
facebook

Distributed DeepLearning at Scale

Soumith Chintala

Facebook AI Research

Overview

- Deep Learning Research at FAIR
- Deep Learning on GPUs
- Deep Learning at scale
- Emerging Trends

Deep Learning Research at Facebook AI Research

Image Intelligence: Classification

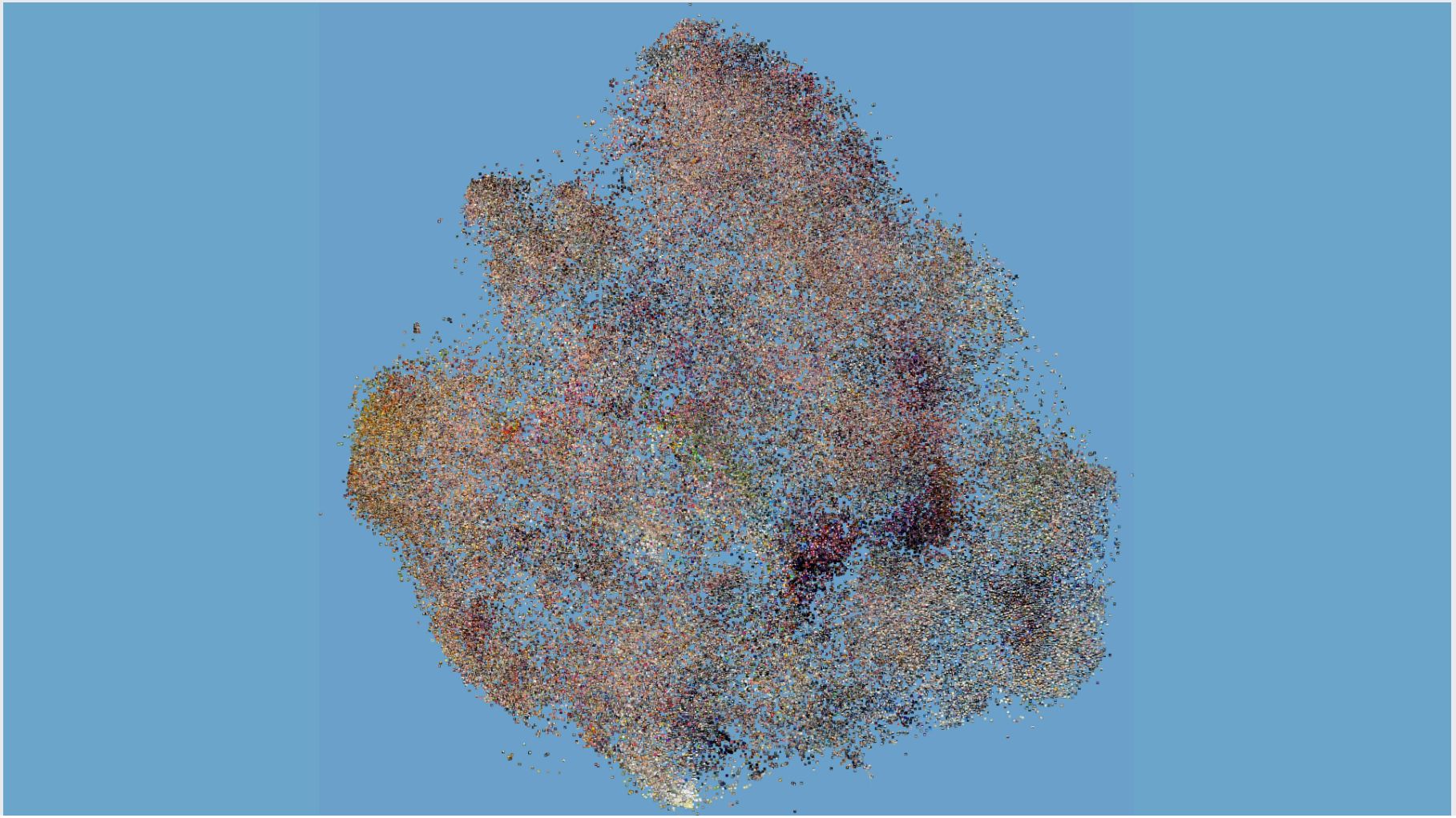


Image Intelligence

Language Translation from Visual Learning

English	French	English	French
oas	oea	uzbekistan	ouzbekistan
infrared	infrarouge	mushroom	champignons
tomatoes	tomates	filmed	serveur
bookshop	librairie	mauritania	mauritanie
server	apocalyptique	pencils	crayons

