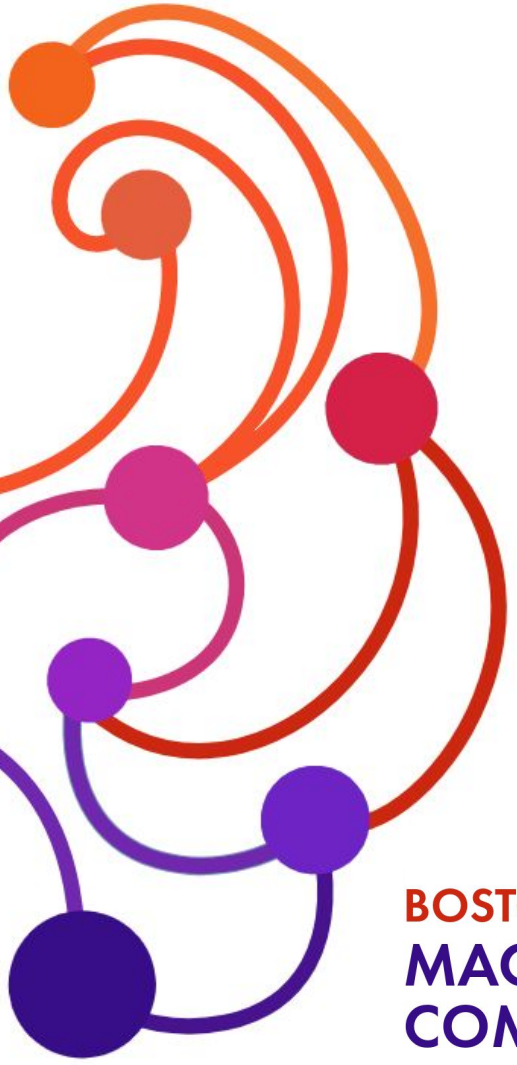


Large Scale Distributed Deep Networks

Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc' Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Andrew Y. Ng

BOSTON UNIVERSITY
MACHINE INTELLIGENCE COMMUNITY

Justin Chen
Nov. 13, 2017

Previous Limitations with Distributed Training

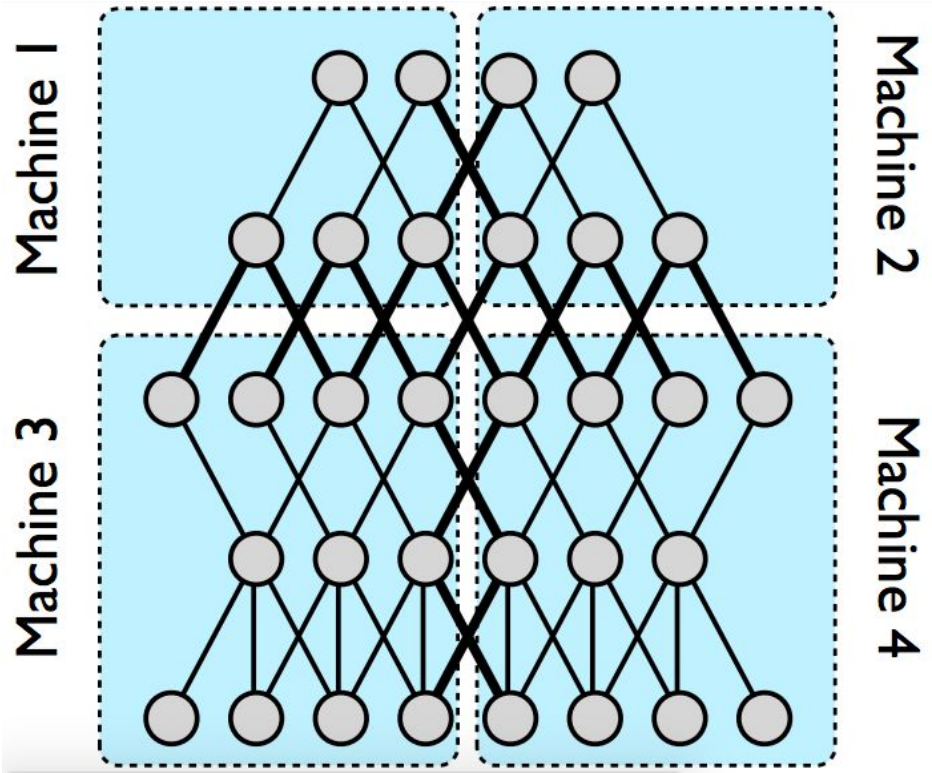
- Large models cannot fit onto a single GPU
- Previous distributed algorithms assumed convexity or sparsity
- MapReduce and GraphLab are not suitable
- Goal of this work:
  - Asynchronous
  - Distributed
  - No assumptions on architecture, sparsity, and convexity

