we leverage existing NIC scatter-gather capabilities to offload costly in-memory copies. We present Cornflakes, a new serialization library that reduces serialization overheads for microsecond-scale network stacks and NICs. Using Cornflakes for a key-value store achieves 1.8-3× higher throughput than prior copying-based serialization libraries.

- **A Case for Replacing TCP (in the Datacenter) — John Ousterhout**
  Although TCP is a tremendously successful transport protocol that has survived 40 years of dramatic technology changes, virtually every aspect of its design is wrong for the datacenter. This talk will discuss the problems with TCP, ranging from its use of connections and streams to its mechanisms for reliable delivery and congestion control. If we are to make significant headway against the "datacenter tax" we must move most datacenter traffic to a new and fundamentally different protocol. The talk will then discuss how the Homa transport protocol solves all of the problems of TCP and suggest a migration strategy for bringing Homa into widespread usage.

3:00 PM  Systems Infrastructure

- **SOL: Safe On-Node Learning in Cloud Platforms — Yawen Wang**
  Cloud platforms run many software agents on each server node to manage different aspects of node operation. In this talk, I will discuss the challenges in improving cloud platforms by infusing online machine learning into server node agents. I will then present SOL, a framework for designing ML-based agents that are safe and robust to a variety of failure conditions that occur in production.

- **Syrup: User-Defined Scheduling across the Stack — Kostis Kaffes**
  I will present Syrup, a framework that enables everyday application developers to specify custom scheduling policies easily and safely deploy them across different layers of the stack over existing operating systems like Linux, bringing the benefits of specialized scheduling to everyone. For example, Syrup allowed us to implement policies that previously required specialized dataplanes in less than 20 lines of code and improve the performance of an in-memory database by 8x without needing any application modification.
4:00 pm   Poster Session