Image Intelligence : Detection

Image Intelligence : Detection

Image Intelligence : Detection

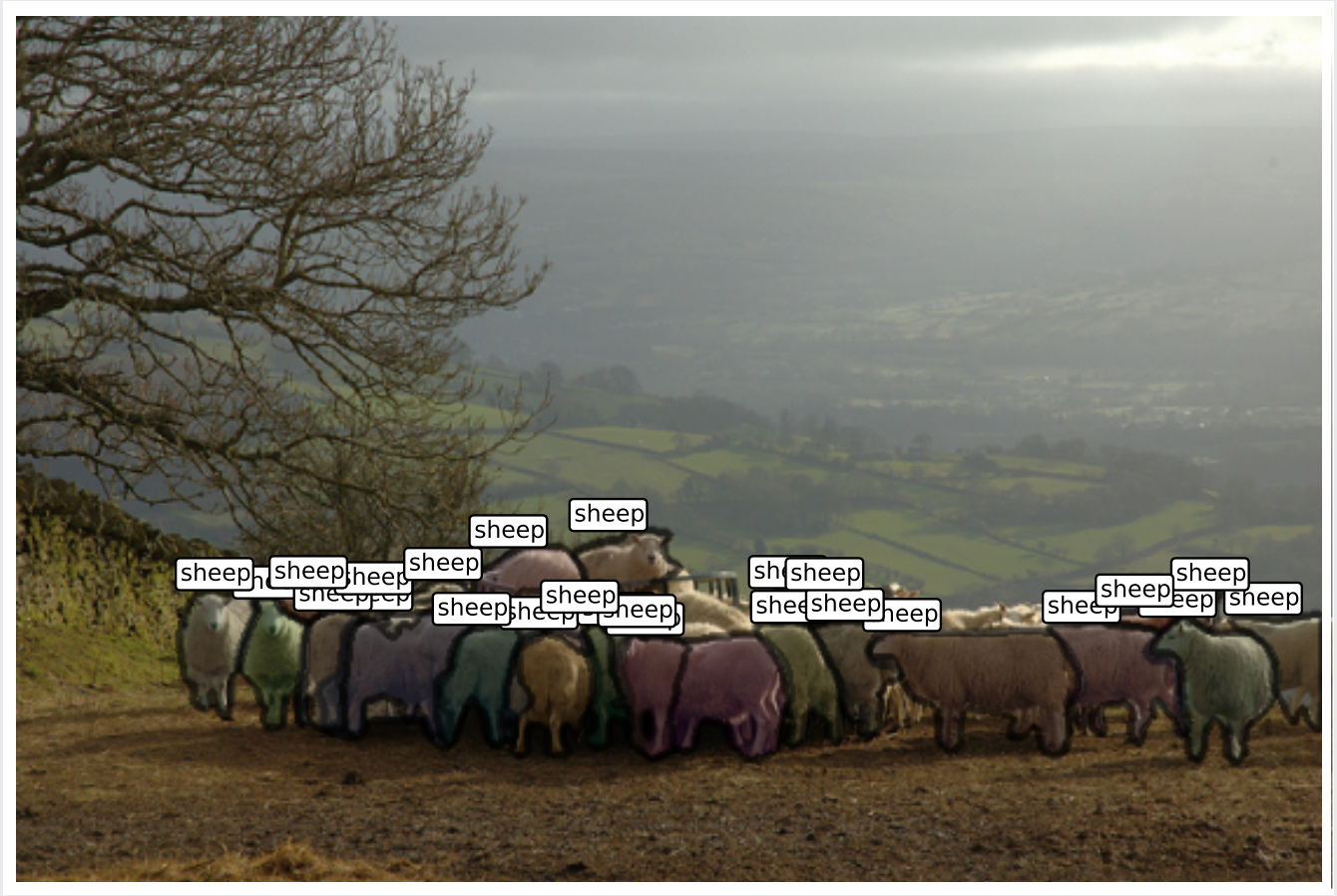


Image Intelligence : Detection

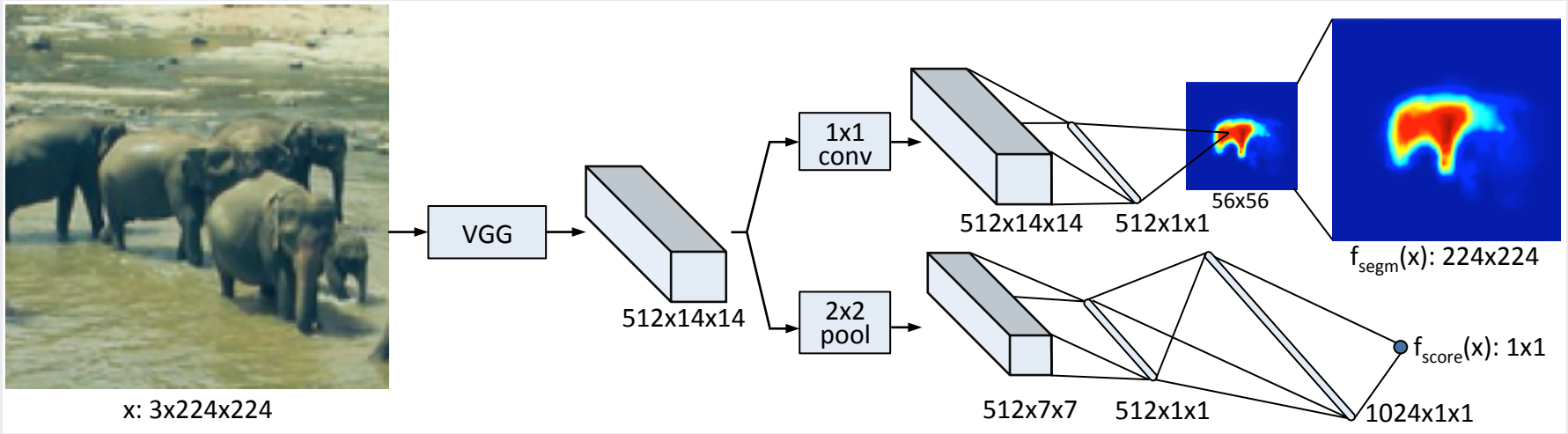
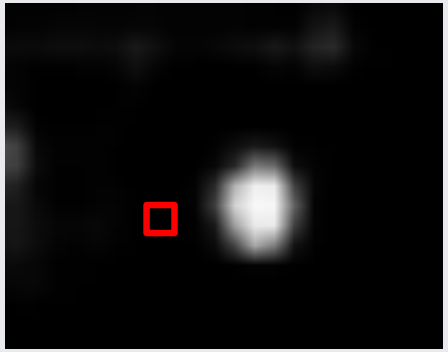


Image Intelligence : Detection

image

scores



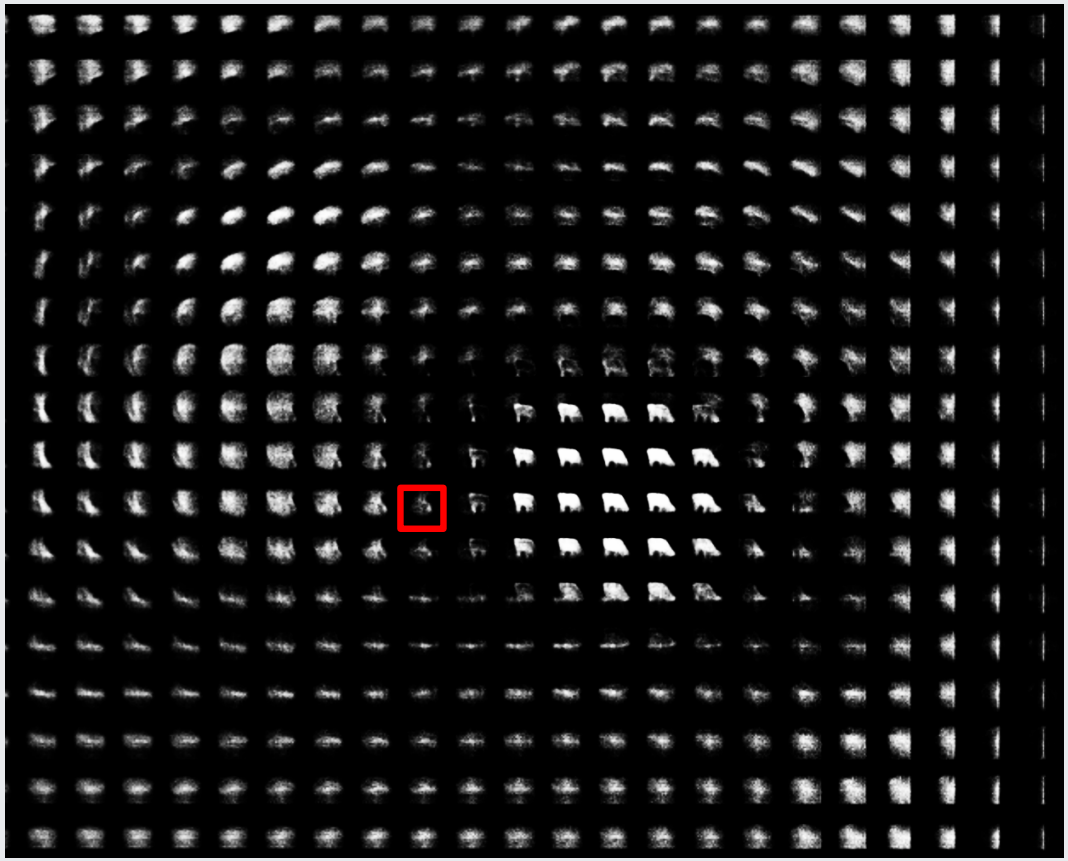
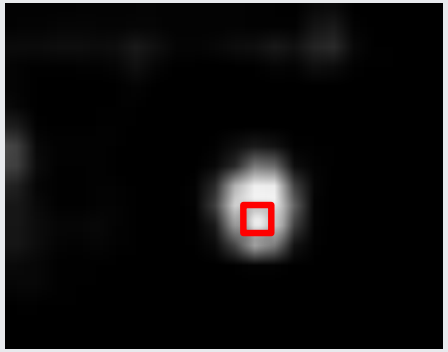