DistBelief Framework

- Eventually became TensorFlow
- Automatic parallelization, synchronization, and communication
- Model parallelism
  - Multithread on single machine
  - Message passing across machines
- Data parallelism
  - Multiple copies of model across cluster

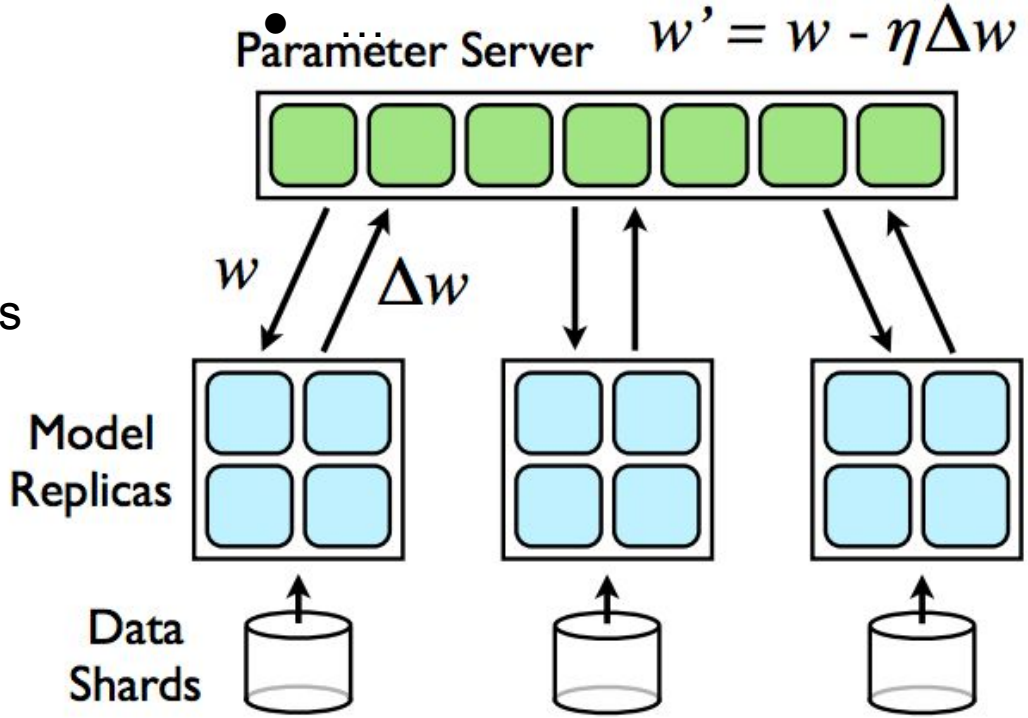
Model Parallelism

- Image depicts a single replicated model across four machines
- Blue box/machine is a **partition**
- Parameters communicated once (thick black lines)
- Node parallelized within partition



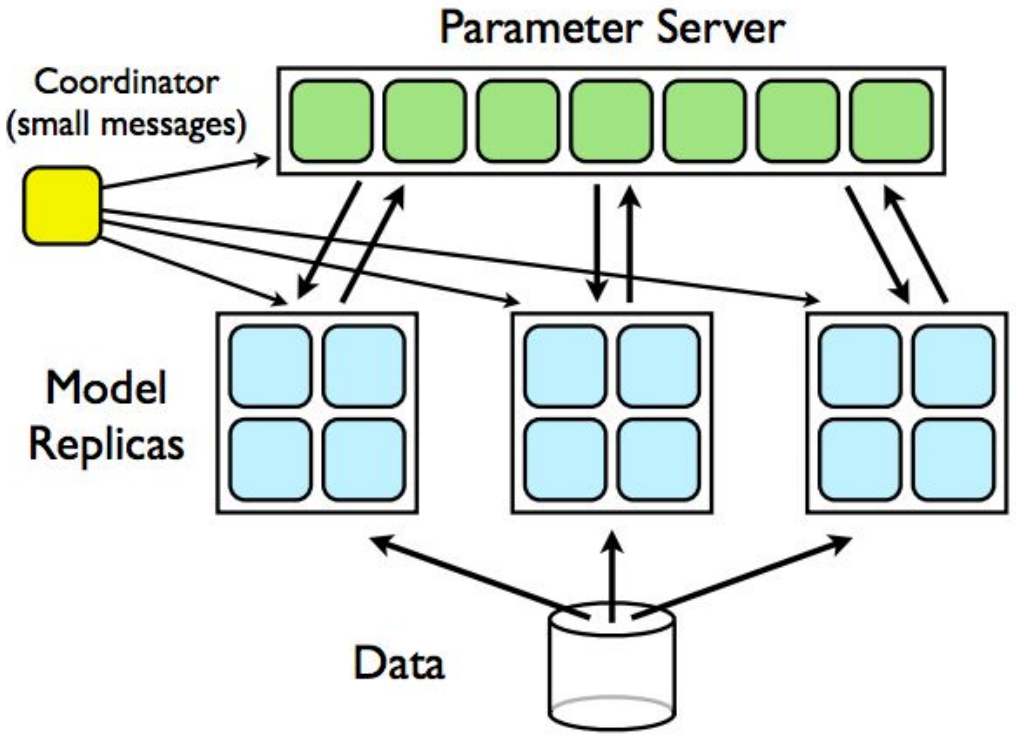
Downpour Stochastic Gradient Descent (SGD)

