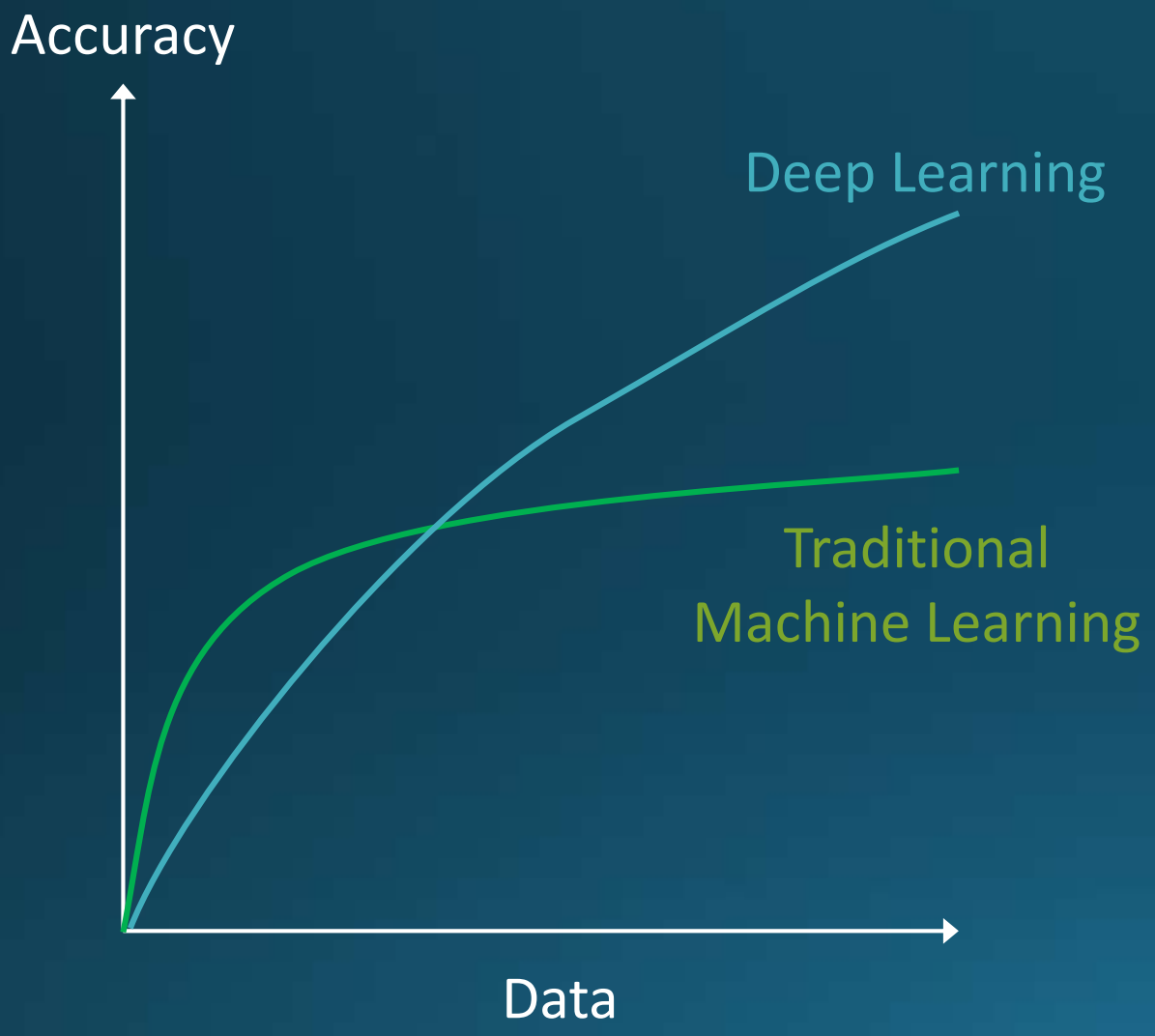
Deep-Learning for the Enterprise

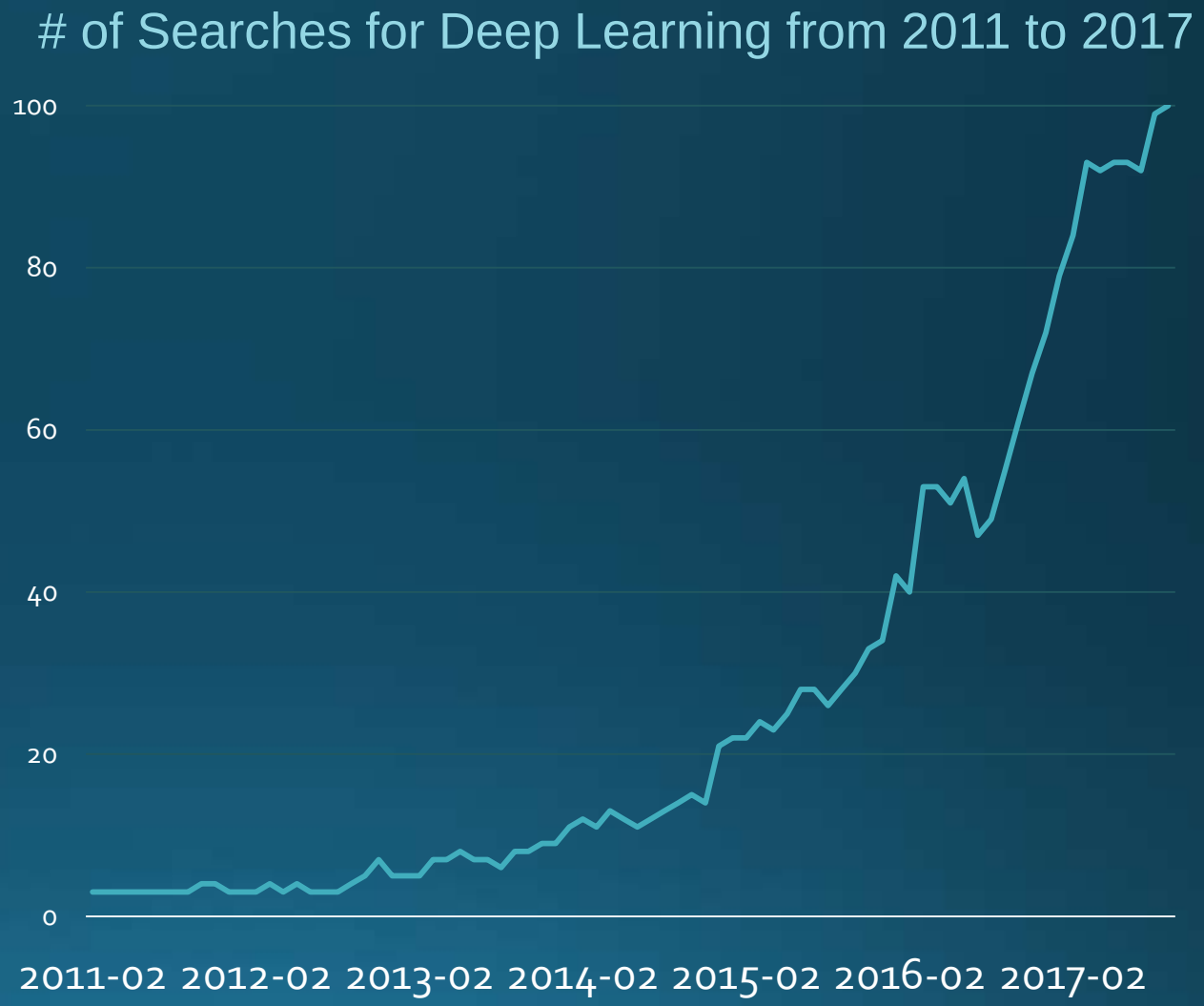
Sumit Gupta
VP, HPC, AI, & Machine Learning
IBM Cognitive Systems

December, 2017

IBM PowerAI

Deep Learning Has Revolutionized Machine Learning





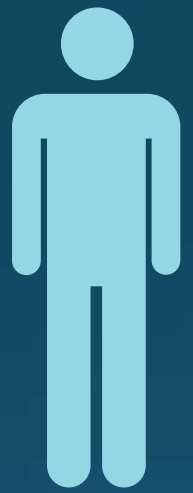
Source: Google Trends. Search term "Deep Learning"

2011

26% Errors

Machine Learning Based

Humans



5% Error

2016

3% Errors

Deep Learning Based

IBM's AI Solutions

PowerAI: Enterprise Distribution of Open-Source AI Frameworks

Developer Ease-of-Use Tools

Open Source Frameworks: Supported Distribution

Faster Training Times via
HW & SW Performance Optimizations

PowerAI Vision: Vision Auto-Model Generator

- Semi-automatic labelling
- Automatically trains deep learning model for labeled input data set
- Enables non-deep learning specialists to build trained AI models