Image Intelligence : Detection

image

scores



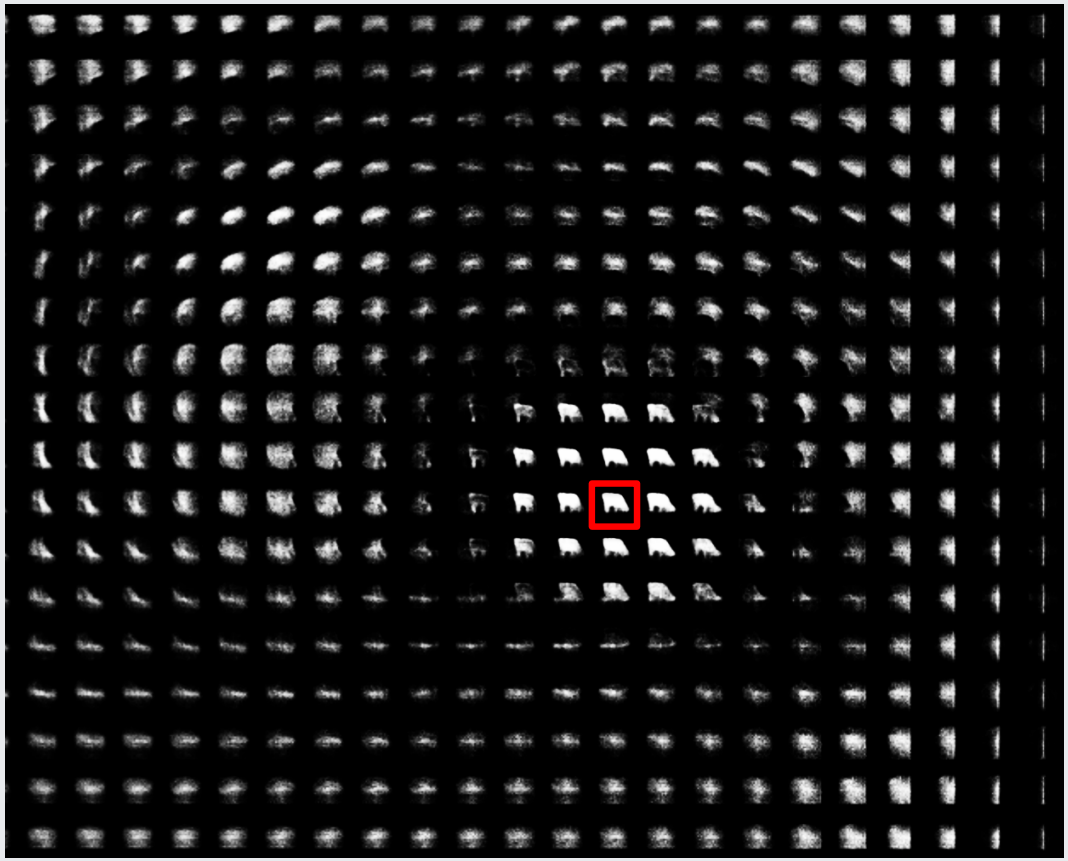


Image Intelligence : Detection

Image Intelligence

https://code.facebook.com/posts/accessibility/

Video Intelligence

Image and Video Generation

Predicting the Future

Natural Language Understanding

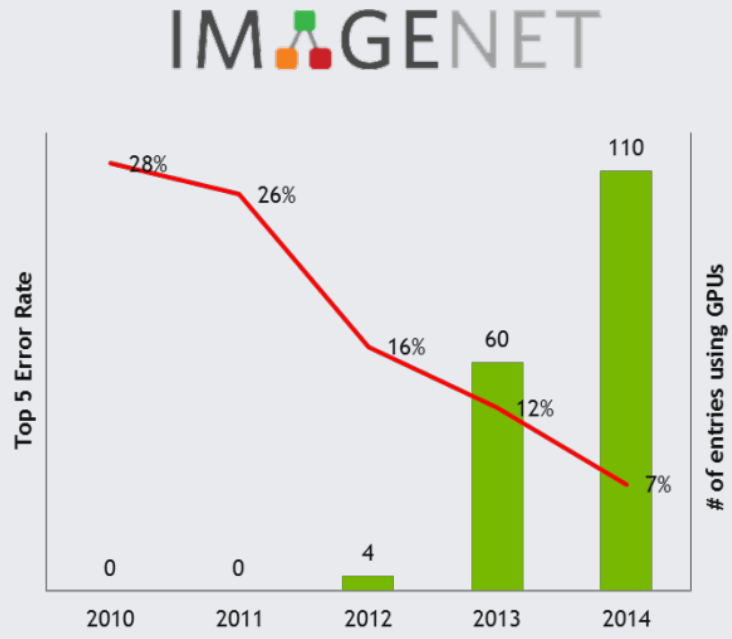
chatbots, personal assistants

- Memory networks
- Language Translation
- Reading, Writing and answering Questions

