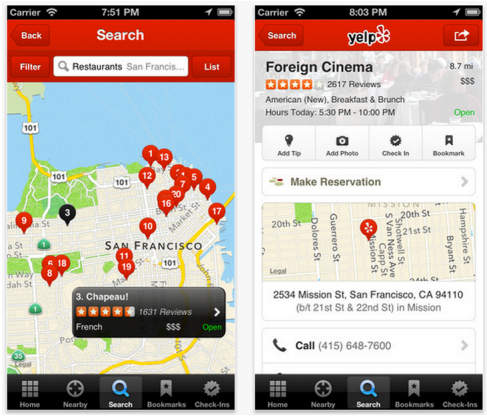
HPC Approach to Training Neural Networks in Deep Learning

Patrick LeGresley
on behalf of the entire
Baidu Silicon Valley AI Lab (SVAIL)

MVAPICH User Group Meeting
August 20, 2015

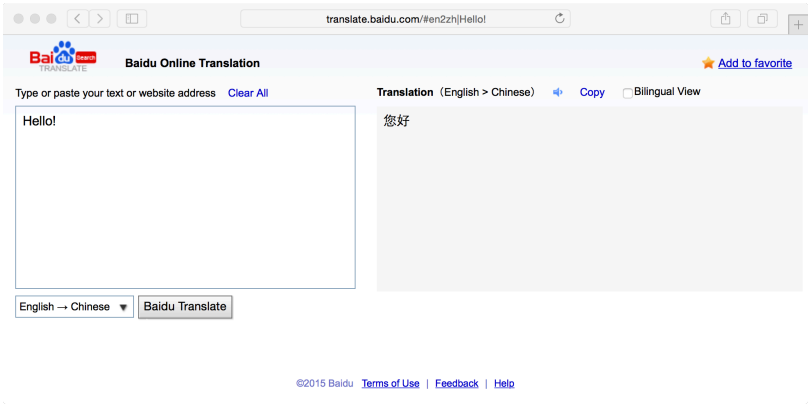
Baidu Research

Machine Learning



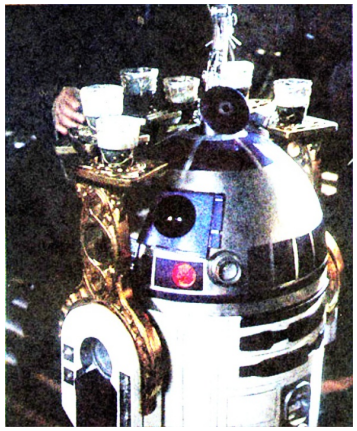
Helps us find things

Assists us



Facilitates communication

Drive us around?

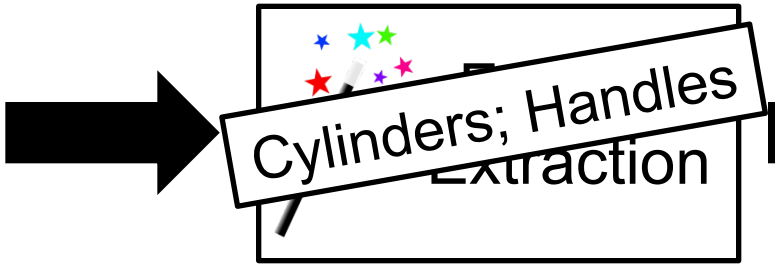


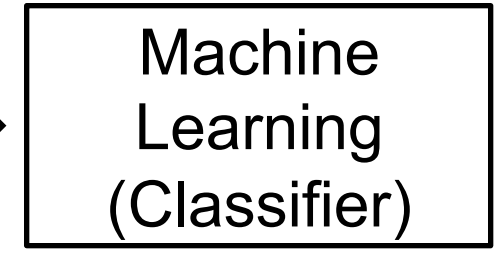
Serve drinks?

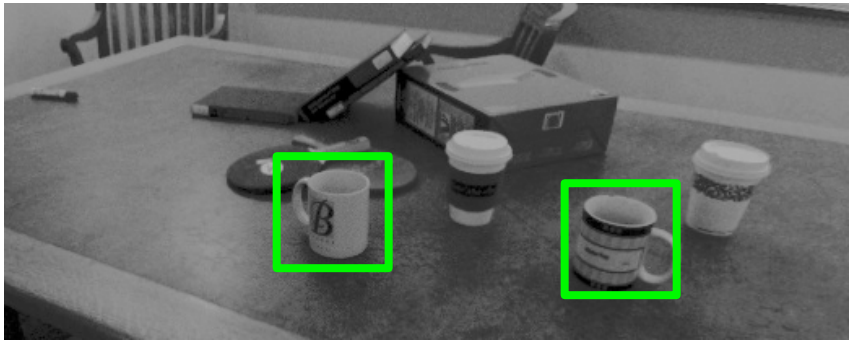
Fold laundry?