- Asynchronous SGD
- Uses Adagrad adaptive learning rate
- Asynchronously push gradients and pull parameters
- Replicated models traverse parameters landscape together in parallel
- Dataset divided evenly into shards



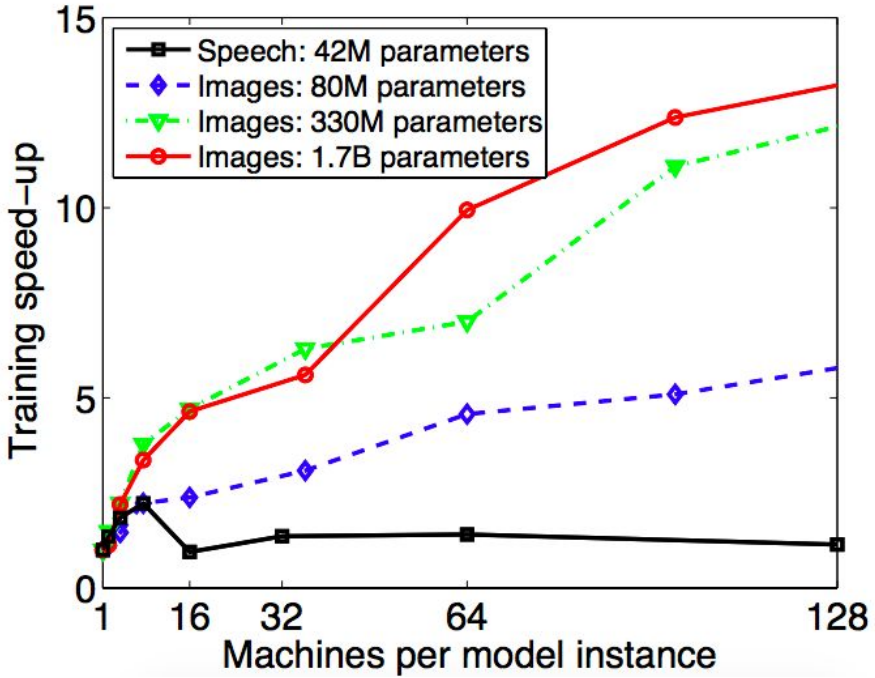
Sandblaster L-BFGS

- Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS)
- Communicate operations instead of values
- **Load balancing** - assign model replicas < 1/N of batch
  - Data parallelism