Deep Learning at Scale

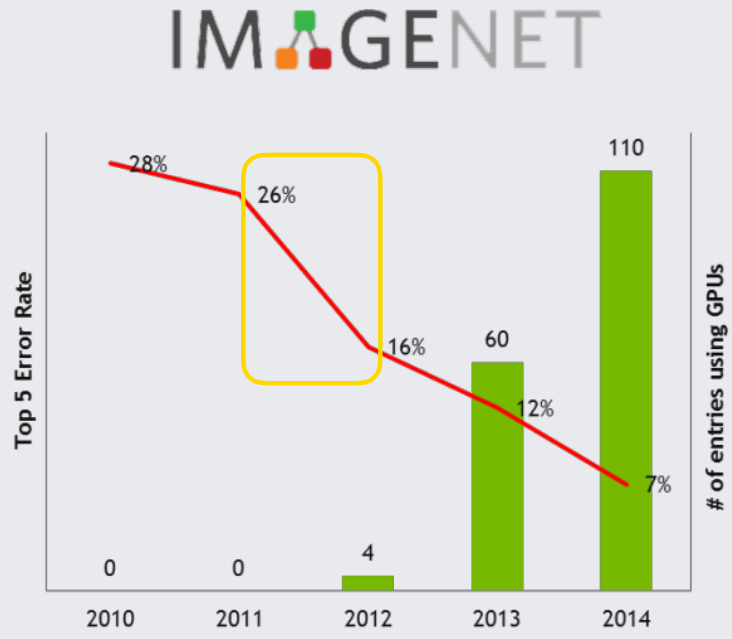
Deep Learning at Scale

GPU-powered Convolution Neural Networks



Deep Learning at Scale

GPU-powered Convolution Neural Networks



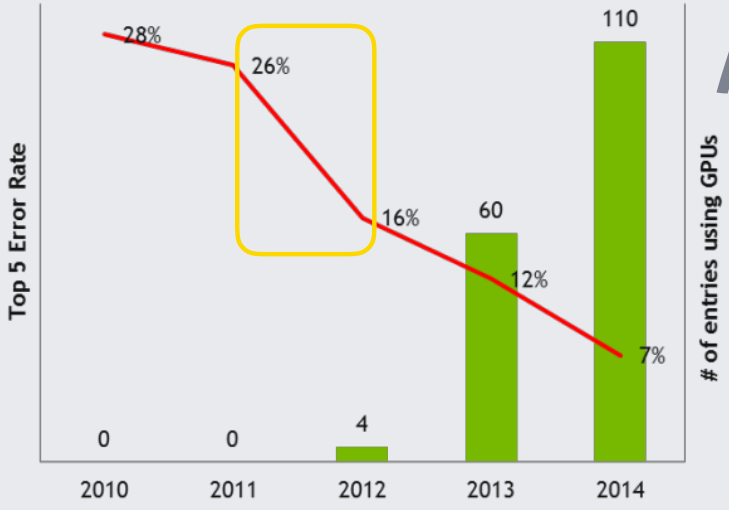
Deep Learning at Scale

GPU-powered Convolution Neural Networks

IMAGENET

cuda-convnet

High-performance C++/CUDA implementation of convolutional neural networks



Alex Khrizevsky

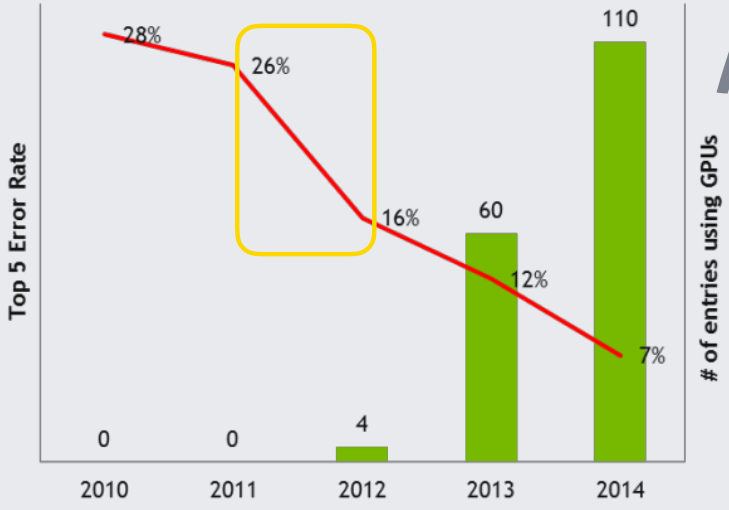
Deep Learning at Scale

GPU-powered Convolution Neural Networks

IMAGENET

cuda-convnet

High-performance C++/CUDA implementation of convolutional neural networks



Alex Khrizevsky

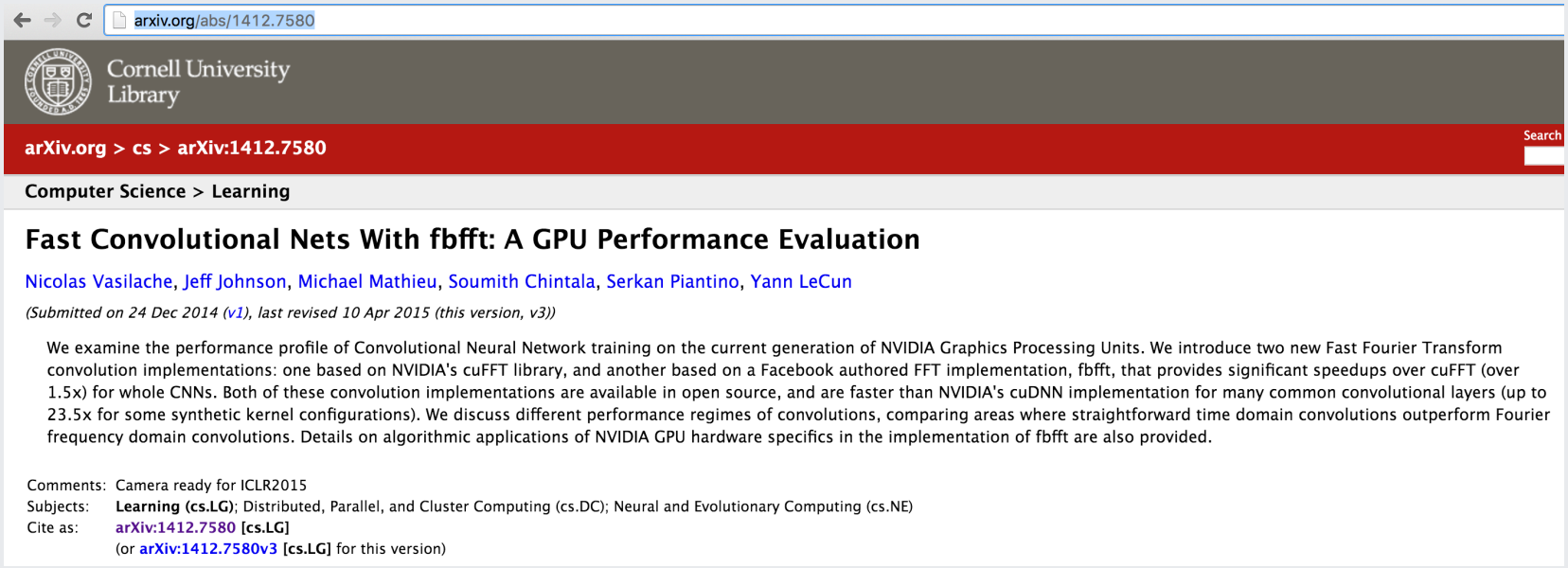
Deep Learning at Scale

GPU-powered Convolution Neural Networks

- Convolutions, GEMM take all the time
  - Faster Convolutions = faster research

Deep Learning at Scale

GPU-powered Convolution Neural Networks



arxiv.org/abs/1412.7580

Cornell University Library

arXiv.org > cs > arXiv:1412.7580

Search

Computer Science > Learning

Fast Convolutional Nets With fbfft: A GPU Performance Evaluation

Nicolas Vasilache, Jeff Johnson, Michael Mathieu, Soumith Chintala, Serkan Piantino, Yann LeCun

(Submitted on 24 Dec 2014 (v1), last revised 10 Apr 2015 (this version, v3))

We examine the performance profile of Convolutional Neural Network training on the current generation of NVIDIA Graphics Processing Units. We introduce two new Fast Fourier Transform convolution implementations: one based on NVIDIA's cuFFT library, and another based on a Facebook authored FFT implementation, fbfft, that provides significant speedups over cuFFT (over 1.5x) for whole CNNs. Both of these convolution implementations are available in open source, and are faster than NVIDIA's cuDNN implementation for many common convolutional layers (up to 23.5x for some synthetic kernel configurations). We discuss different performance regimes of convolutions, comparing areas where straightforward time domain convolutions outperform Fourier frequency domain convolutions. Details on algorithmic applications of NVIDIA GPU hardware specifics in the implementation of fbfft are also provided.

