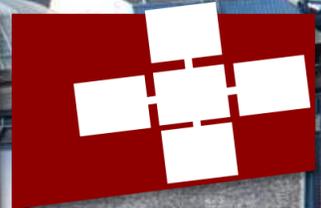


Torsten Hoefler

Deep500: An HPC Deep Learning Benchmark and Competition

Birds of a Feather, SC18, Nov. 2018, Dallas, TX



EuroMPI'19

September 11-13 2019

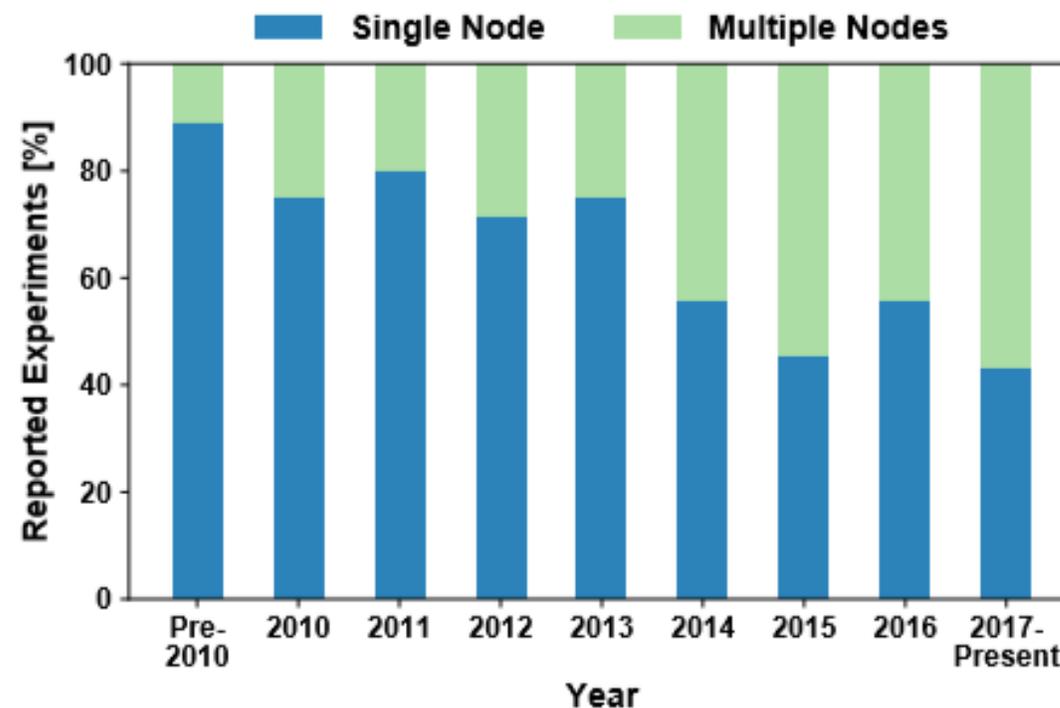
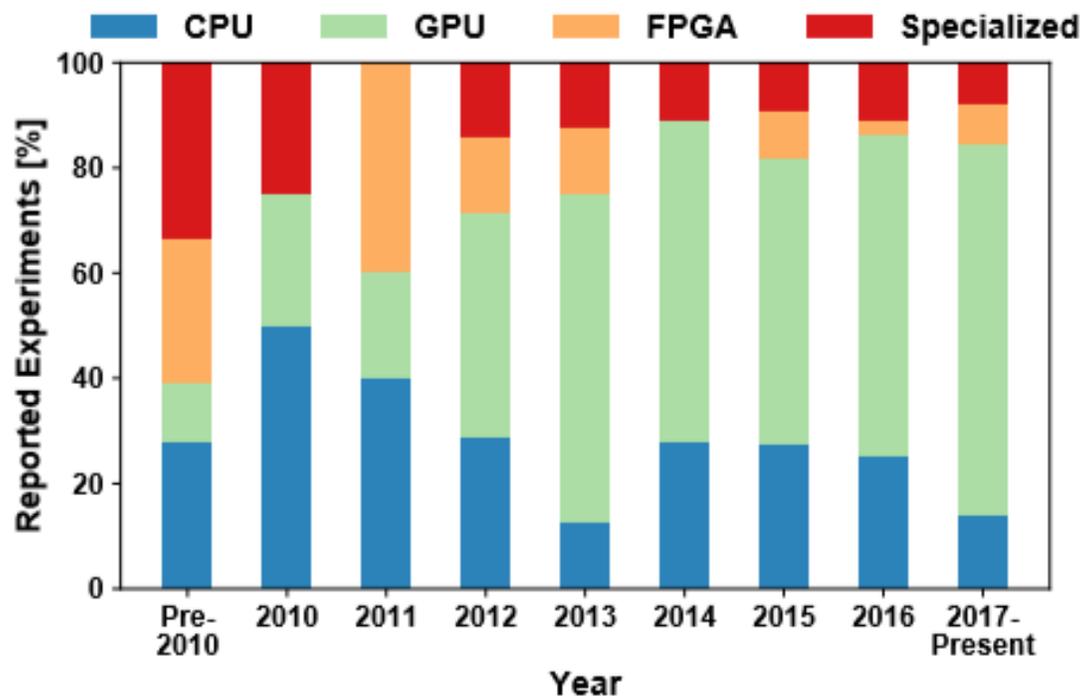
Zurich, Switzerland