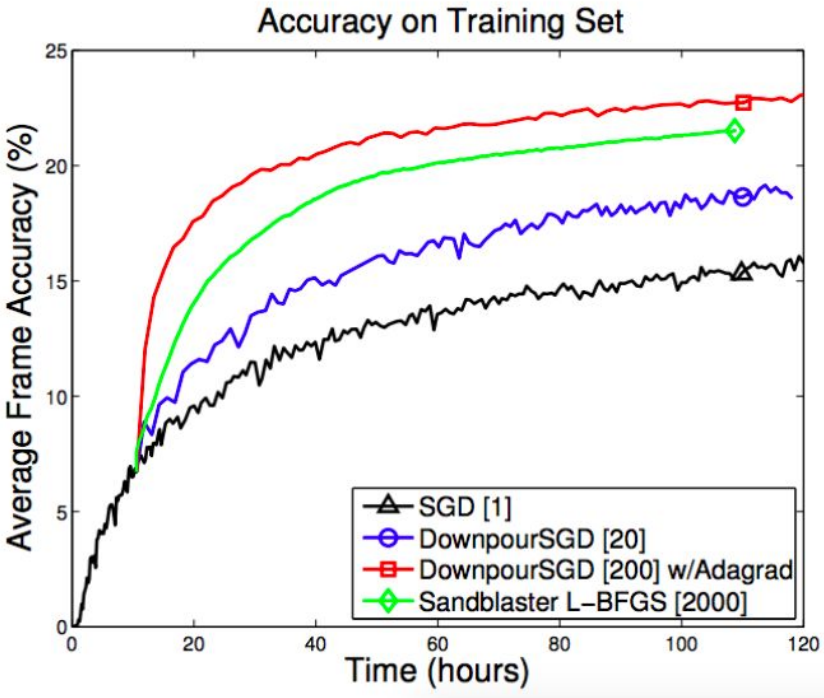


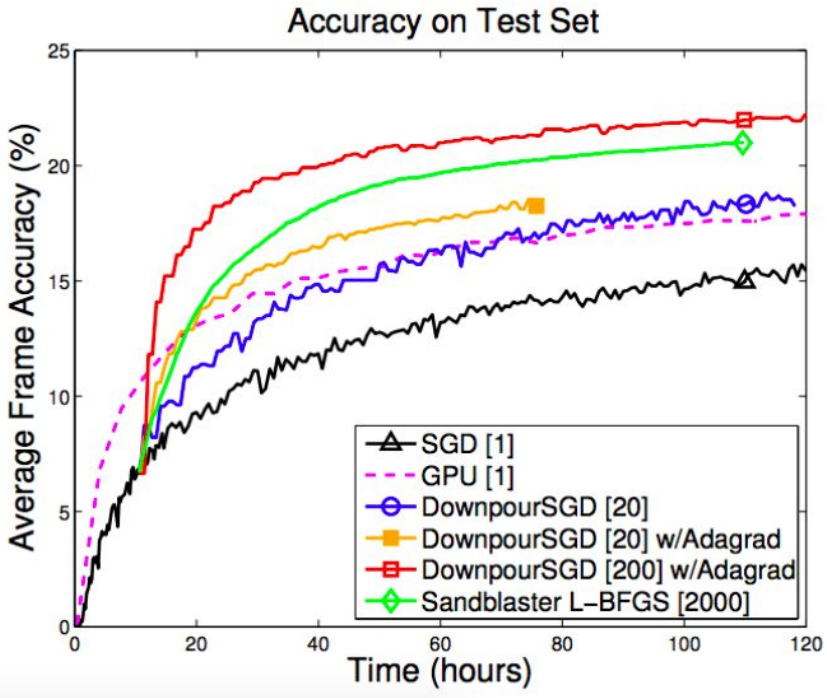
Machines per Model Instance

- Convnet for ImageNet
- RNN for speech
- Local connectivity - models that are not densely connected are more amenable for distributed training

