**CRUNCH Seminars at Brown, Division of Applied Mathematics**

**Friday – October 4, 2019**

## Benchmarking TPU, GPU, and CPU Platforms for Deep Learning

Khemraj Shukla

Training deep learning models is compute-intensive and there is an industry-wide trend towards hardware specialization to improve performance. To systematically benchmark deep learning platforms, we introduce ParaDnn, a parameterized benchmark suite for deep learning that generates end-to-end models for fully connected (FC), convolutional (CNN), and recurrent (RNN) neural networks. Along with six real-world models, we benchmark Google's Cloud TPU v2/v3, NVIDIA's V100 GPU, and an Intel Skylake CPU platform. We take a deep dive into TPU architecture, reveal its bottle- necks, and highlight valuable lessons learned for future specialized system design. We also provide a thorough compari-son of the platforms and find that each has unique strengths for some types of models. Finally, we quantify the rapid performance improvements that specialized software stacks provide for the TPU and GPU platforms.