<https://eurompi19.inf.ethz.ch>

Submit papers by April 15th!

Trends in deep learning: hardware and multi-node

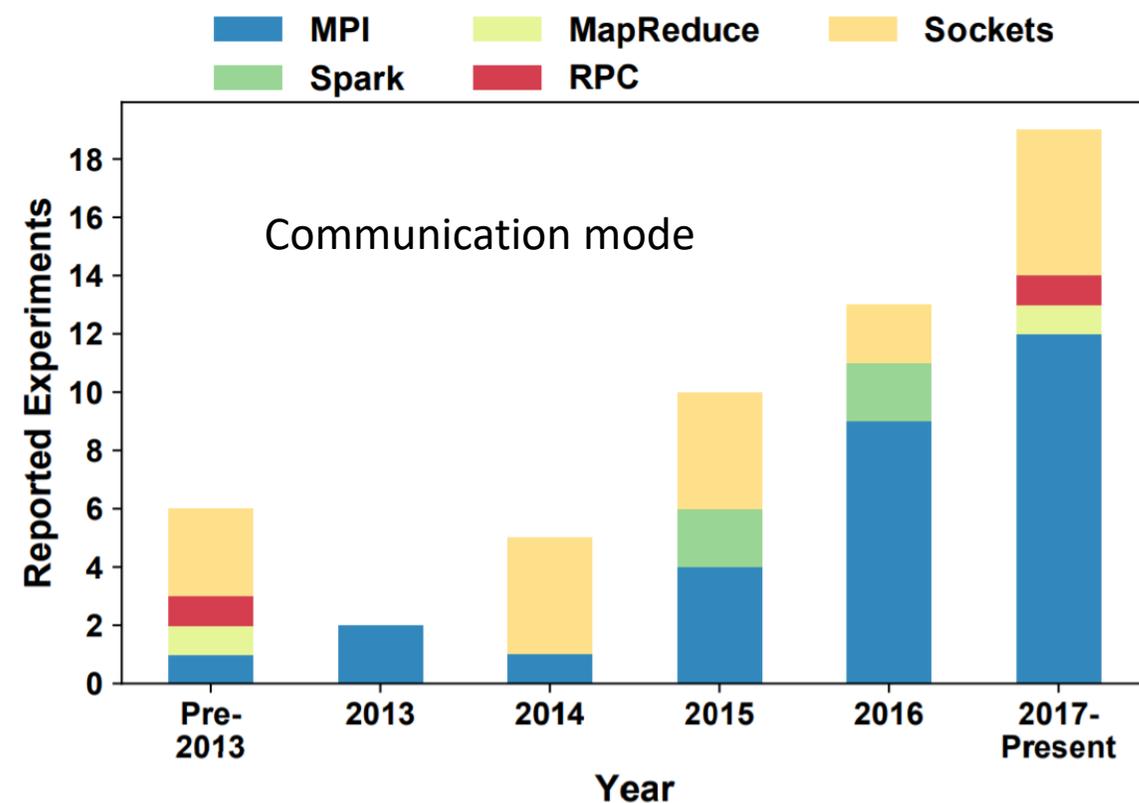