Baidu Research

Machine Learning in Practice

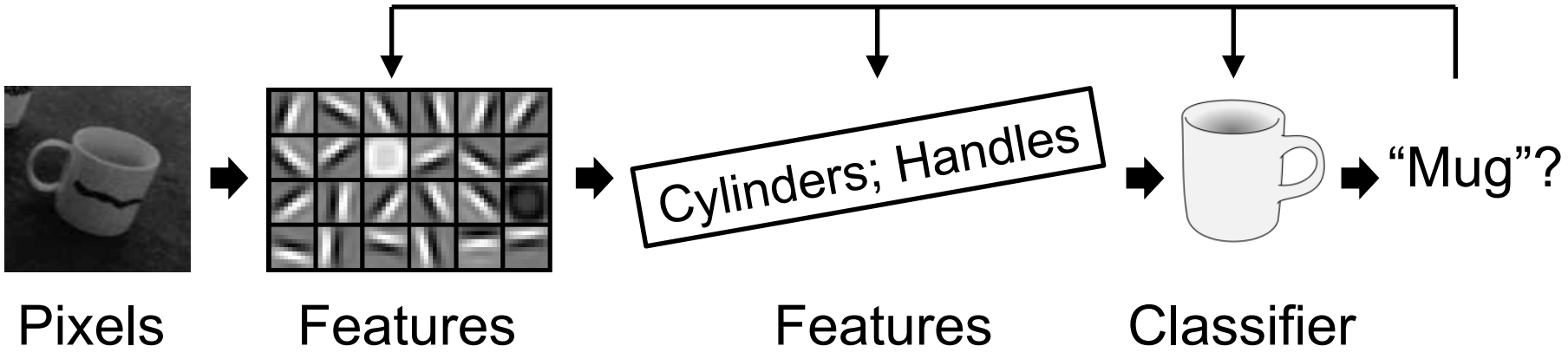






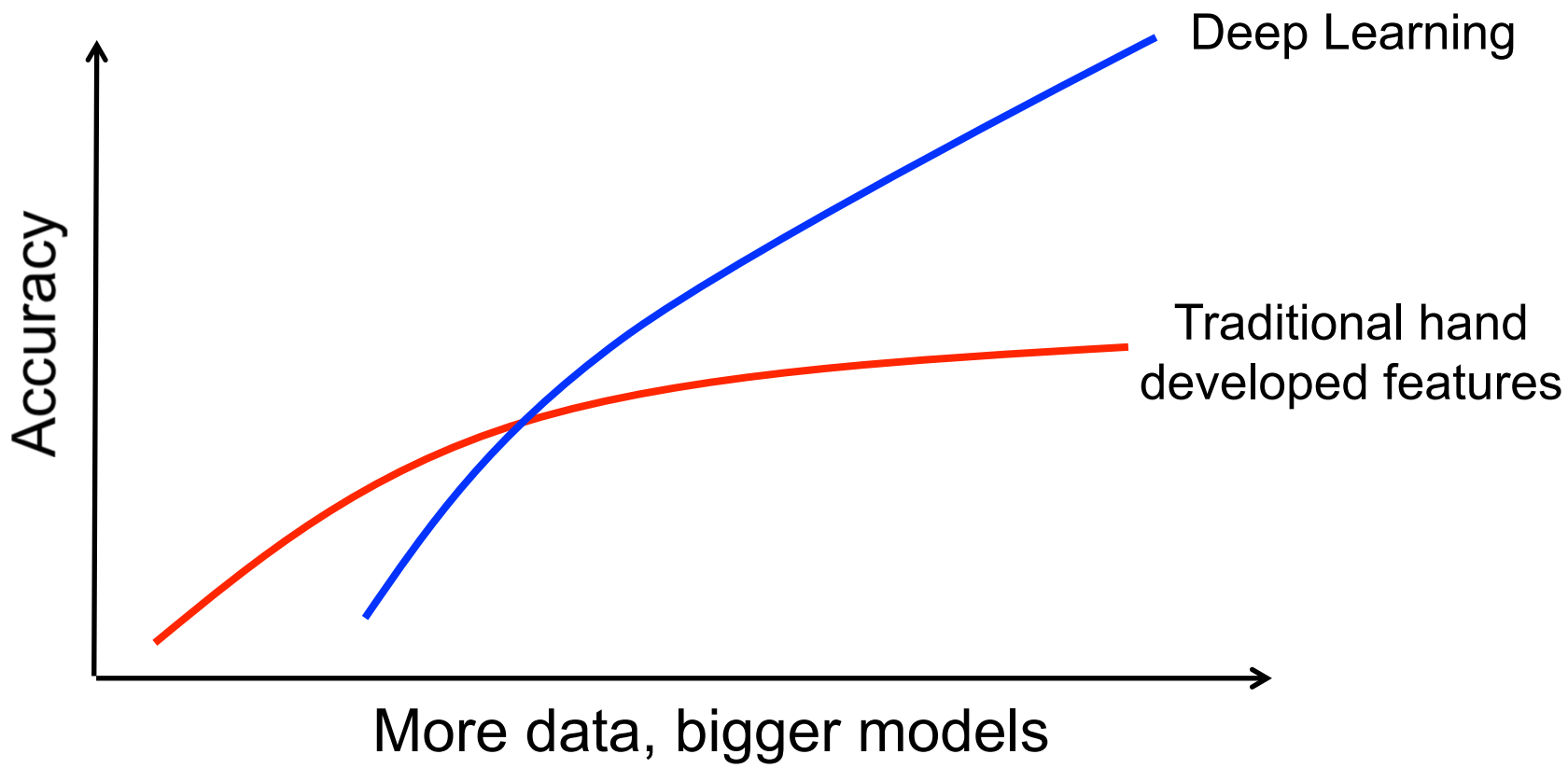
Baidu Research

Deep Learning



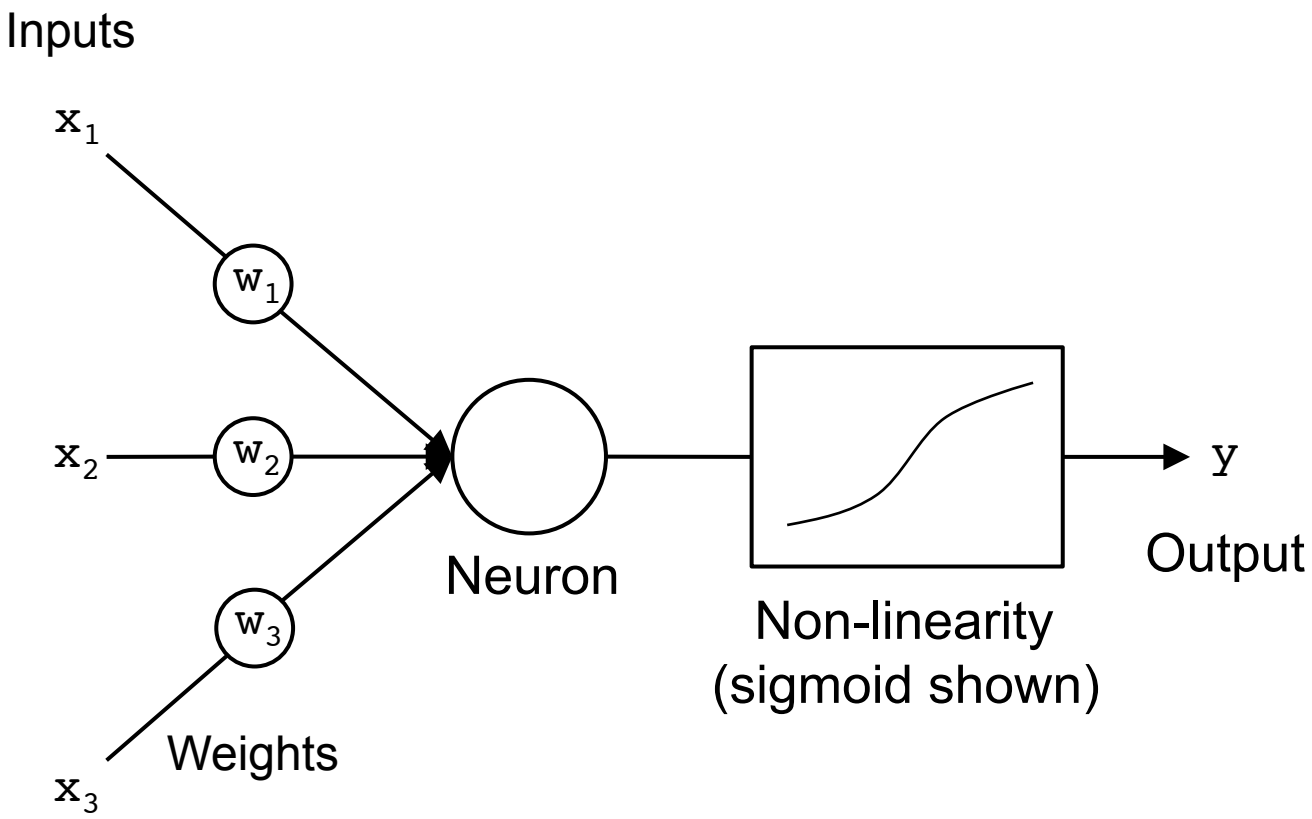
- Instead of feature extraction and hand tuning via domain experts, use lots of data and let model determine the features