Image Labeling and Preprocessing

Video Labeling Service

Custom Learning for Image Classification

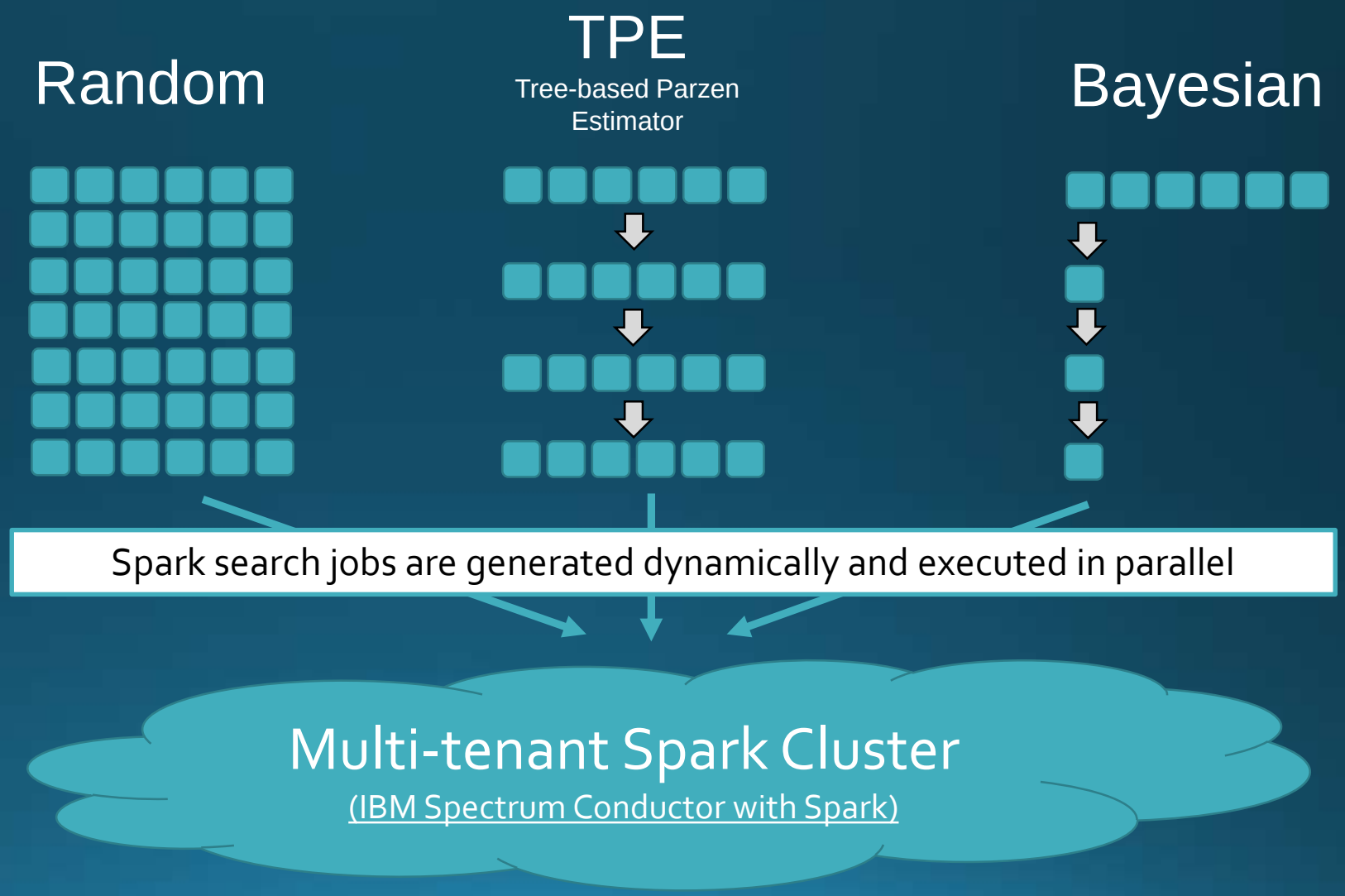
Custom Learning for Object Detection

Self-defined Training with visualized monitoring

Inference API deployment

Auto Hyper-Parameter Tuning/Search

- Hyper-parameters
  - Learning rate
  - Decay rate
  - Batch size
  - Optimizer:
    - GradientDecedent,
    - Adadelta,
    - Momentum,
    - RMSProp
    - …
  - Momentum (for some optimizers)
  - LSTM hidden unit size (for models which use LSTM)