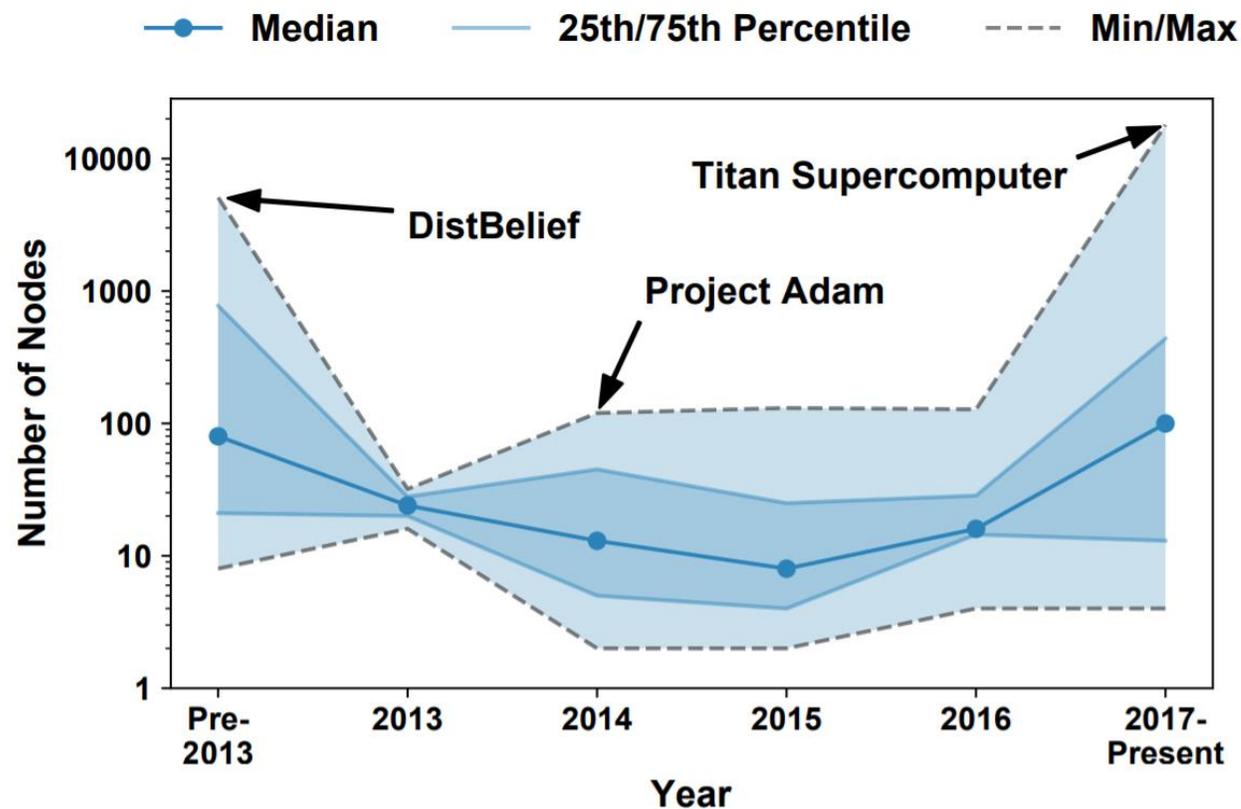
The field is moving fast – trying everything imaginable – survey results from 240 papers in the area of parallel deep learning