Deep Learning



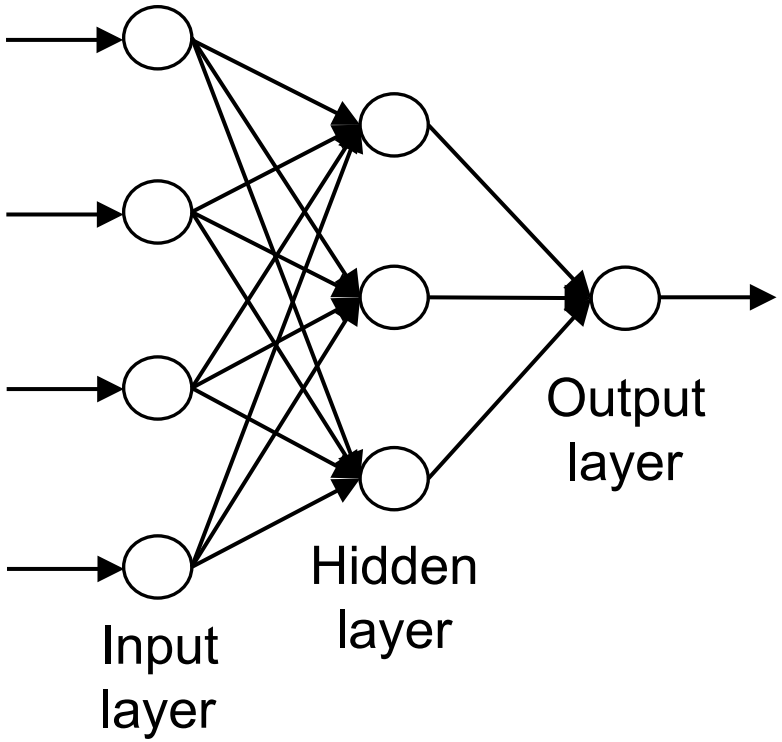
Baidu Research

Artificial Neuron



Baidu Research

Artificial Neural Network



- In a realistic network there will typically be many layers, each with millions of neurons, resulting in billions of connections
- ***But fundamentally just a complicated function that takes an input and computes an output***

Baidu Research

Why Speech?

Baidu Research

Speech is a Universal Interface

Samsung Galaxy Gear Smartwatch

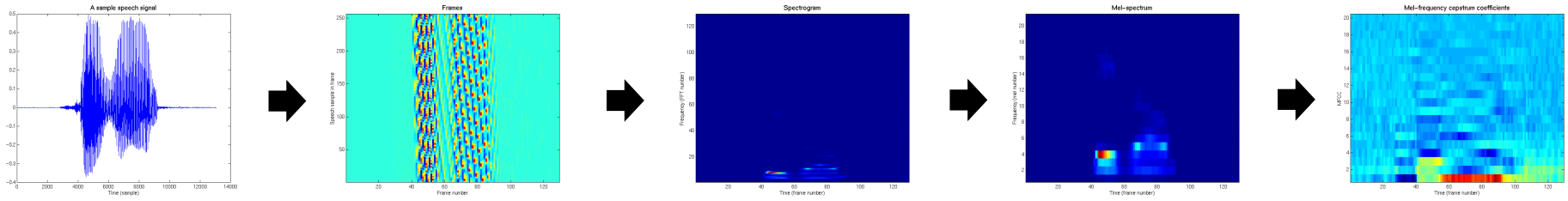
Amazon Echo

Baidu Eye

Baidu Research

Traditional Approaches for Speech

- Hands on development of feature identification by domain experts
- Small, incremental improvements
- Some approaches may not generalize well, e.g. tonal vs. non-tonal languages

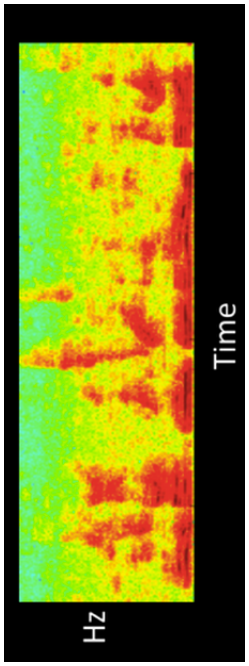


Mel-Frequency Cepstrum Coefficients (MFCC)