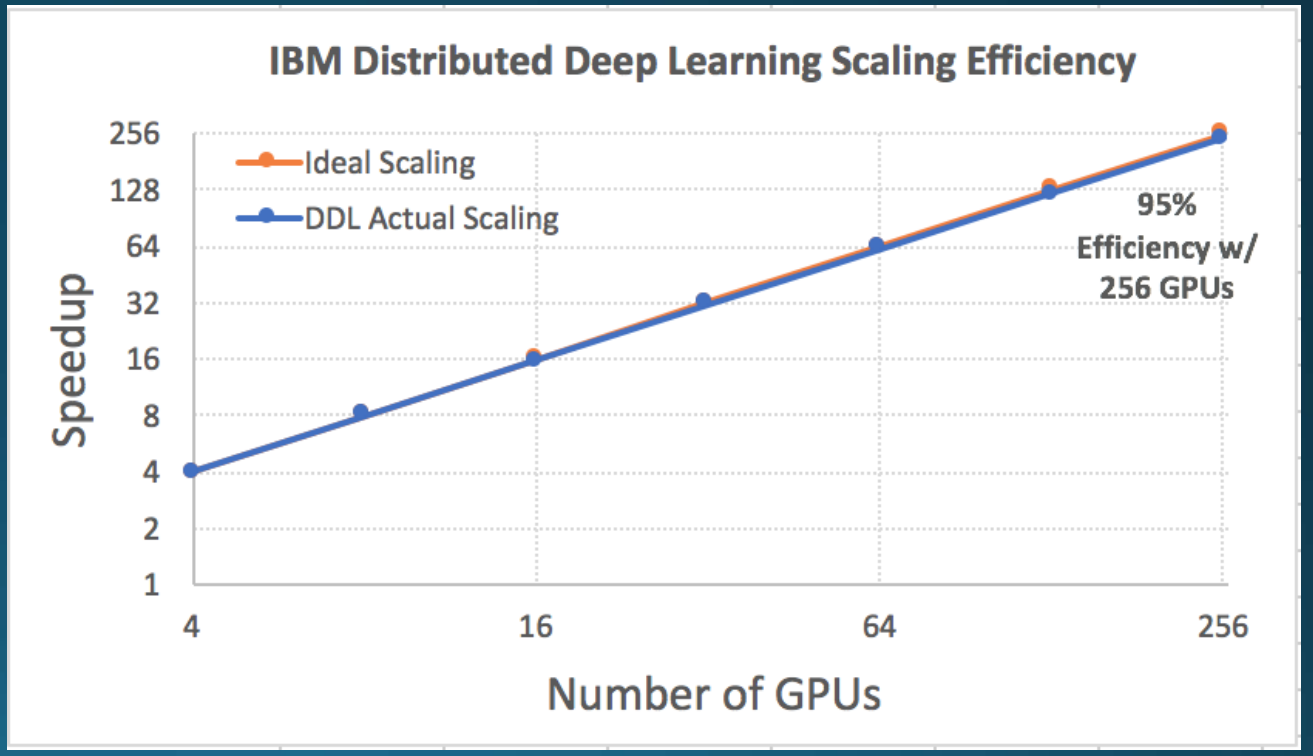
Distributed Deep Learning (DDL)