Deep Learning is largely on distributed memory today!

Trends in **distributed** deep learning: node count and communication

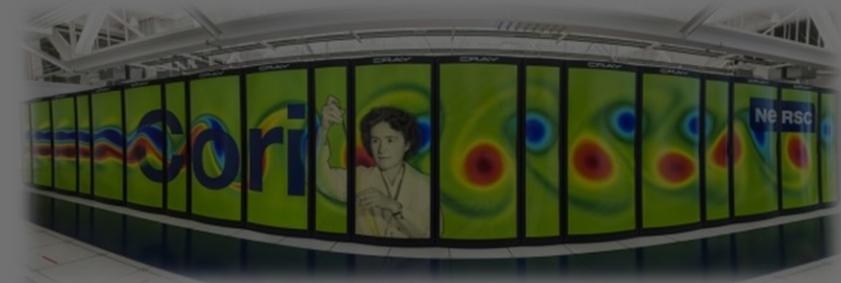
The field is moving fast – trying everything imaginable – survey results from 240 papers in the area of parallel deep learning



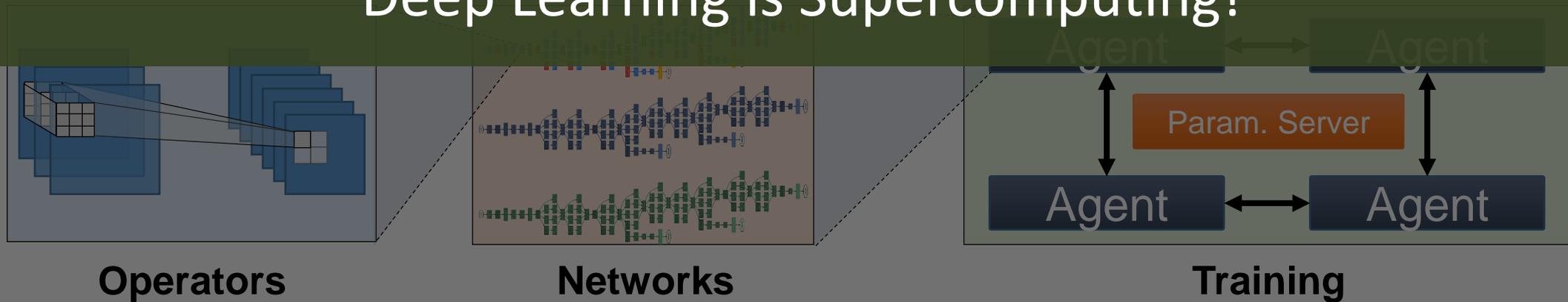
Deep Learning research is converging to MPI!

Parallelism in Deep Learning

- Individual operators
- Network parallelism
- Optimization algorithm
- Distributed training



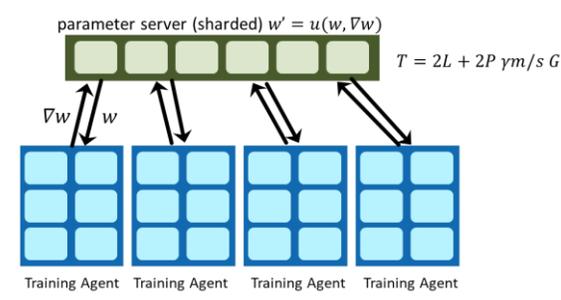
Deep Learning is Supercomputing!



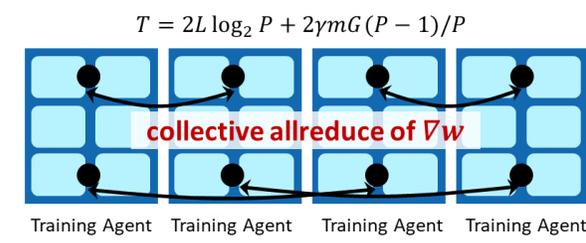
Challenges

Different:

- Communication schemes
 - Model consistency requirements
 - Software stacks and feature sets
- ## Need to define:
- Open datasets from computational sciences
 - Metrics robust to methods (or freeze methods)
 - Standard benchmarking infrastructure



Centralized



Decentralized



System	Operators		Networks			Training			Dist. Training				
	Sta	Cus	Def	Eag	Com	Tra	Dat	Opt	Cus	PS	Dec	Asy	Cus
(L) cuDNN	👍	👍	👎	👎	👎	👎	👎	👎	👎	👎	👎	👎	👎
(L) MKL-DNN	👍	👍	👎	👎	👎	👎	👎	👎	👎	👎	👎	👎	👎
(F) TensorFlow [1]	👍	👍	👍	👍	👍	👍	👍	UR	👍	👍	👍	👍	👍
(F) Caffe, Caffe2 [20]	👍	👍	👍	👍	👍	👍	👍	UR	👍	👍	👍	👍	👍
(F) [Py]Torch [10, 34]	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍
(F) MXNet [6]	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍
(F) CNTK [45]	👍	👍	👍	👍	👍	👍	👍	UR	👍	👍	👍	👍	👍
(F) Theano [4]	👍	👍	👍	👍	👍	👍	👍	👍	👎	👎	👎	👎	👎
(F) Chainer[MN] [43]	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍	👍
(F) Darknet [37]	👍	👍	👍	👍	👍	👍	👍	👍	👎	👎	👎	👎	👎
(F) DL4j [42]	👍	👍	👍	👍	👍	👍	👍	UR	👍	👍	👍	👍	👍
(F) DSSTNE	👍	👍	👍	👍	👍	👍	👍	UR	👍	👎	👎	👎	👎
(F) PaddlePaddle	👍	👍	👍	👍	👍	👍	👍	UR	👍	👍	👍	👍	👍
(F) TVM [7]	👍	👍	👍	👍	👍	👍	👍	👍	👎	👎	👎	👎	👎
(E) Keras [8]	👍	👍	👎	👎	👎	👎	👎	UR	👍	👎	👎	👎	👎
(E) Horovod [41]	👎	👎	👎	👎	👎	👎	👎	👍	👍	👍	👍	👍	👍
(E) TensorLayer [14]	👍	👍	👎	👎	👎	👎	👎	UR	👍	👍	👍	👍	👍
(E) Lasagne	👍	👍	👎	👎	👎	👎	👎	UR	👍	👍	👍	👍	👍
(E) TFLearn [11]	👍	👍	👎	👎	👎	👎	👎	👍	👍	👍	👍	👍	👍