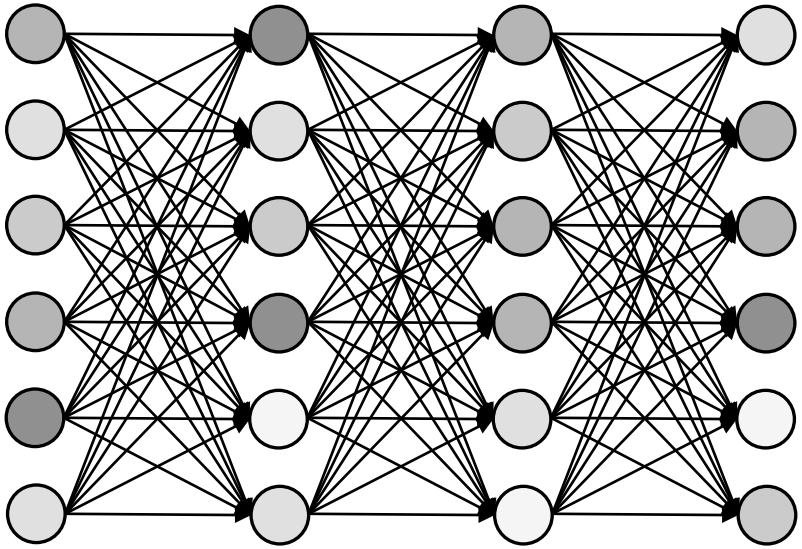
Minh N. Do, University of Illinois

Baidu Research

Deep Learning for Speech



Linear spectrogram



Artificial Neural Network

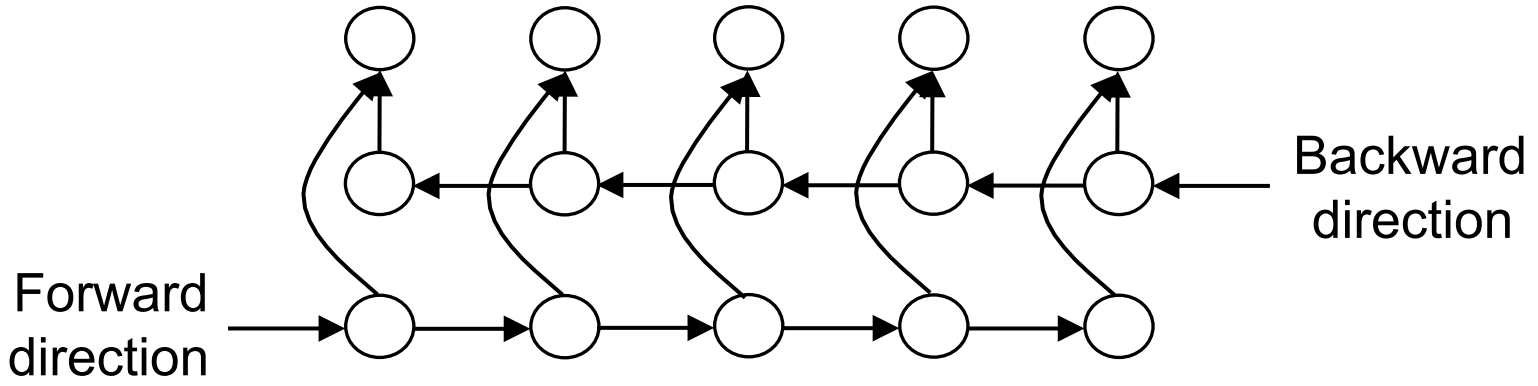
call home and tell them i'll be late

Transcription

Baidu Research

Bidirectional Recurrent Neural Networks

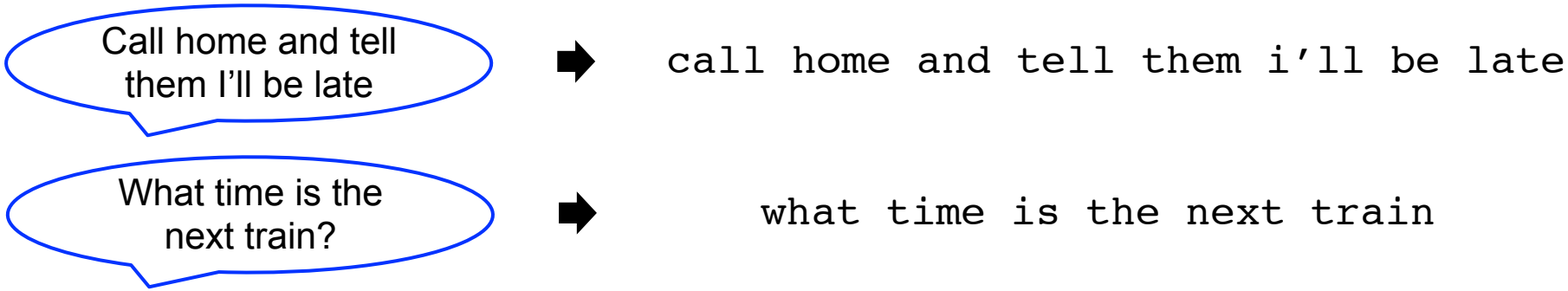
- Combines the outputs of two RNNs, one going in each direction of the input
- Useful where temporal dependencies are important



Baidu Research

Model Training

- Process of computing the values for the weights in the network

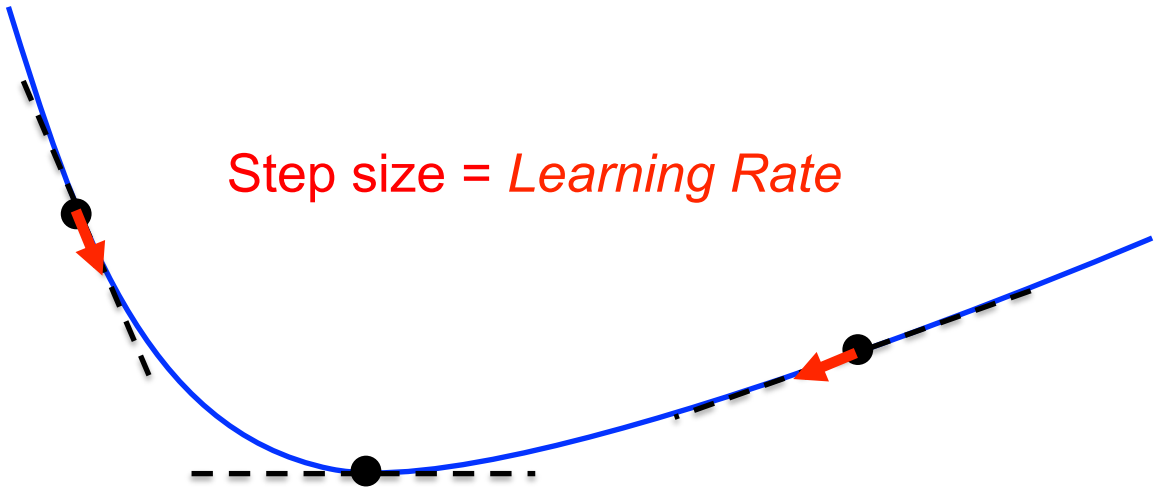


Labeled training data

- Formulated as a large scale optimization problem: adjust the weights to minimize the error for the labeled data

Baidu Research

Gradient Based Optimization

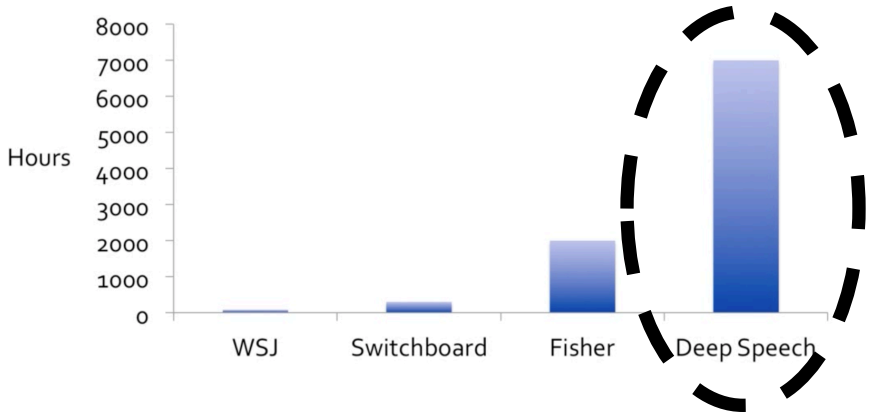


- Gradient of the output error with respect to all of the weight values can be computed analytically using a process called *back propagation* (the chain rule)
- Typically split data into *minibatches* and use Stochastic Gradient Descent (SGD) so working set fits in memory