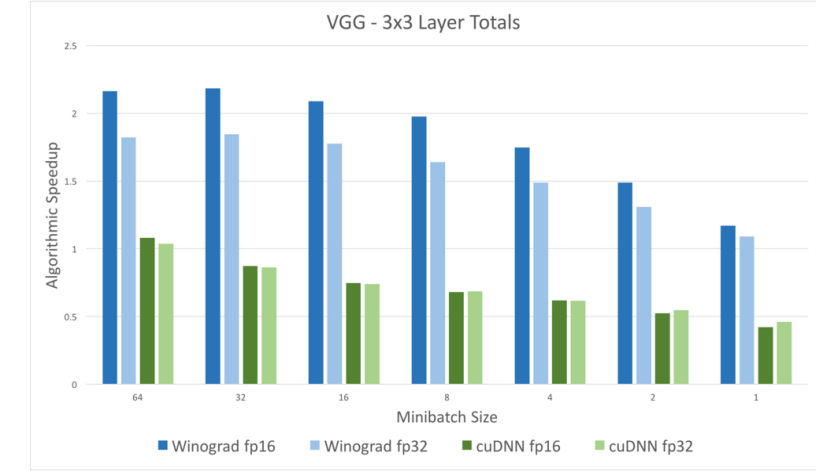
Comments: Camera ready for ICLR2015
Subjects: **Learning (cs.LG)**; Distributed, Parallel, and Cluster Computing (cs.DC); Neural and Evolutionary Computing (cs.NE)
Cite as: **arXiv:1412.7580 [cs.LG]**
(or **arXiv:1412.7580v3 [cs.LG]** for this version)

Deep Learning at Scale

GPU-powered Convolution Neural Networks

Winograd transform based Convolutions

nervana



Deep Learning at Scale

GPU-powered Convolution Neural Networks

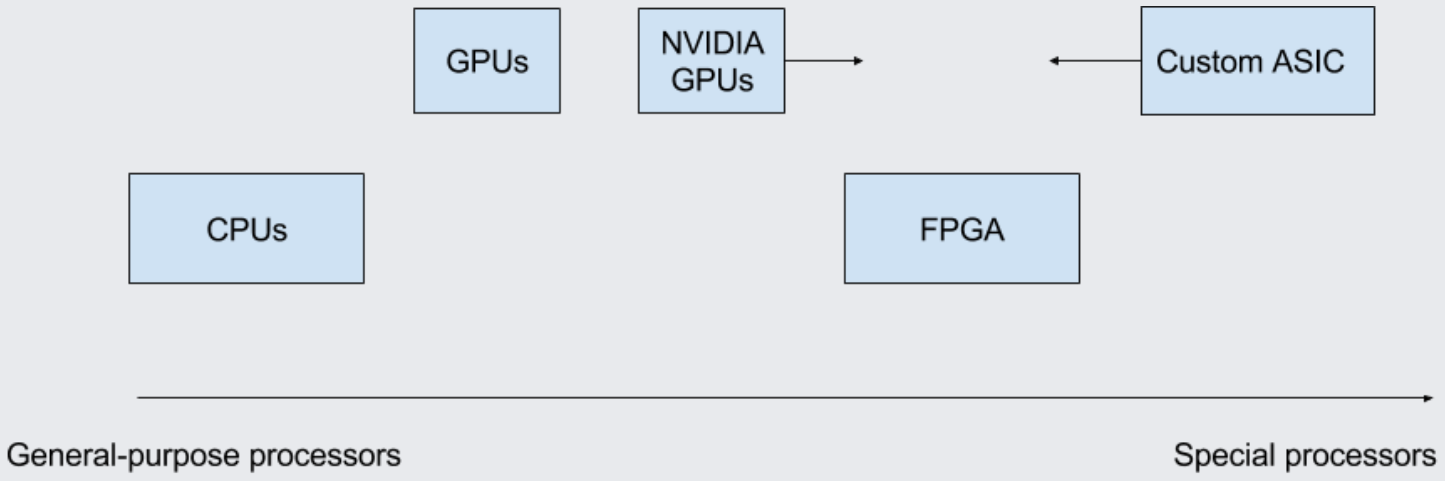
- The standard in deep learning:

NVIDIA GPUs + CUDA + CuDNN

Deep Learning at Scale

GPU-powered Convolution Neural Networks

- Exotic new hardware!
  - Custom chips (Yunji Chen et. al., Nervana Systems)