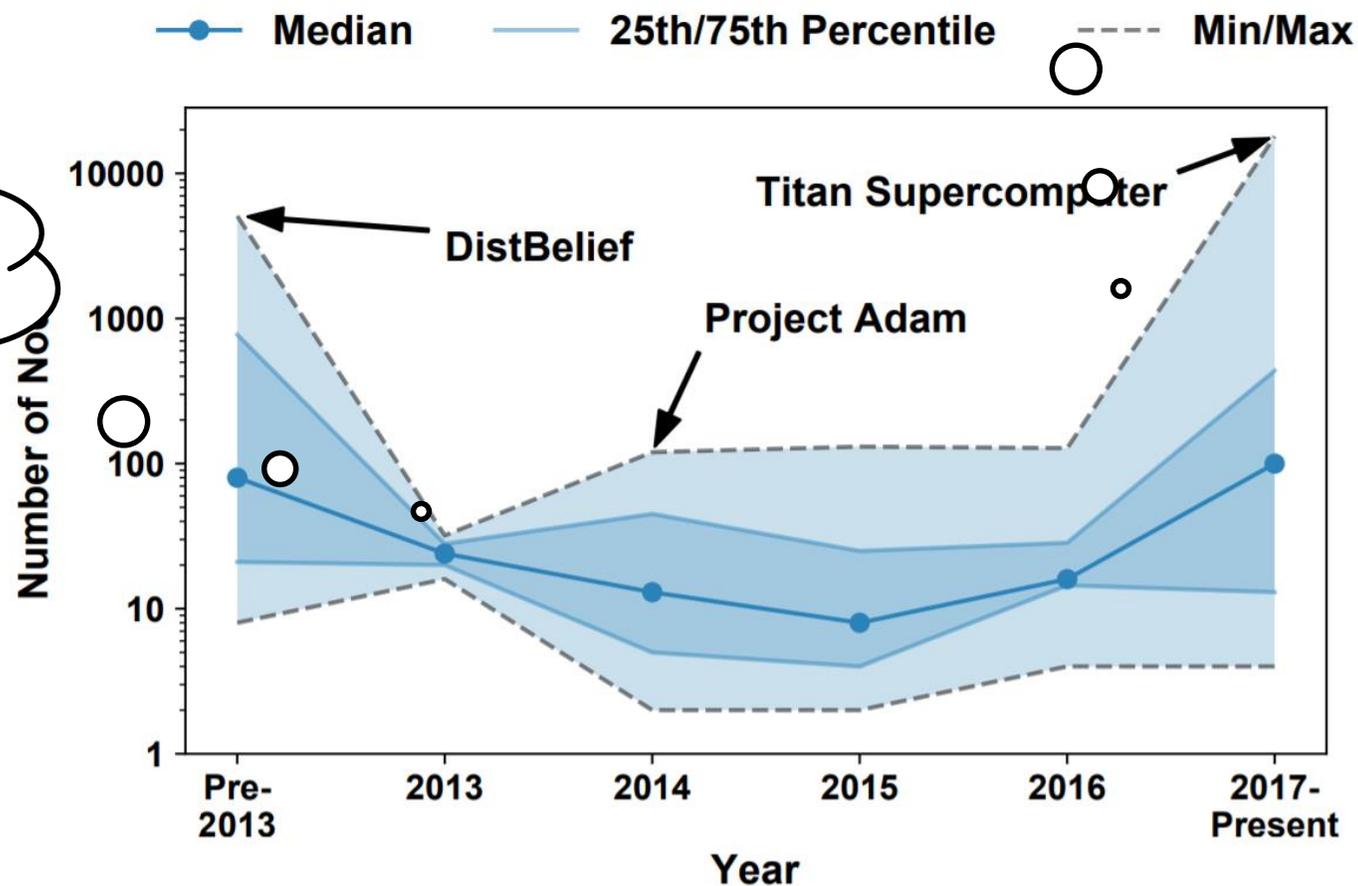
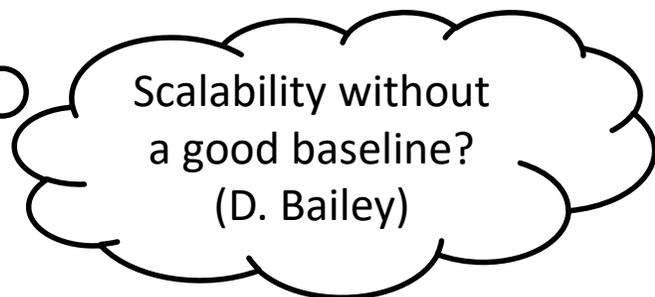
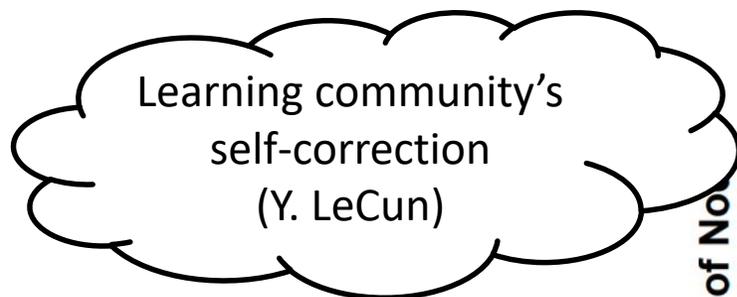
So how to not do this

