Petuum: A New Platform for Distributed Machine Learning on Big Data
Carnegie Mellon University, *Institute for Infocomm Research, A*STAR

**Objectives**
- Support modern opti.+probabilistic methods
- Systematic approach to Big Data+Model problems

**ML program vs Traditional program**
- Opti/MCMC-centric ML program
- Traditional operation-centric program

**ML-centric view**
- Iterative-convergent ML program:
  \[
  \arg \max_\theta L(\{x_i, y_i\}_{i=1}^{N} : \bar{\theta}) + \Omega(\bar{\theta})
  \]
- for (t = 1 to T) {
  doThings()
  \[\theta^{t+1} = g(\bar{\theta}, \Delta_t, \bar{\theta}(D))\]
  doOtherThings()
}

**More properties of ML programs**
- Model parameters not independent
- Careful model-parallelism needed for stability
- Uneven parameter convergence – opportunity to prioritize

**ML parallelization**
- Exploit self-healing with async-like parameter access, enforce staleness bounds to ensure model convergence
- “Eagerly” push out parameters to keep staleness distribution small (see theory)

**Strads: Programmable Structure-Aware Scheduling**
- Prioritizes parameters based on convergence rate
- Avoid updating non-independent model parameters in parallel (see theory)
- Prioritization makes dependency analysis easier!

**Bösen: Programmable Bounded-Async K-V store**
- Exploit self-healing with async-like parameter access, enforce staleness bounds to ensure model convergence
- “Eagerly” push out parameters to keep staleness distribution small (see theory)

**Theoretical Guarantees**
- Bösen bounded-async is important for quality
- Strads scheduling is nearly-ideal

**Performance**
- Bösen improves throughput while maintaining iter quality
- Strads improves iter quality while maintaining throughput

**App library:** DNN/CNN, MedLDA, Sparse Coding, Random Forest, Distance Metric Learning, Multiclass Logistic Regression, and more ...

---

http://petuum.org