Deep Learning at Scale

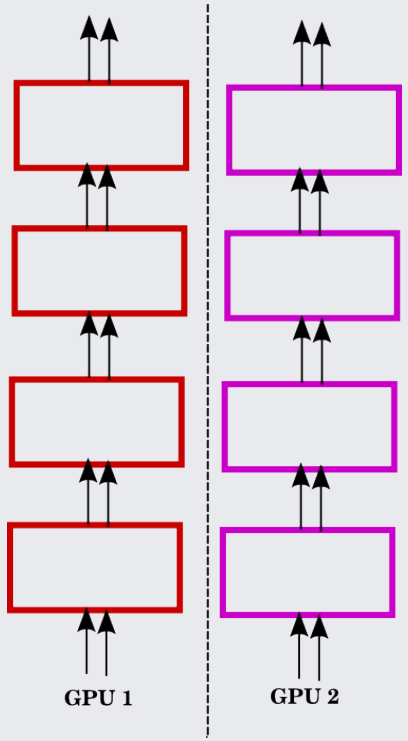
Multi-GPU Training

- Use multiple GPUs on single machine

Deep Learning at Scale

Multi-GPU Training

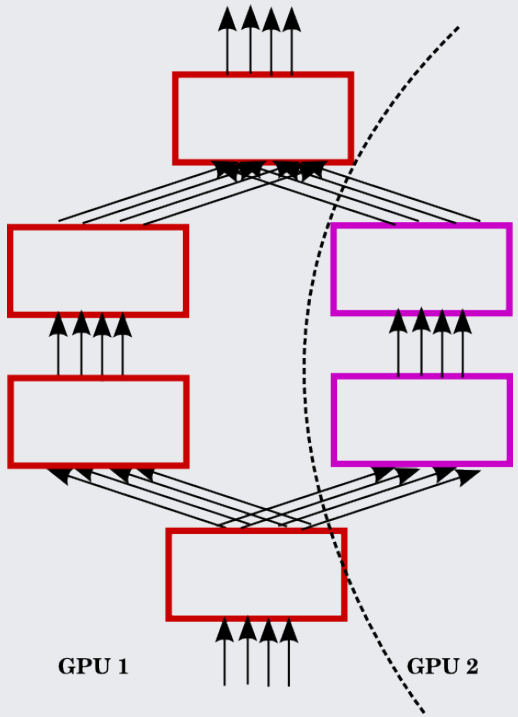
- Data parallel



Deep Learning at Scale

Multi-GPU Training

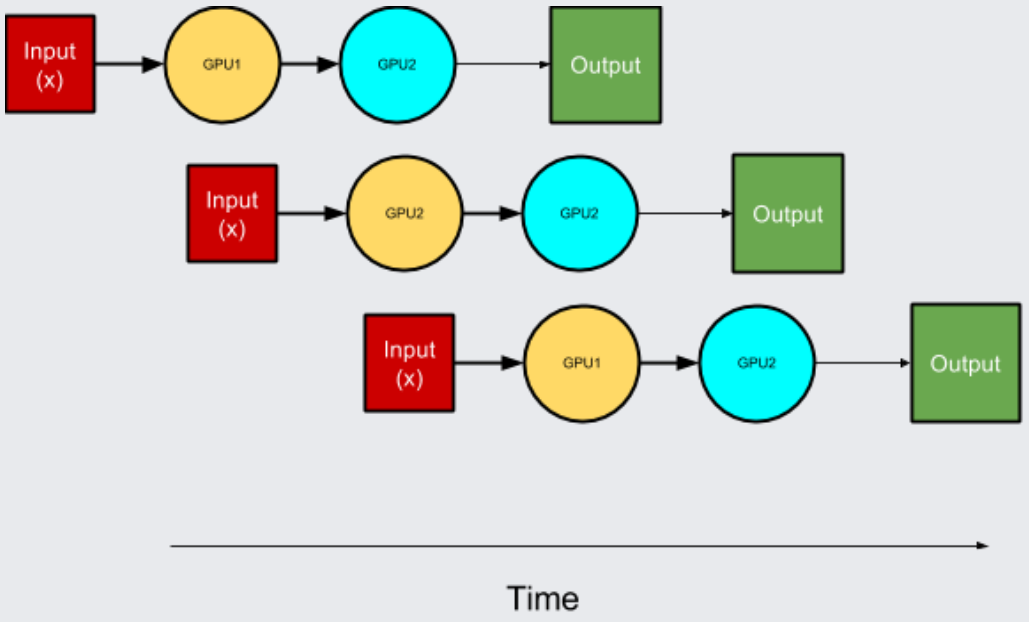
- Model parallel



Deep Learning at Scale

Multi-GPU Training

- Pipeline-parallel



Deep Learning at Scale

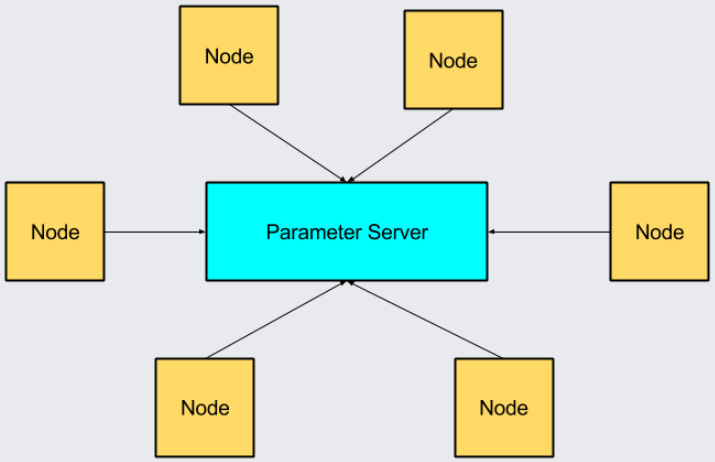
Multi-GPU Training

Bottleneck: interconnects

Deep Learning at Scale

Multi-Machine Training

- Multi-machine SGD

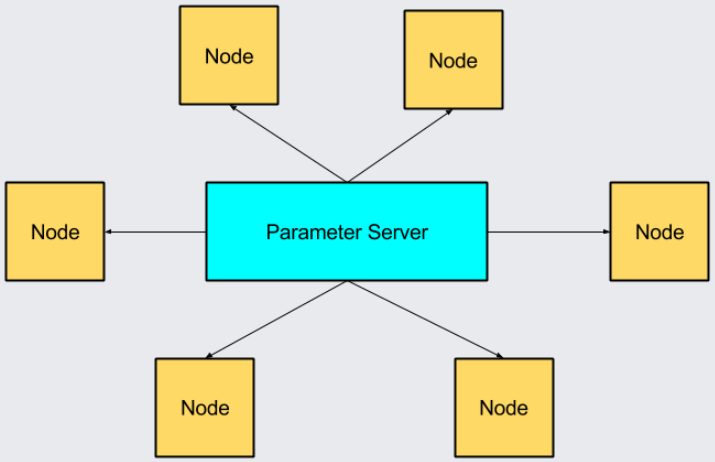


Send gradients

Deep Learning at Scale

Multi-Machine Training

- Multi-machine SGD

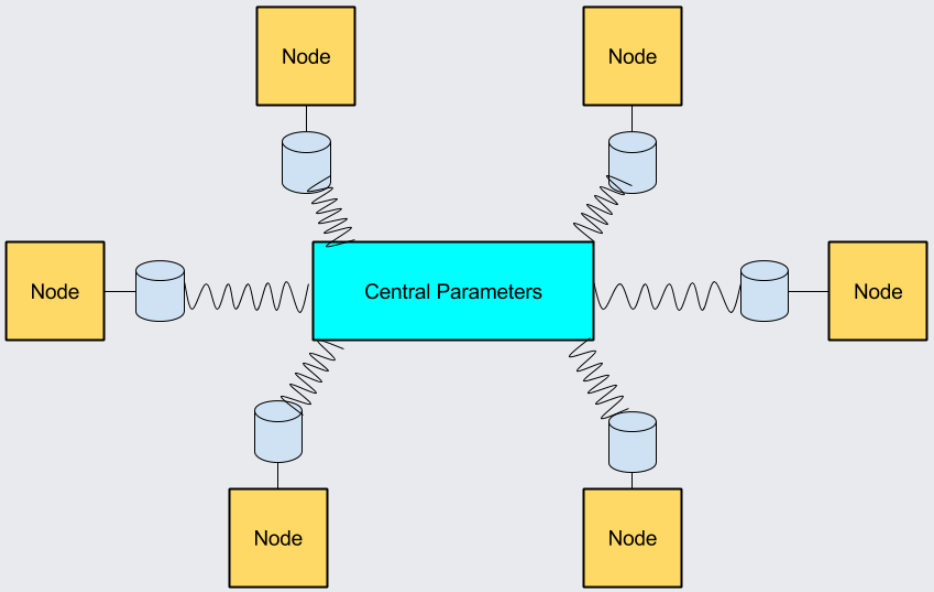


Send Weights

Deep Learning at Scale

Multi-Machine Training

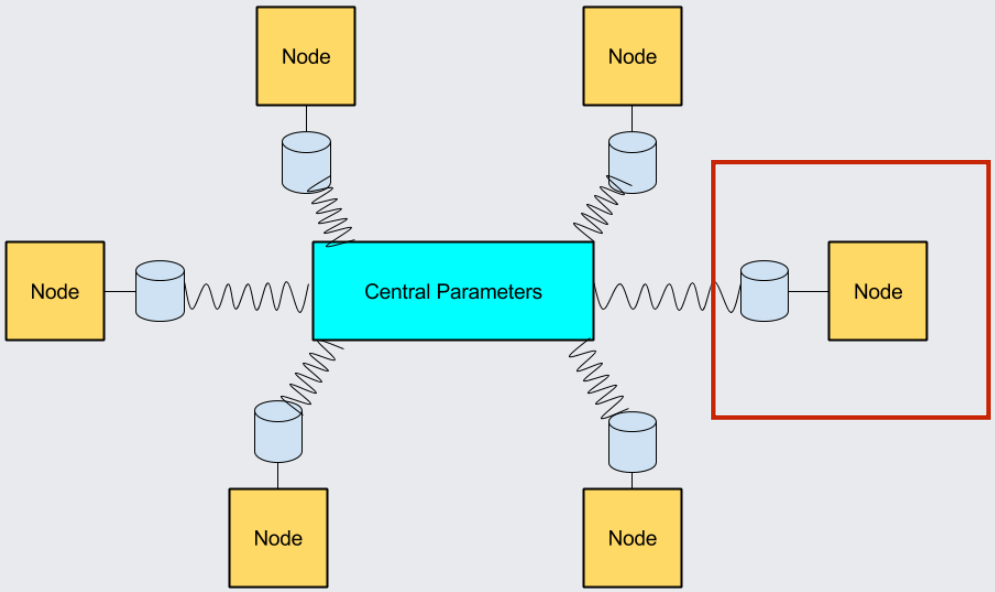
- Elastic Averaging SGD! (Sixin Zhang, Anna Choromanska, Yann LeCun)