“Twelve ways to fool the masses when reporting performance of deep learning workloads”
(my humorous guide to floptimize deep learning)



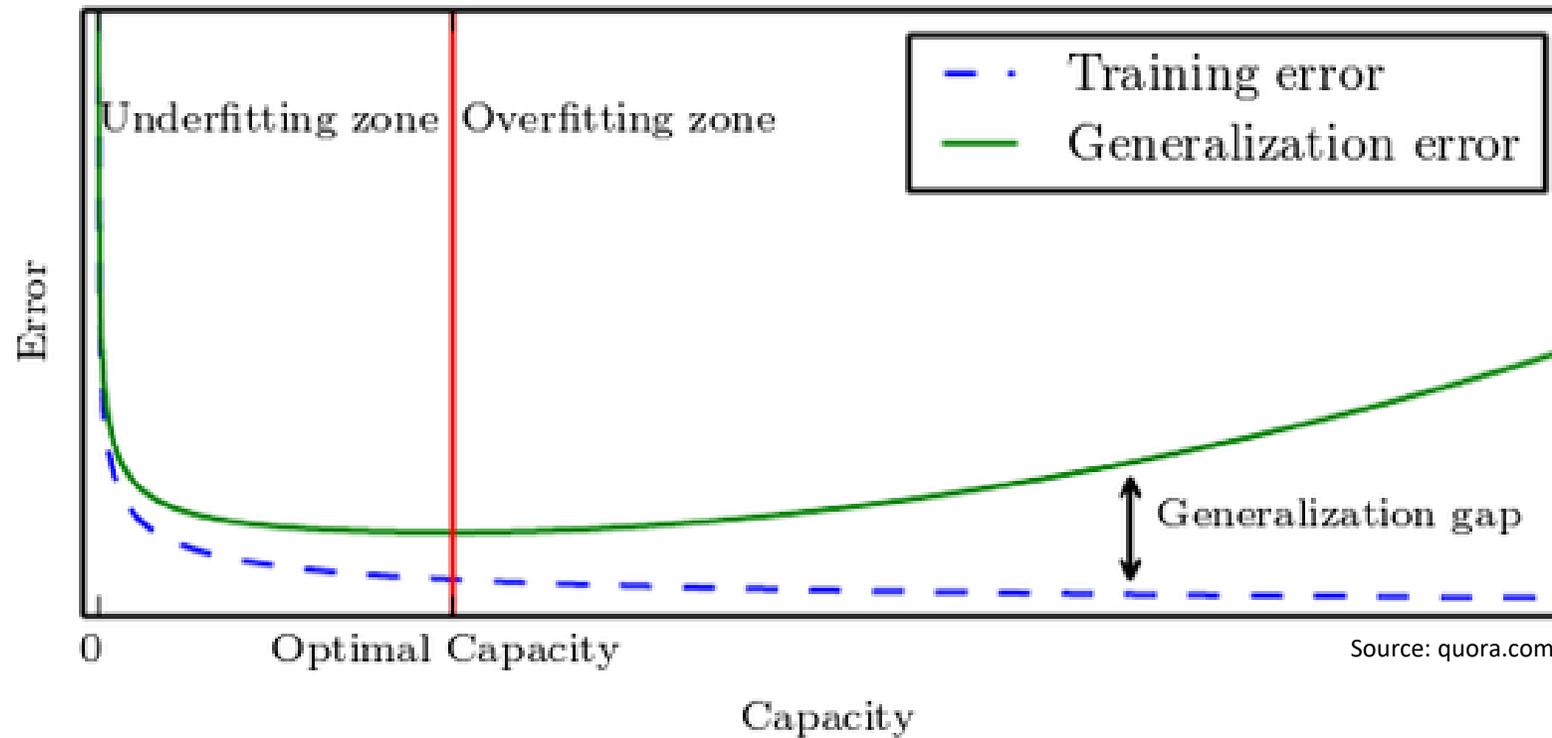
1) Ignore accuracy when scaling up!