Reducing Training Time from Weeks to Hours

16 Days Down to 7 Hours:

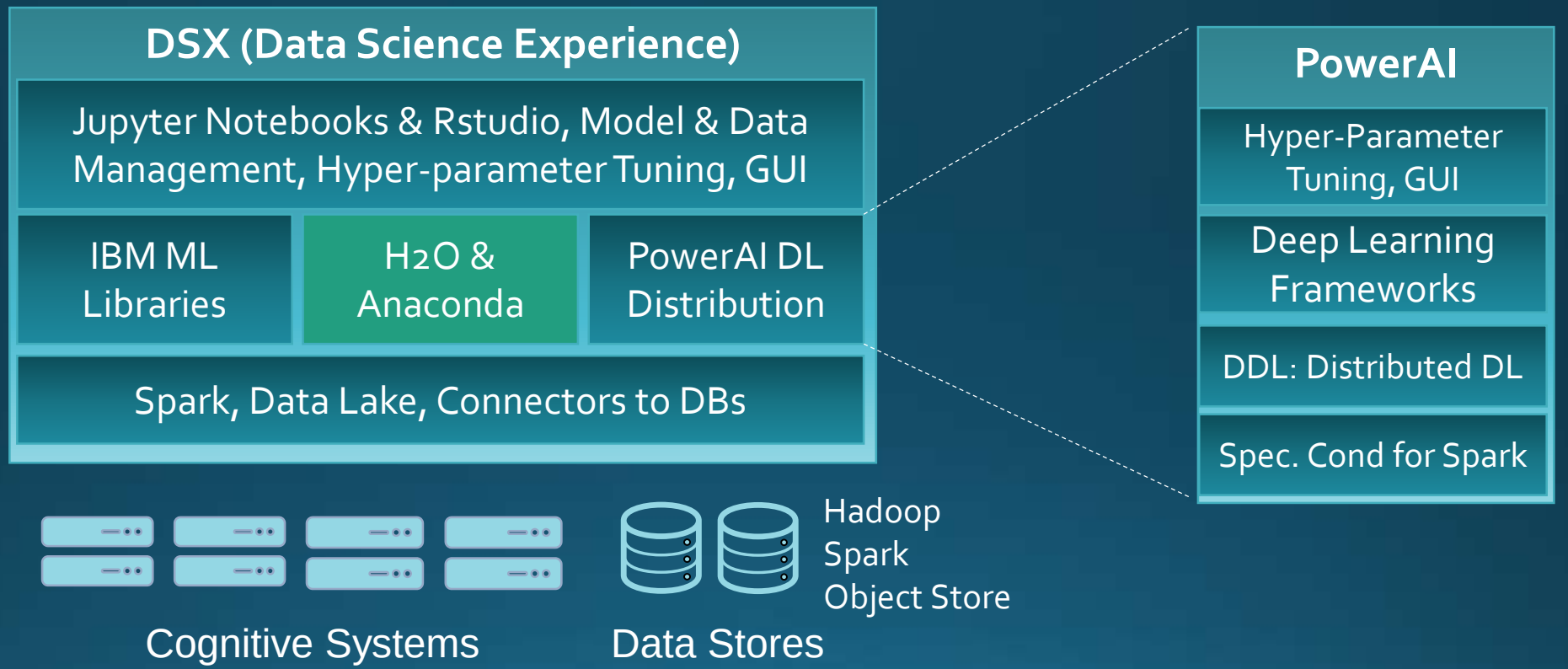
58x Faster

Near Ideal Scaling to 256 GPUs and Beyond



ResNet-101, ImageNet-22K, Caffe with PowerAI DDL, Running on Minsky (S822Lc) Power System