Baidu Research

Training Data and Computational Costs

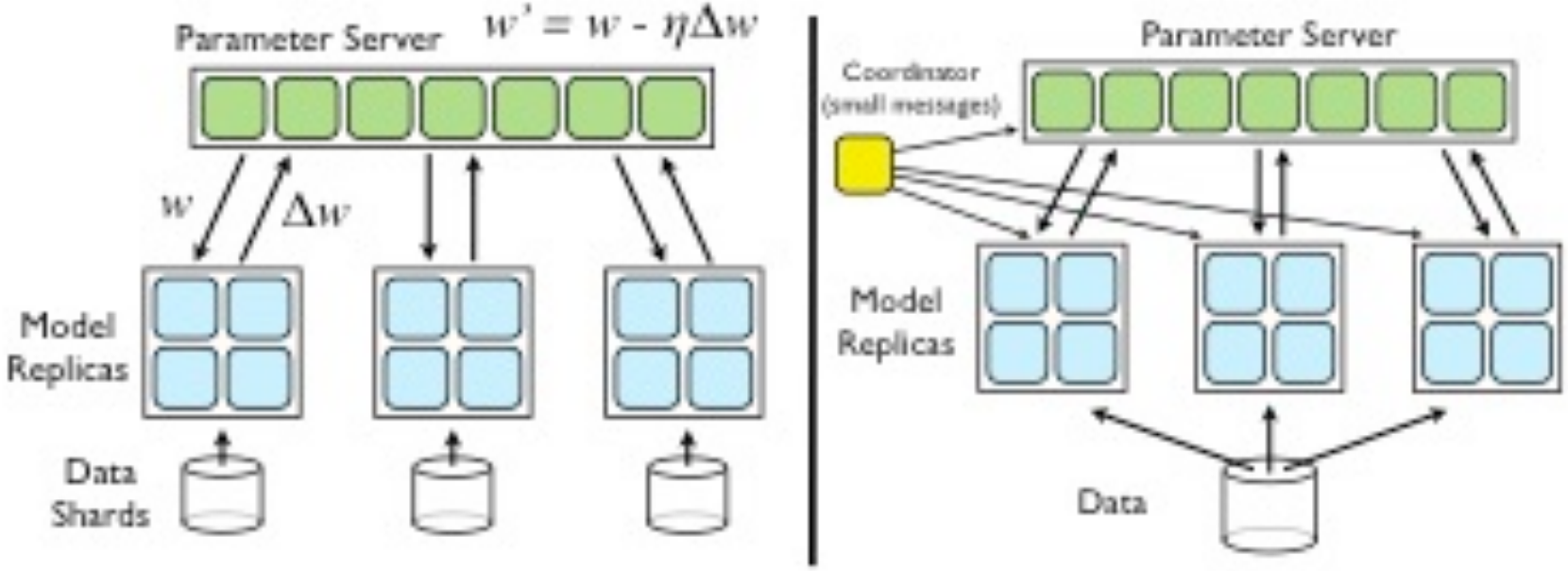
- Thousands of hours of labeled training data
- Can also add noise and create *synthesized data*



- Training one model requires 10s of exaFLOP of computations (using 16 or 32 bit floating point)
- Want to train many models to experiment with different layer configurations, non-linearities, etc.

Baidu Research

Traditional Approach



DistBelief (Dean et al. 2012)

Baidu Research

Deterministic Results

- Lock-less, asynchronous approaches achieve performance at the expense of determinism
- These approaches can also mask weird bugs
- We feel that results should be replicable and that non-determinism is a bug
- Both performance and determinism can be achieved by adopting HPC practices

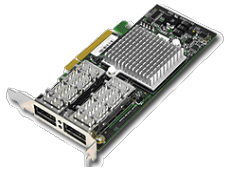
Baidu Research

SVAIL HPC Infrastructure

- Software: CUDA, MPI, Majel (SVAIL internal library)
- Hardware:

NVIDIA GeForce
GTX Titan X

Titan X x8



Mellanox Interconnect

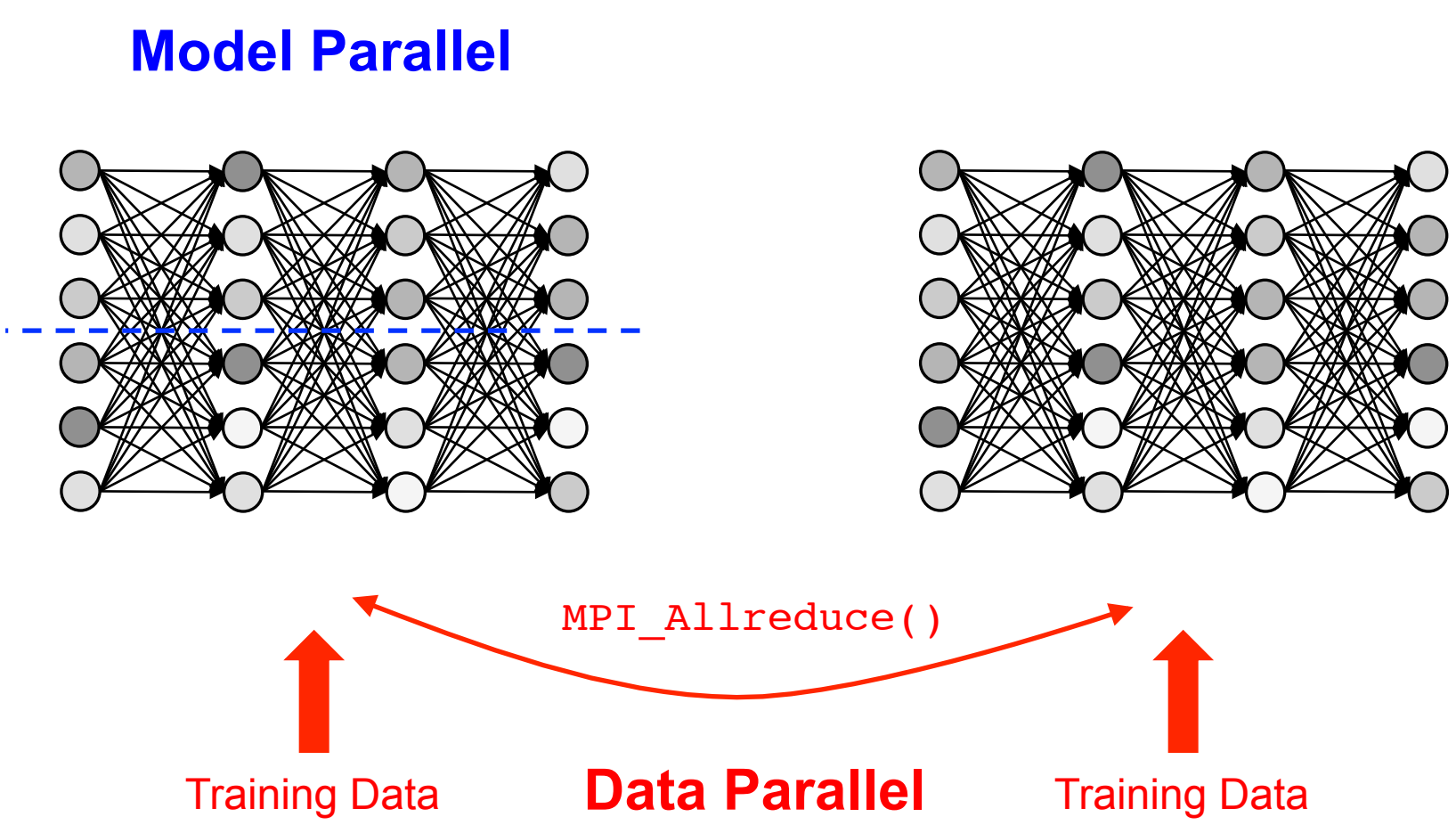
Baidu Research

Majel

- C++ library for multi-dimensional, contiguous in memory arrays
- Python bindings
- CPU / GPU aware
- Manages memory pool for dynamic memory, and provides a GPU fallback allocator
- Supports various vector / matrix operations, and convolution
  - Use NVIDIA and Nervana Systems libraries for GPU
- Reduction operations (on single GPU and distributed CPU or GPU memory)