- Too obvious for this audience
 - Was very popular in 2015!
- Surprisingly many (still) do this



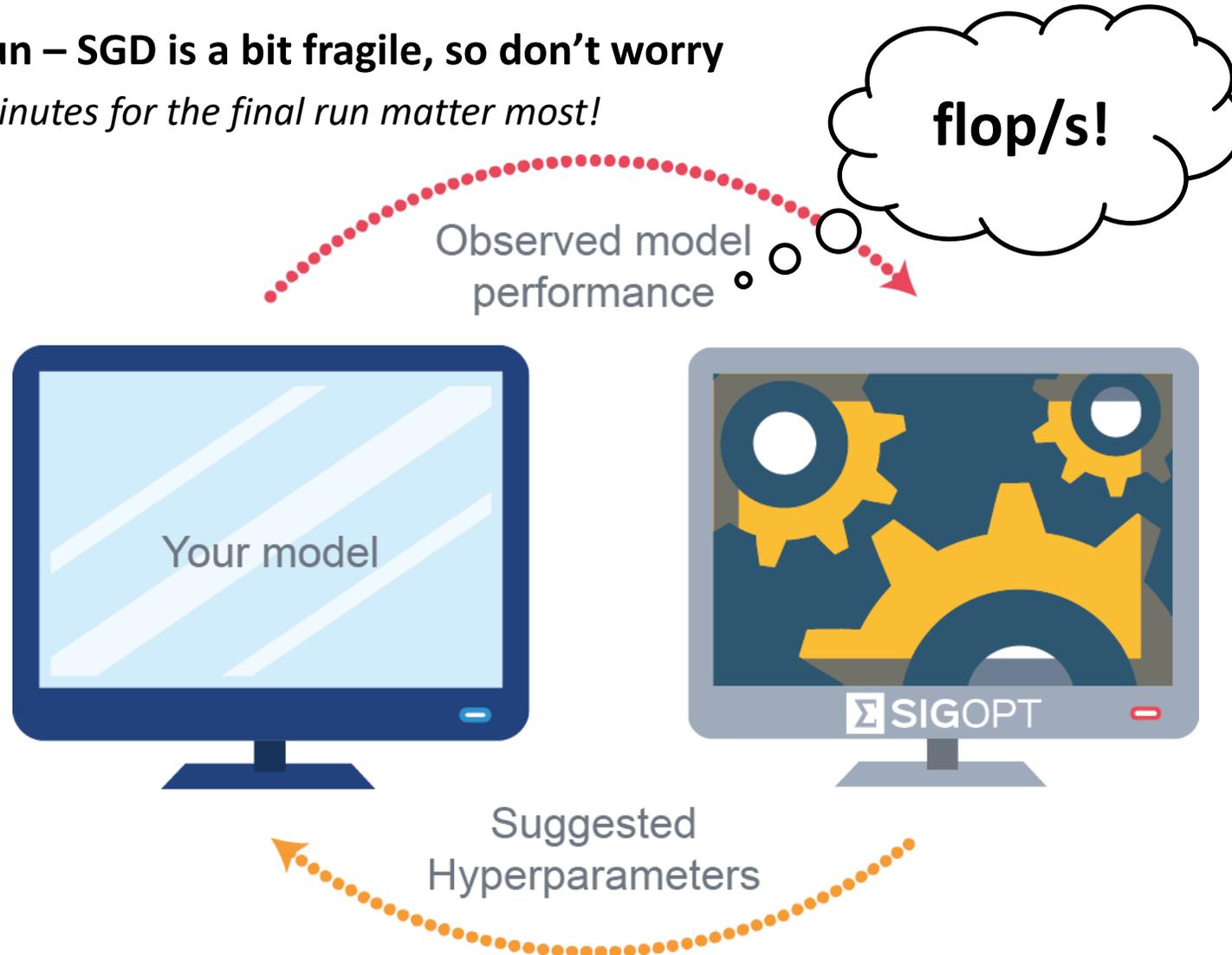
2) Do not report test accuracy!