Deep Learning at Scale

Multi-Machine Training

- Elastic Averaging SGD!

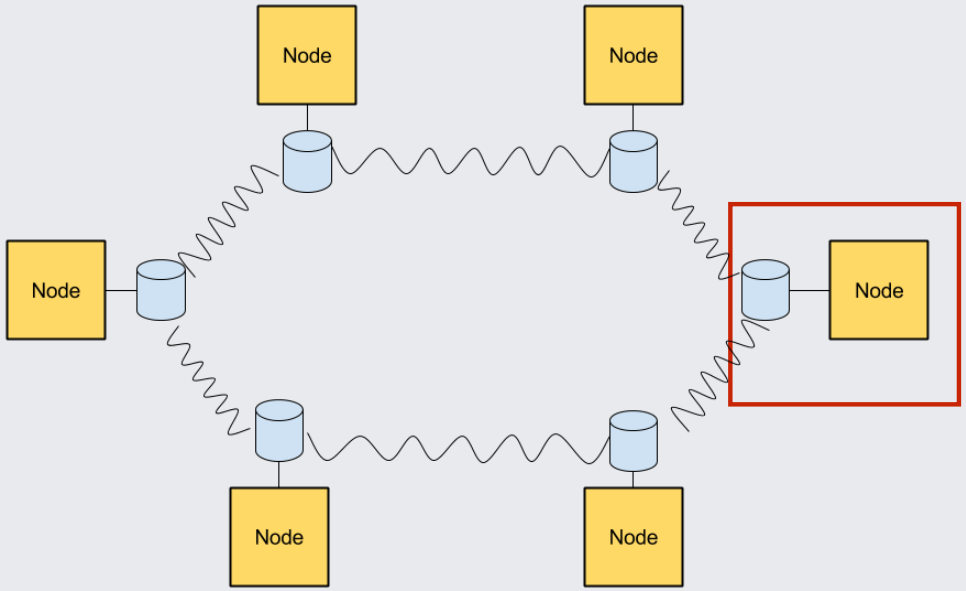


Train synchronously
Occasionally, check with master
Dont go too far from everyone else

Deep Learning at Scale

Multi-Machine Training

- Elastic Averaging SGD!



Train synchronously
Occasionally, check with neighbors
Dont go too far from everyone else

Deep Learning at Scale

Multi-Machine Training

- Elastic Averaging SGD!
  - Empirical speedup of SquareRoot(N)
    - N = number of nodes
  - No communication overhead with pre-fetching
    - 128 GPUs (32 clients * 4 GPUs)
    - Sharded parameters over 64 CPU servers
    - Tau = 10, prefetch = 5
    - zero overhead

Deep Learning at Scale

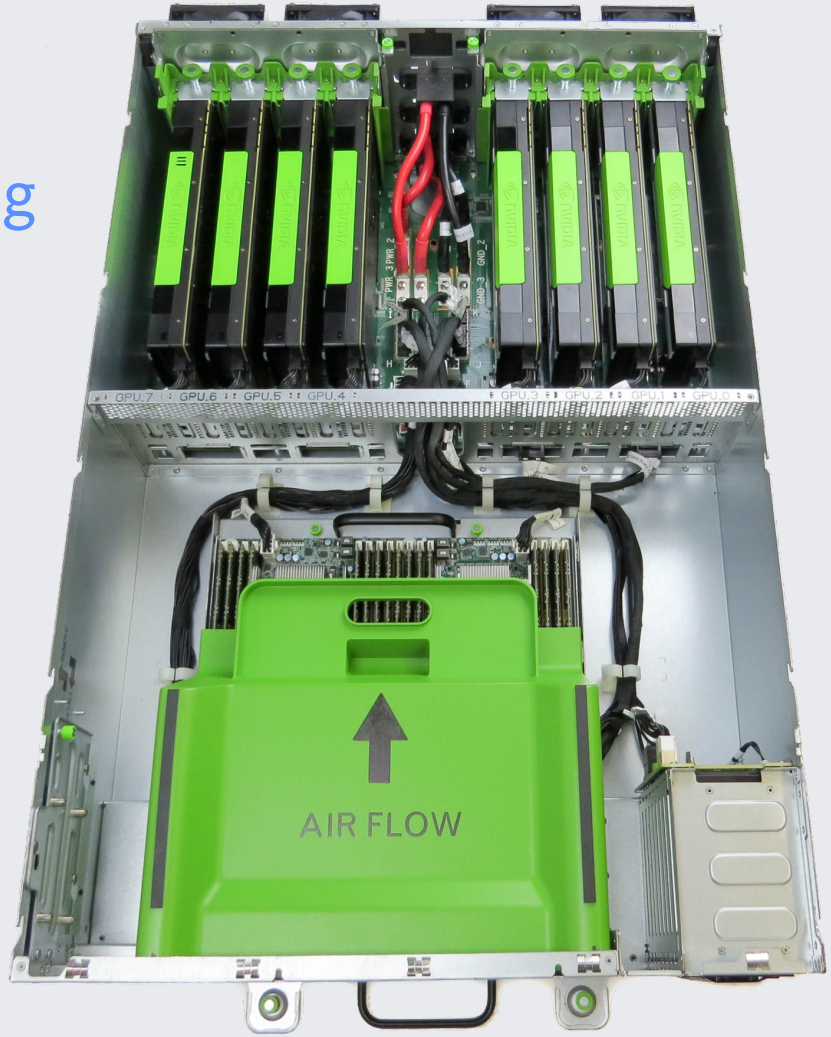
Multi-Machine Training

- Elastic Averaging SGD!
  - Fun fact: Trained AlexNet in 5 epochs of Imagenet data
  - Good success in training Vision and Text networks