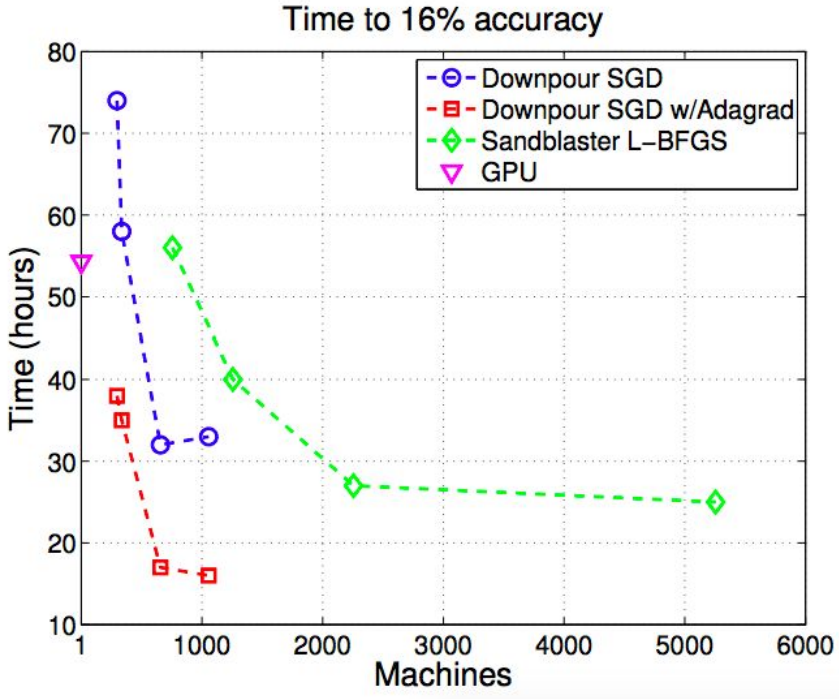


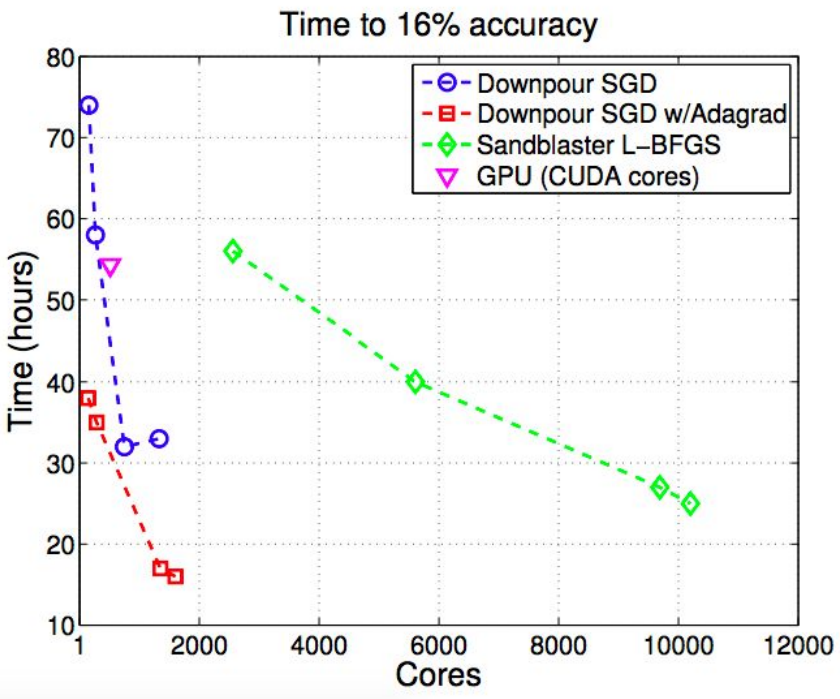
Accuracy over Time



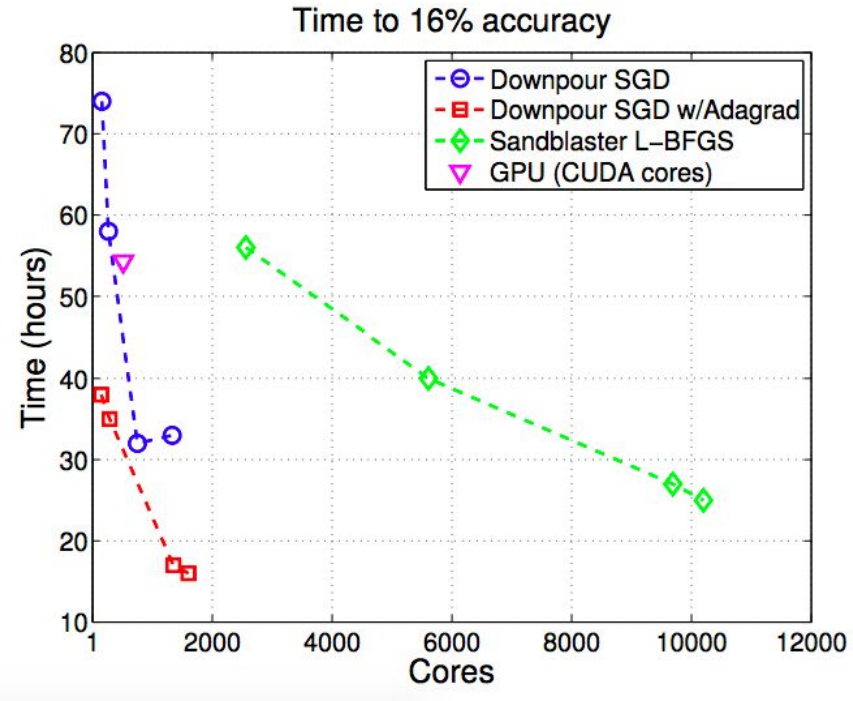


Training Time per Machine





Training Time per Cores



References and Further Reading

1. Dean, Jeffrey, et al. "**Large scale distributed deep networks**." Advances in neural information processing systems. 2012.
2. Sergeev, Alex, et al. "**Meet Horovod: Uber's Open Source Distributed Deep Learning Framework for TensorFlow**." https://eng.uber.com/horovod/ (2017).
3. Black, Alex, et al. "**Distributed Deep Learning, Part 1: An Introduction to Distributed Training of Neural Networks**. "http://engineering.skymind.io/distributed-deep-learning-part-1-an-introduction-to-distributed-training-of-neural-networks (2017).
4. **Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training**
5. Micikevicius, Paulius, et al. "**Mixed Precision Training**." arXiv preprint arXiv:1710.03740 (2017).