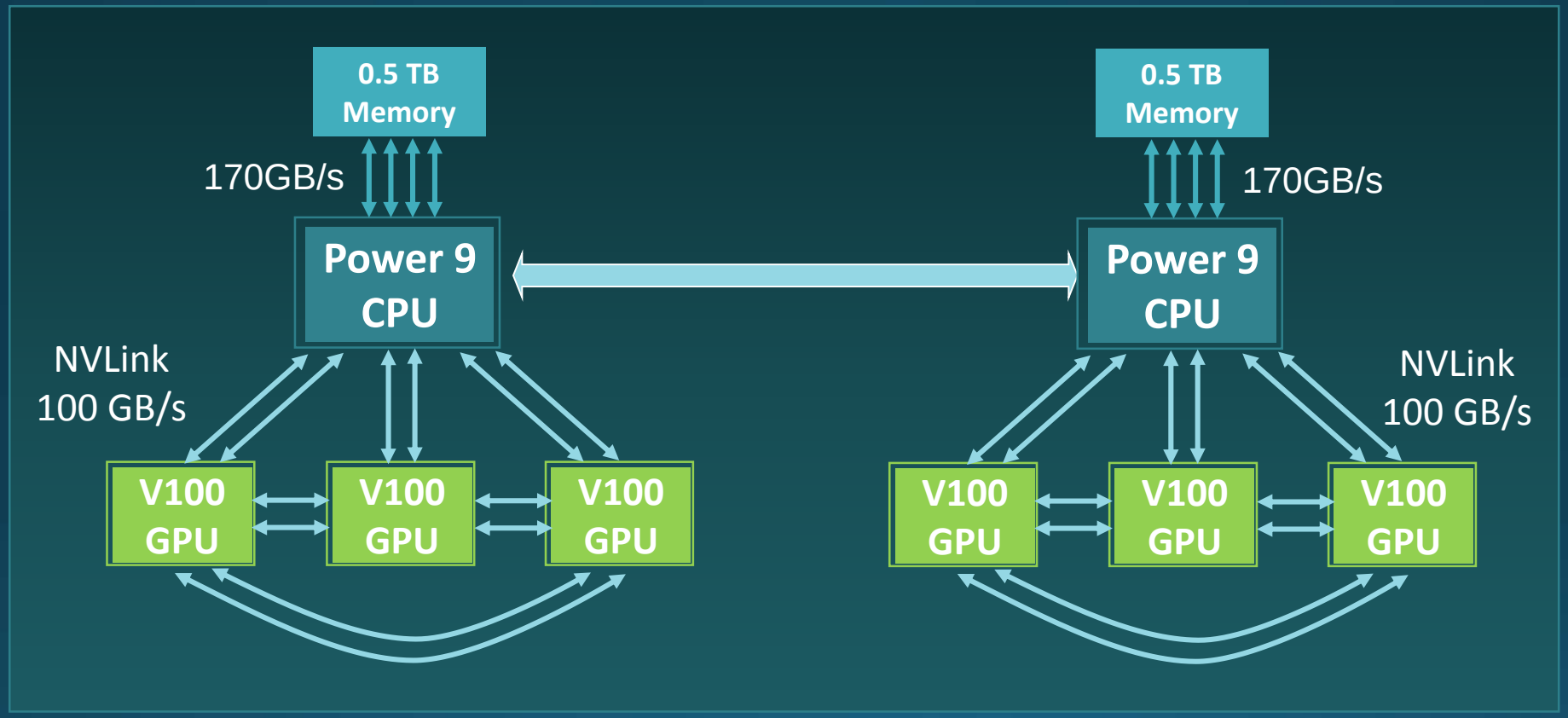
IBM AI / Data Science Workbench



Legend: Non-IBM Products

Announcing New Deep Learning Server

IBM AC922 Power System

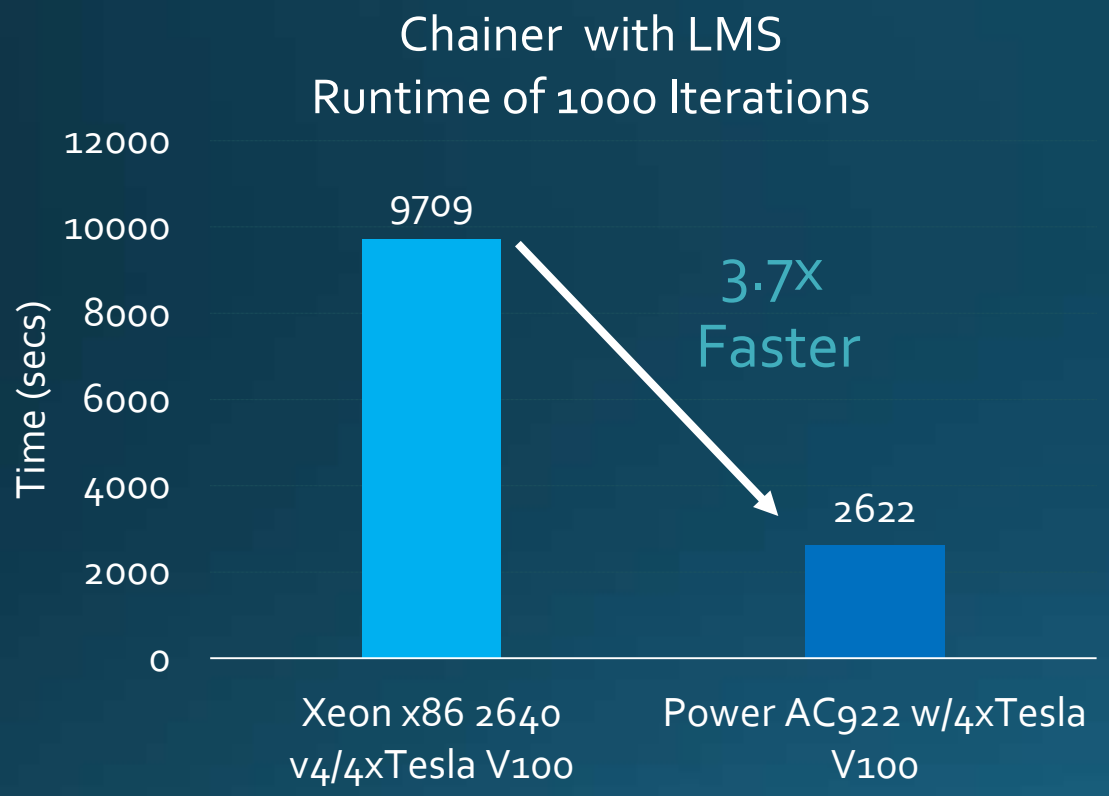


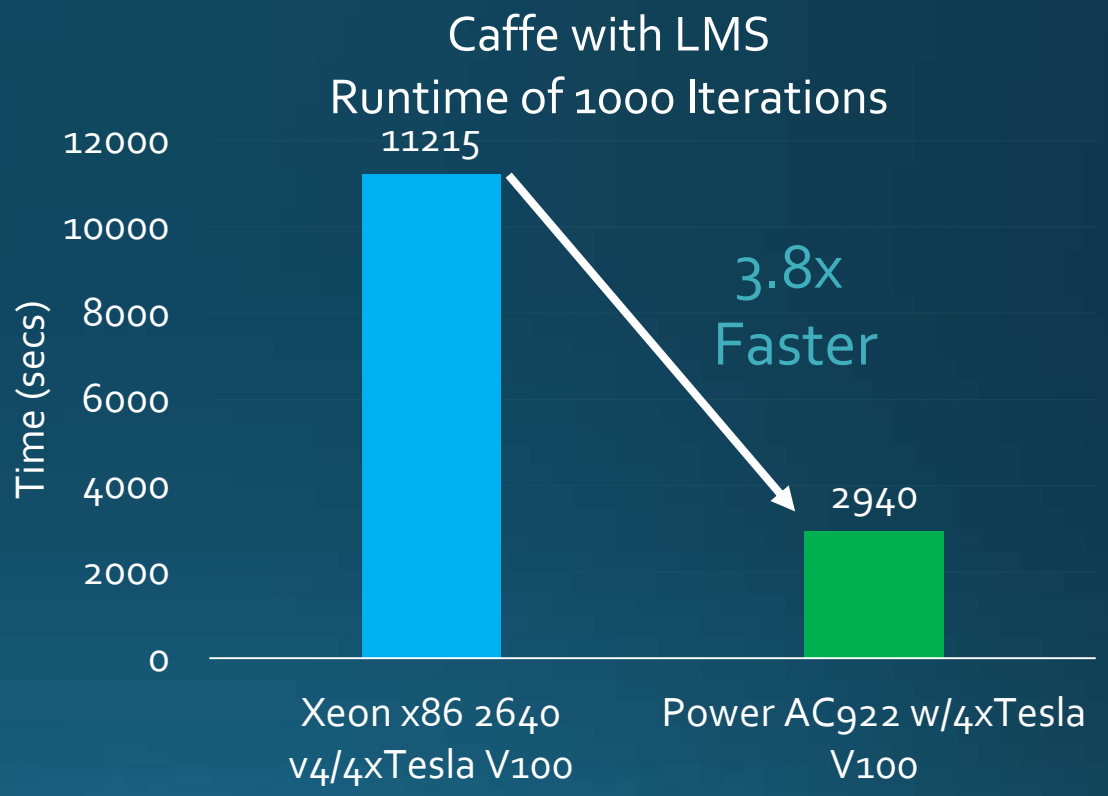
6x Faster CPU-GPU Data Communication

Enables Large Models with Large Input Data

POWER9 with Tesla V100
3.8x Faster than x86 GPU servers

Large Model Support (LMS) Utilizes Fast CPU—GPU NVLink

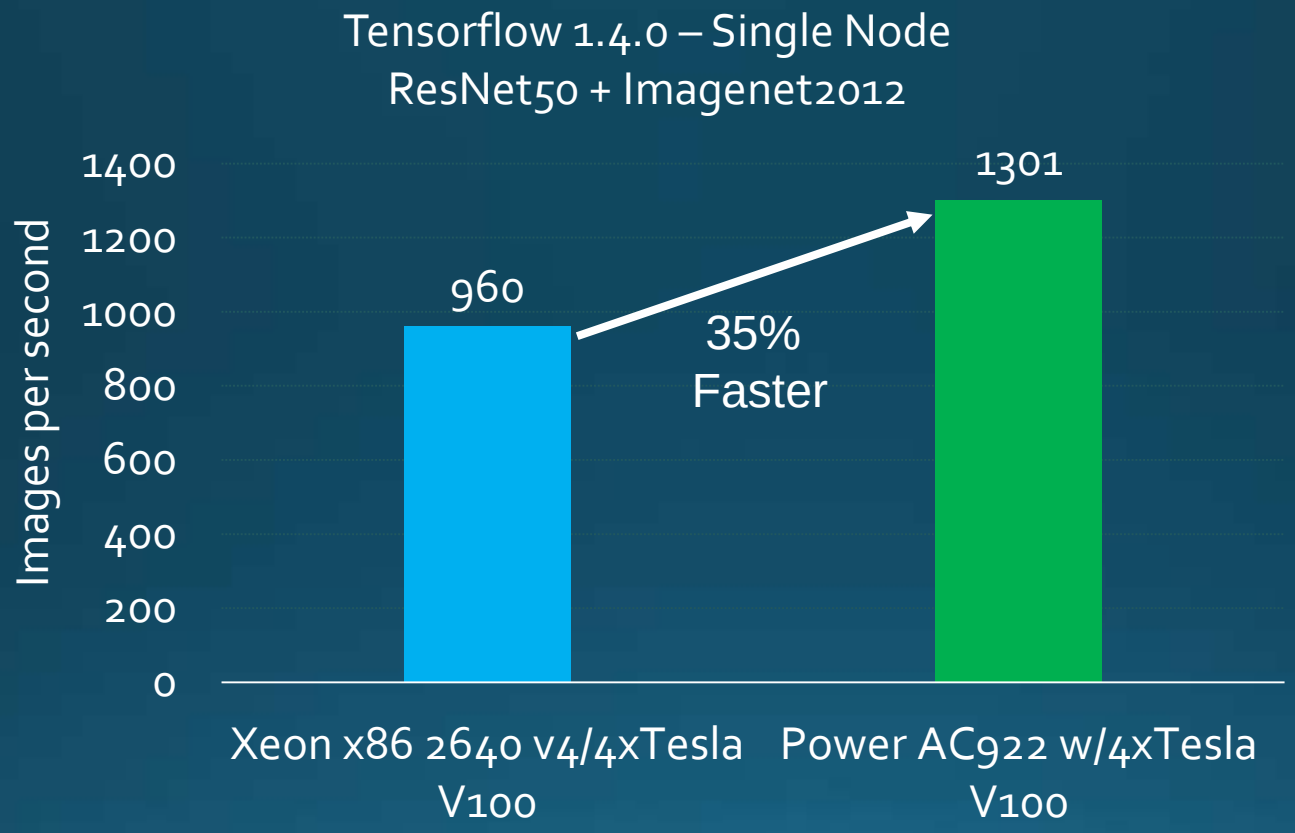




- Hardware: Power AC922; 40 cores (2 x 20c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU Pegas 1.0. Competitive stack: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04.
- Chainer: IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model on Enlarged Imagenet Dataset (2560x2560) .