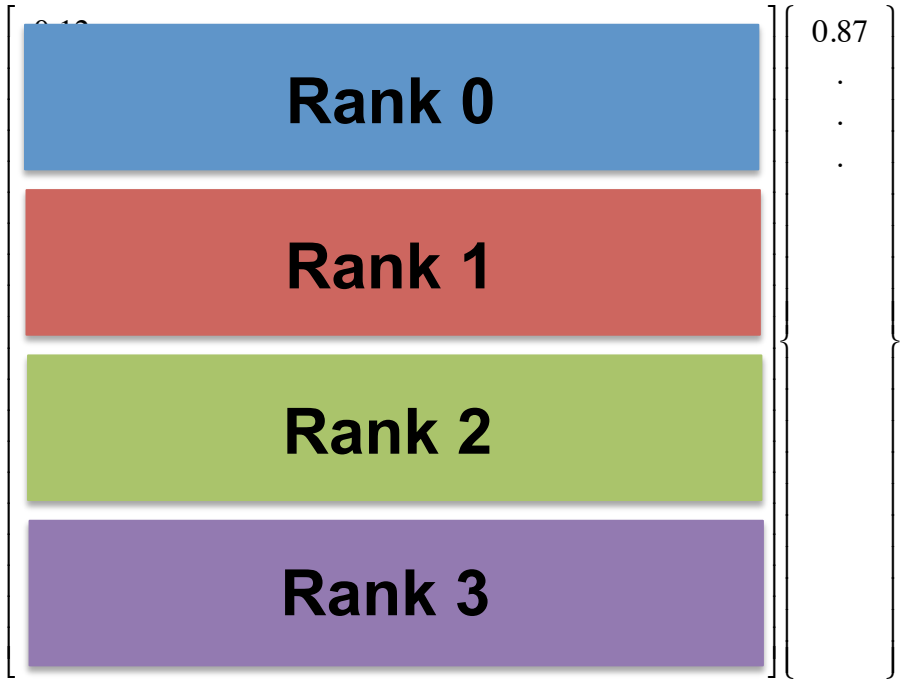
Baidu Research

Parallelism



Baidu Research

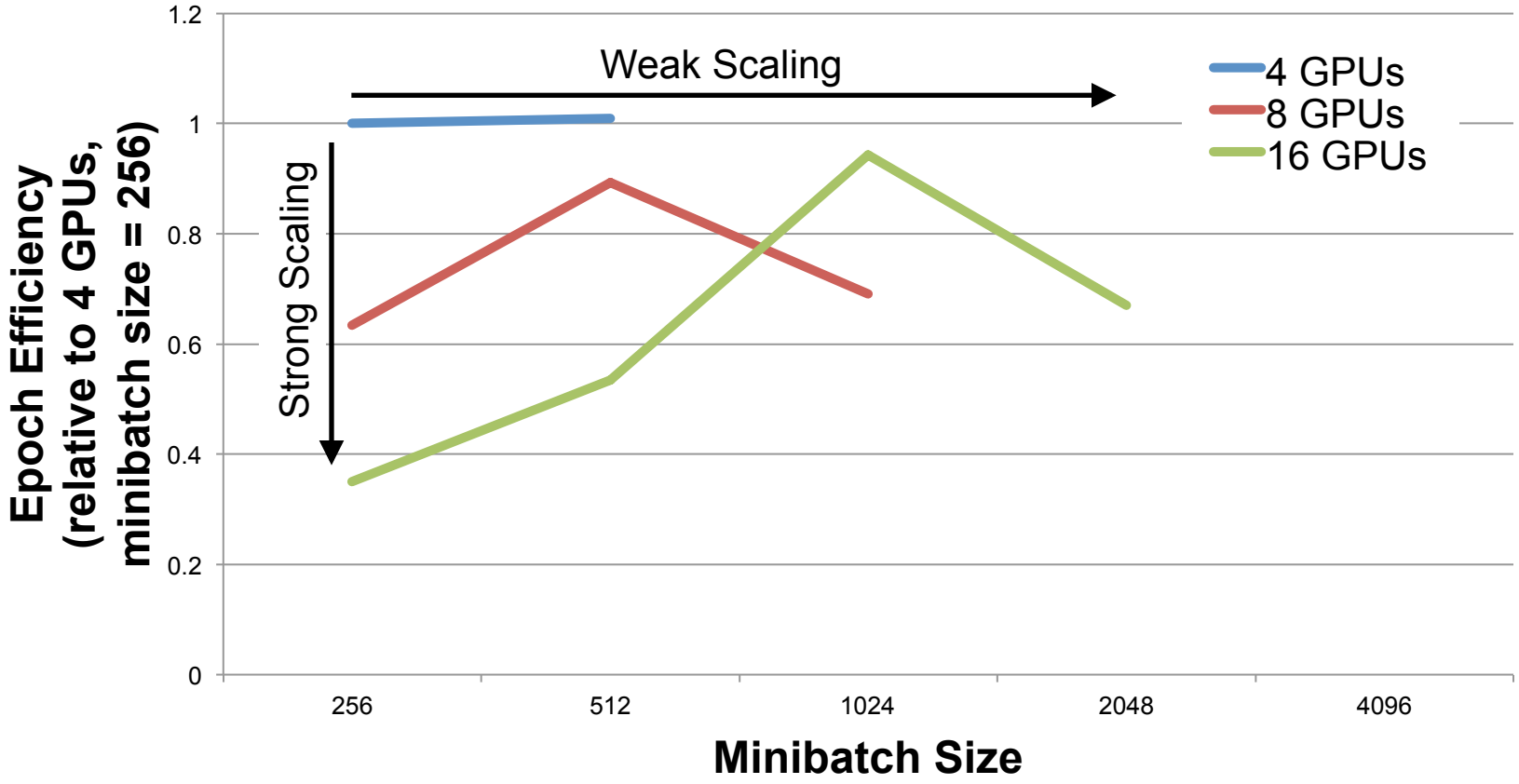
Distributed Array



- Reduces per GPU memory requirement
- More latency sensitive
- Can be combined with model / data parallel
- Better scalability