Big Sur

Open Compute for Deep Learning

- Serviceability
- Thermal Efficiency
- Performance

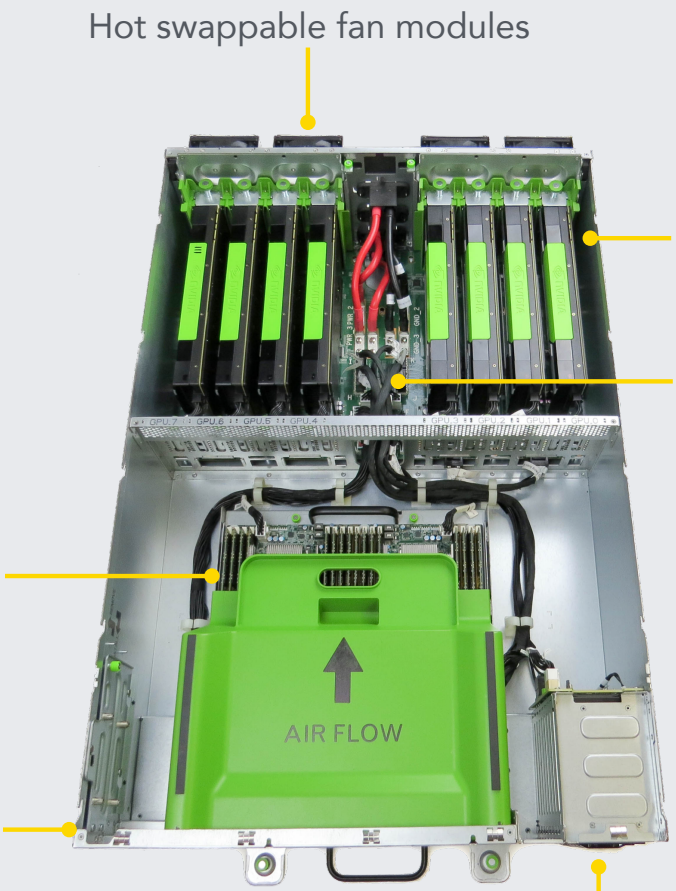


Big Sur

Open Compute for Deep Learning

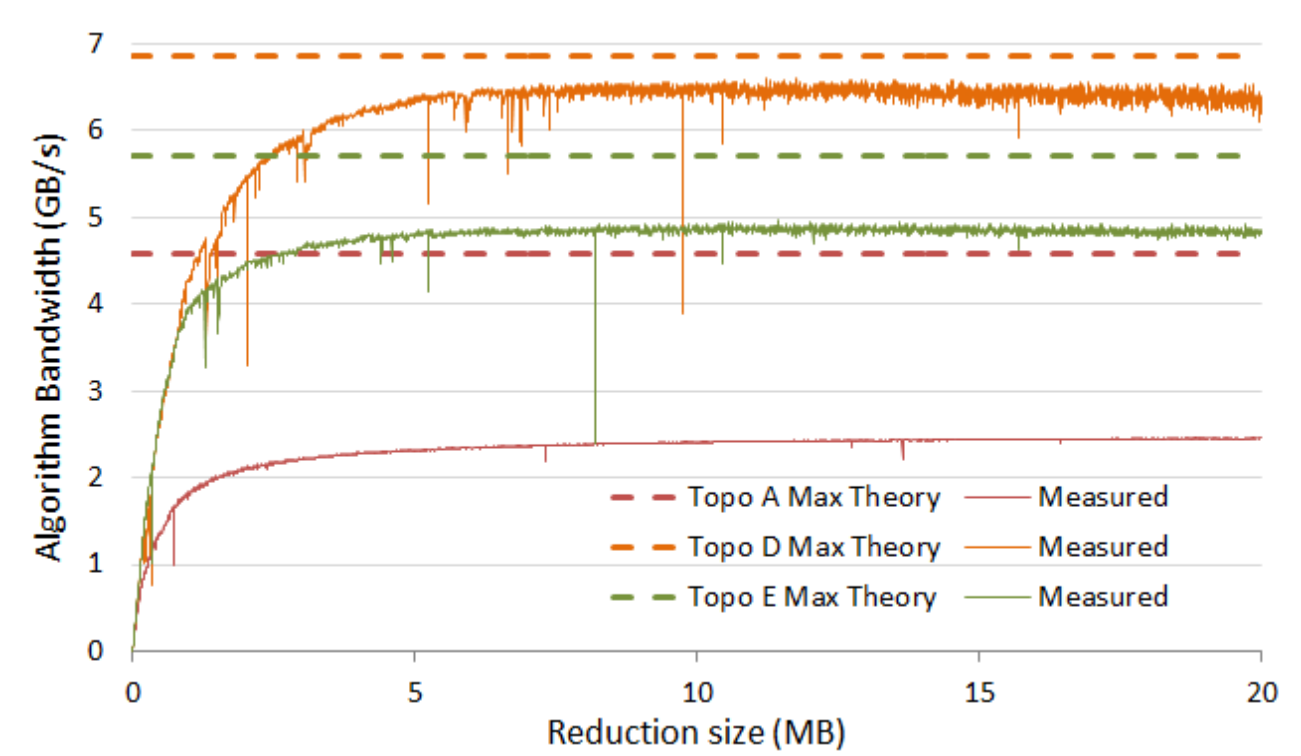
GPU removal using 2 thumb screws

Swap PCI-e Topologies with incredible ease



Big Sur

PCI-e Topologies — Matter!



Big Sur

PCI-e Topologies — Matter!

Network	#GPUs	batch size	#batches	E to A SpdUp	D to A SpdUp	D to E SpdUp
Alexnet, batch norm	8	1024	20	2%	5%	3%
vgg	8	512	10	-1%	0%	1%
vggstress	8	512	10	-3%	7%	10%
googlenet	8	512	10	0%	-2%	-2%


torch

Emerging Trends

Emerging Trends

Efficient Collectives + Imperative Programs

- Data / Model / Pipeline parallel seems sufficient
- Torch (nn / autograd / distlearn)
- Caffe

Emerging Trends

Computational Graph Toolkits

- Intel CnC, Caffe, TensorFlow, MXNet, Theano
- Graph placement hints + execution
- DSLs to write the computation graphs

Silver Bullet

Imperative Language + Graph Compiler

- Best of both worlds
- Hard problem of automatic graph placement
- Limited heuristic-driven success

Presence at GTC 2016

If you want to chat in-person, drop us an email

- Big Sur Hardware
  - Kevin Lee kevinlee@fb.com
  - Doug Wimer dwimer@fb.com
  - Soumith Chintala soumith@fb.com
- Multi-GPU / Multi-machine Training
  - Nicolas Vasilache ntv@fb.com
  - Jeff Johnson jhj@fb.com
  - Soumith Chintala soumith@fb.com
- Computation Graphs, Automatic Placement
  - Jeff Johnson jhj@fb.com
  - Andrew Tulloch tulloch@fb.com
  - Yangqing Jia jiayq@fb.com
  - Soumith Chintala soumith@fb.com


facebook