- Training accuracy is sufficient isn't it?



3) Do not report all training runs needed to tune hyperparameters!

- Report the best run – SGD is a bit fragile, so don't worry
At the end, the minutes for the final run matter most!



So how to not do this

“Twelve ways to fool the masses when reporting performance of deep learning workloads”
(my humorous guide to floptimize deep learning)



Other Results

- SparCML: a sparse reduction protocol to implement faster reductions in parallel systems with sparse input vectors [arXiv'18, NIPS'18]

- Using deep learning to create learnable representations of code [NIPS'18]
 - State of the art in predicting fastest hardware mapping and algorithm classification

- Accelerating convolution operators using micro-batches [Cluster'18]
 - Key technique: Use ILP and Dynamic Programming

- Parallelism modeling of deep learning, from operator to distributed training on supercomputers [arXiv'18]

SPARCML: High-Performance Sparse Communication for Machine Learning

Neural Code Comprehension: A Learnable Representation of Code Semantics

Accelerating Deep Learning Frameworks with Micro-batches

arXiv:1802.09941v2 [cs.LG] 15 Sep 2018

Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis

TAL BEN-NUN and TORSTEN HOEFLER, ETH Zurich, Switzerland

Deep Neural Networks (DNNs) are becoming an important tool in modern computing applications. Accelerating their training is a major challenge and techniques range from distributed algorithms to low-level circuit design. In this survey, we describe the problem from a theoretical perspective, followed by approaches for its parallelization. We present trends in DNN architectures and the resulting implications on parallelization strategies. We then review and model the different types of concurrency in DNNs: from the single operator, through parallelism in network inference and training, to distributed deep learning. We discuss asynchronous stochastic optimization, distributed system architectures, communication schemes, and neural architecture search. Based on those approaches, we extrapolate potential directions for parallelism in deep learning.

CCS Concepts • General and reference → Surveys and overviews; • Computing methodologies → Neural networks; Parallel computing methodologies; Distributed computing methodologies

Additional Key Words and Phrases: Deep Learning, Distributed Computing, Parallel Algorithms

ACM Reference Format: Tal Ben-Nun and Torsten Hoefler. 2018. Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis. 47 pages.

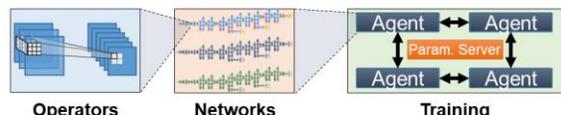
1 INTRODUCTION

Machine Learning, and in particular Deep Learning^[14], is rapidly taking over a variety of aspects in our daily lives. At the core of deep learning lies the Deep Neural Network (DNN), a construct inspired by the interconnected nature of the human brain. Trained properly, the expressiveness of DNNs provides accurate solutions for problems previously thought to be unsolvable, merely by observing large amounts of data. Deep learning has been successfully implemented for a plethora of fields, ranging from image classification^[10], through speech recognition^[6] and medical diagnosis^[4], to autonomous driving^[22] and defeating human players in complex games^[21].

Since the 1980s, neural networks have attracted the attention of the machine learning community^[14]. However, DNNs' rise into prominence was tightly coupled to the available computational power, which allowed to exploit their inherent parallelism. Consequently, deep learning managed to outperform all existing approaches in speech recognition^[14] and image classification^[10], where the latter (AlexNet) increased the accuracy by a factor of two, sparking interest outside of the community and even academia.

As datasets increase in size and DNNs in complexity, the computational intensity and memory demands of deep learning increase proportionally. Training a DNN to competitive accuracy today

Authors' address: Tal Ben-Nun, talbn@inf.ethz.ch, Torsten Hoefler, htorg@inf.ethz.ch, ETH Zurich, Department of Computer Science, Zurich, 8006, Switzerland.



Operators
Networks
Training