Baidu Research

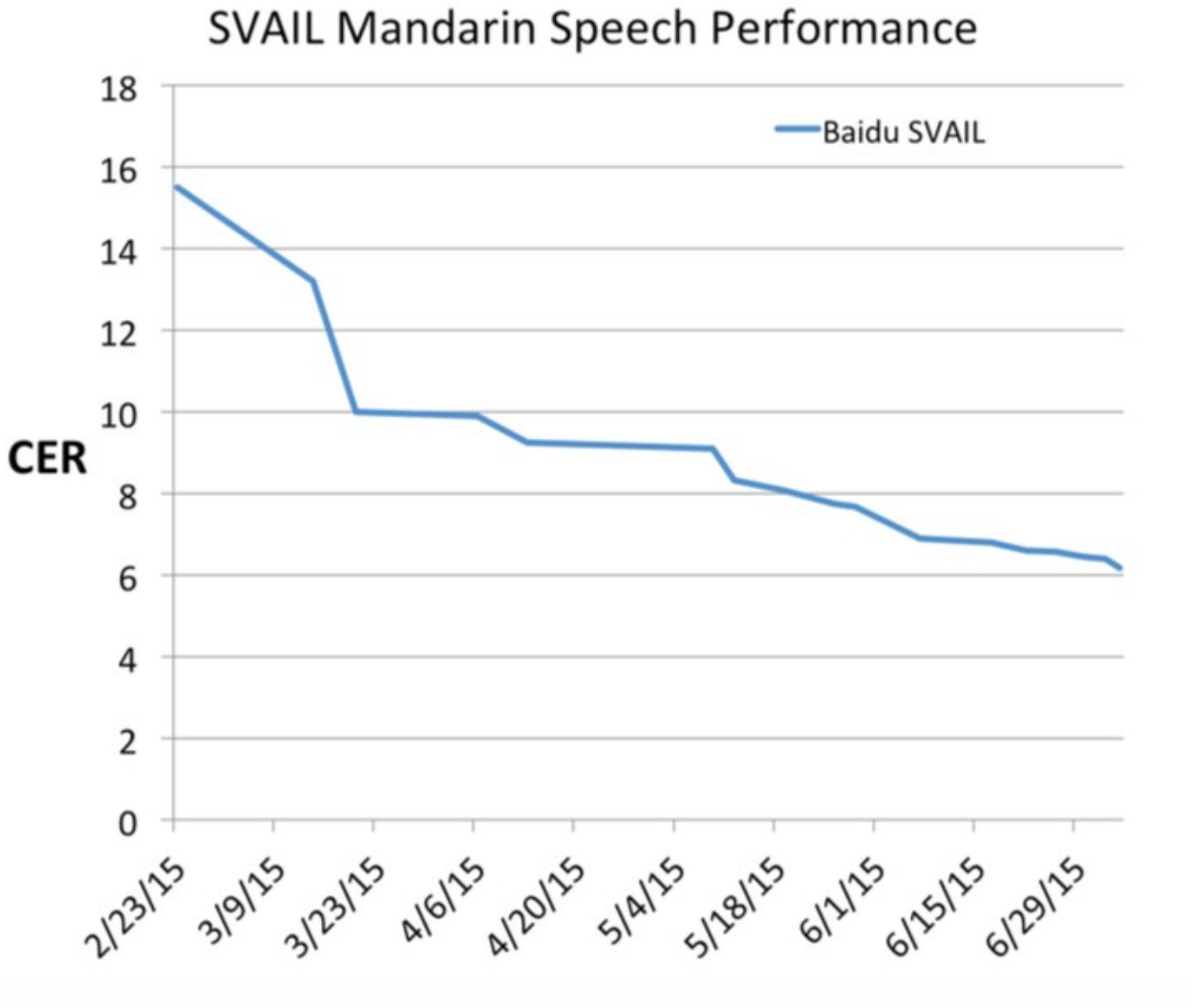
Performance



- Efficiency
  - 60-65% of GPU FMA peak using a single GPU
  - ~50% of peak using 8 GPUs in one node
- Model training time is a few days

Baidu Research

Accuracy



CER = Character Error Rate

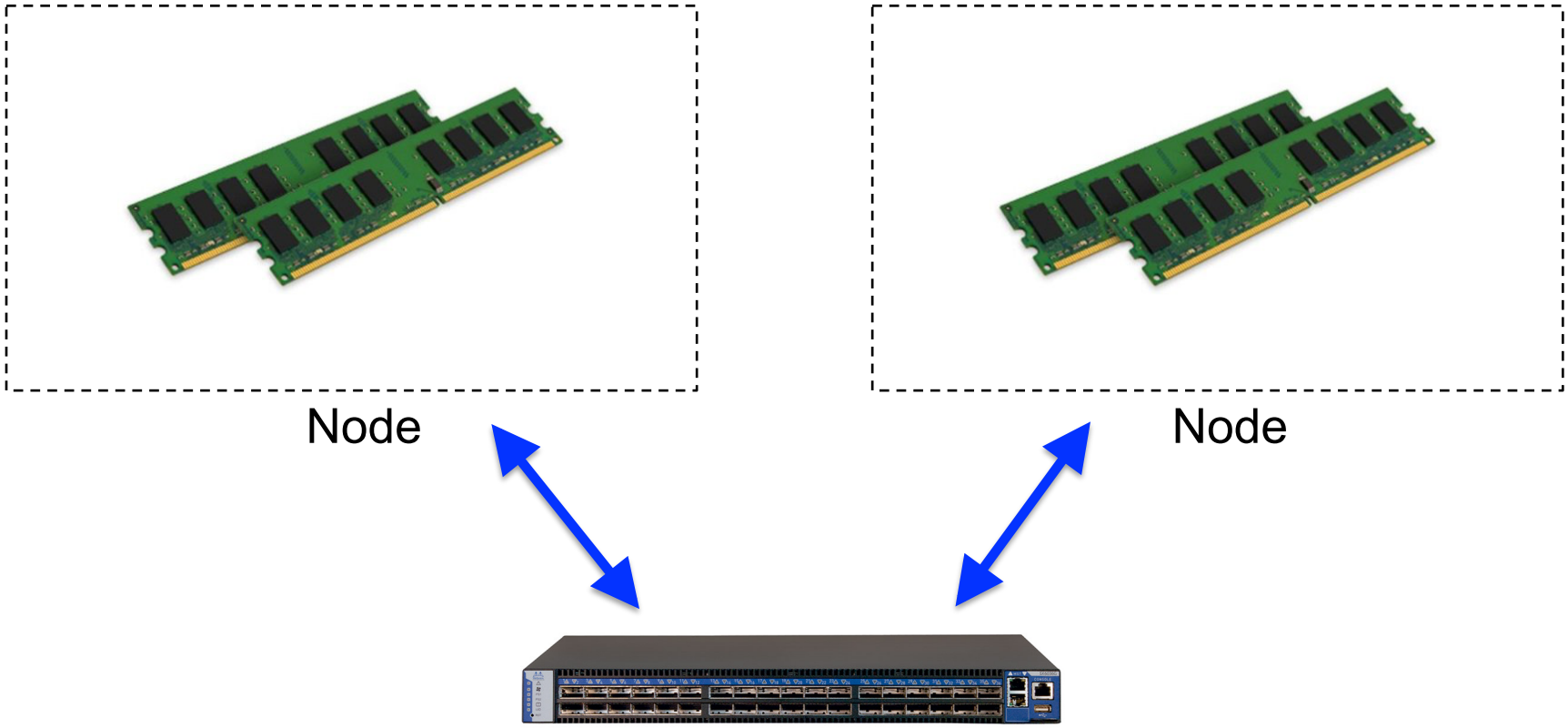
Baidu Research

Challenges and Opportunities

- Memory Allocation Takeover
- Machine Topology
- CUDA Aware MPI

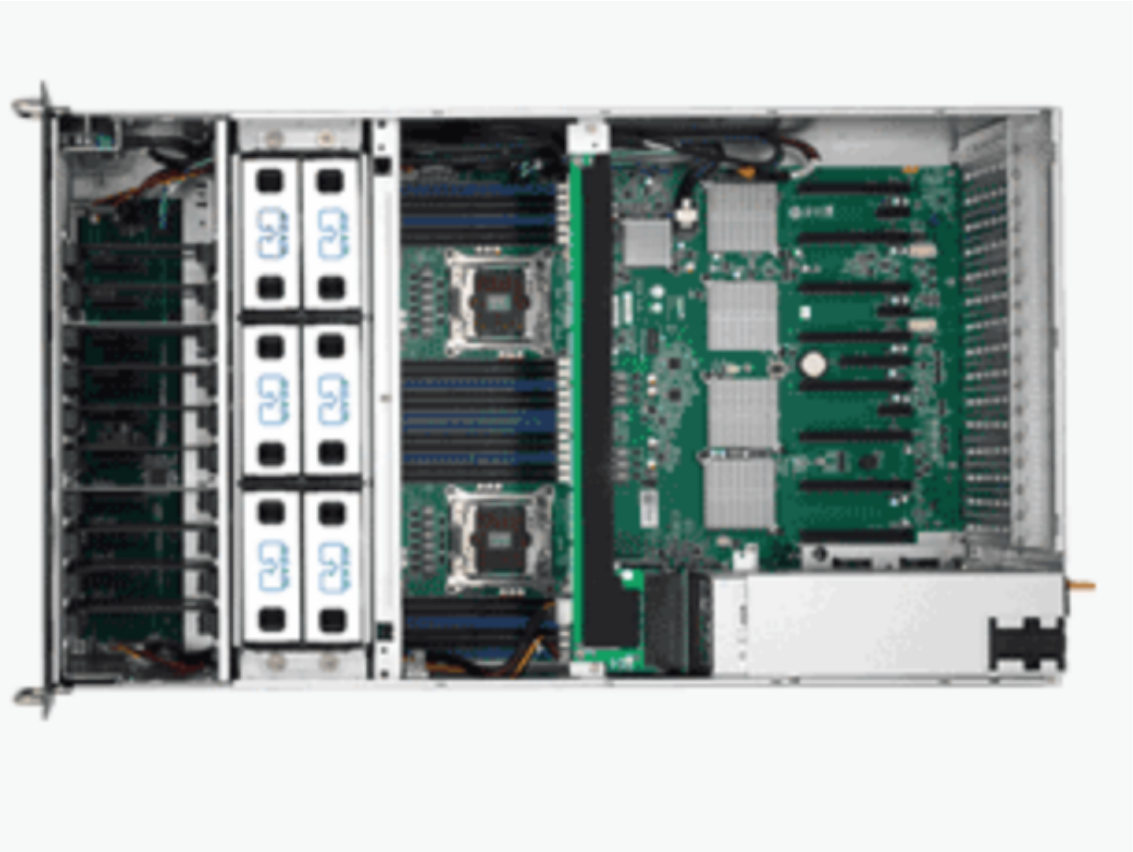
Baidu Research

Memory Allocation Takeover



- Possible to provide finer grain control over when it is used?

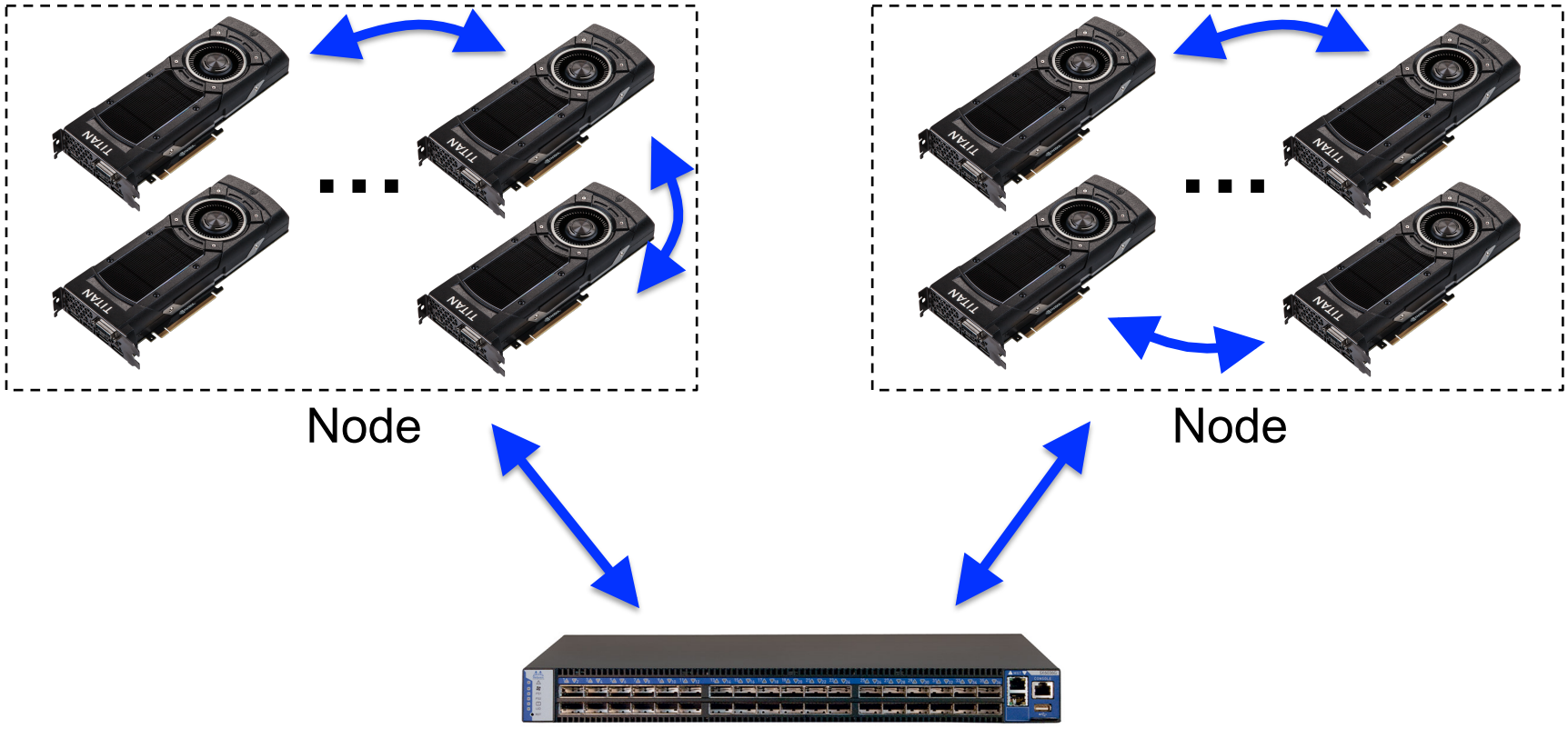
Machine Topology



- What common things can be done to make life easier for application developers?

Baidu Research

CUDA Aware MPI



- Seems immature in terms of both features and performance

Baidu Research

Summary

- Concepts in neural networks are remarkably simple
- The systems team in our AI lab works toward achieving
  - Deterministic results
  - Application performance ≥ 50% of hardware theoretical peak
  - Strong scaling for multiple GPUs across multiple nodes
- **AI, via HPC, will impact hundreds of millions of people!**

- *We are always looking for more systems researchers with backgrounds in parallel computing, computer architecture, storage, etc.*

Baidu Research

Find Out More

- Deep Learning Tutorial
  http://deeplearning.stanford.edu/tutorial/

- Deep Speech: Scaling up end-to-end speech recognition
  http://arxiv.org/abs/1412.5567

- Baidu Research: Making Progress in Multi-Lingual Speech Recognition
  http://usa.baidu.com/multi-lingual-speech-recognition/

Baidu Research

Speech Transcription Demo