  - Software: Chainverv3 /LMS/Out of Core with CUDA 9 / CuDNN7 with patches found at https://github.com/cupy/cupy/pull/694 and https://github.com/chainer/chainer/pull/3762
- Caffe Results: IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240x2240) .
  - Software: IBM Caffe with LMS Source code: https://github.ibm.com/TUNG/trlcaffe/tree/1.0-ibm-blc-bm-fix-hang+-p9collateral based on the branch "1.0-ibm-blc-bm-fix-hang+" (base for PowerAI R4) and a PR#5972 from BVLC/Caffe (for supporting cudnn7).

Learn More at
www.ibm.biz/poweraideveloper

POWER9 with Tesla V100
35% Faster than x86 GPU servers



Preliminary Results

- IBM Internal Measurements running 1000 iterations of HPM Resnet50) on 1.2M images and validation on 50K images with Dataset from ILSVRC 2012 aka Imagenet 2012.
- Hardware: Power AC922; 40 cores (2 x 20c chips), POWER9 with NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU ; Red Hat Enterprise Linux 7.4 for Power Little Endian (POWER9). Competitive stack: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4xTesla V100 GPU, Ubuntu 16.04.
- Software: Tensorflow 1.4 framework and HPM Resnet50. Found at mldl-repo-local-esp-5.0.0-5rc4.ppc64le.rpm and https://github.com/tensorflow/benchmarks.gif with the following parameters:Batch-Size: 64 per GPU ; Iterations: 1100; Data: synthetic and imagenet; local-parameter-device: gpu; variable-update: replicated

Machine Learning

Input

Feature Extraction

Features

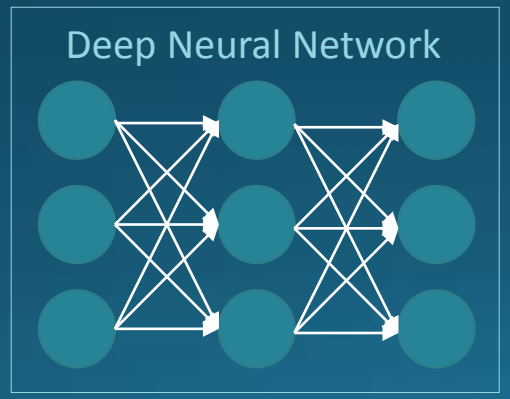
Machine Learning Algorithms

Classification

Output

Deep Learning

Input

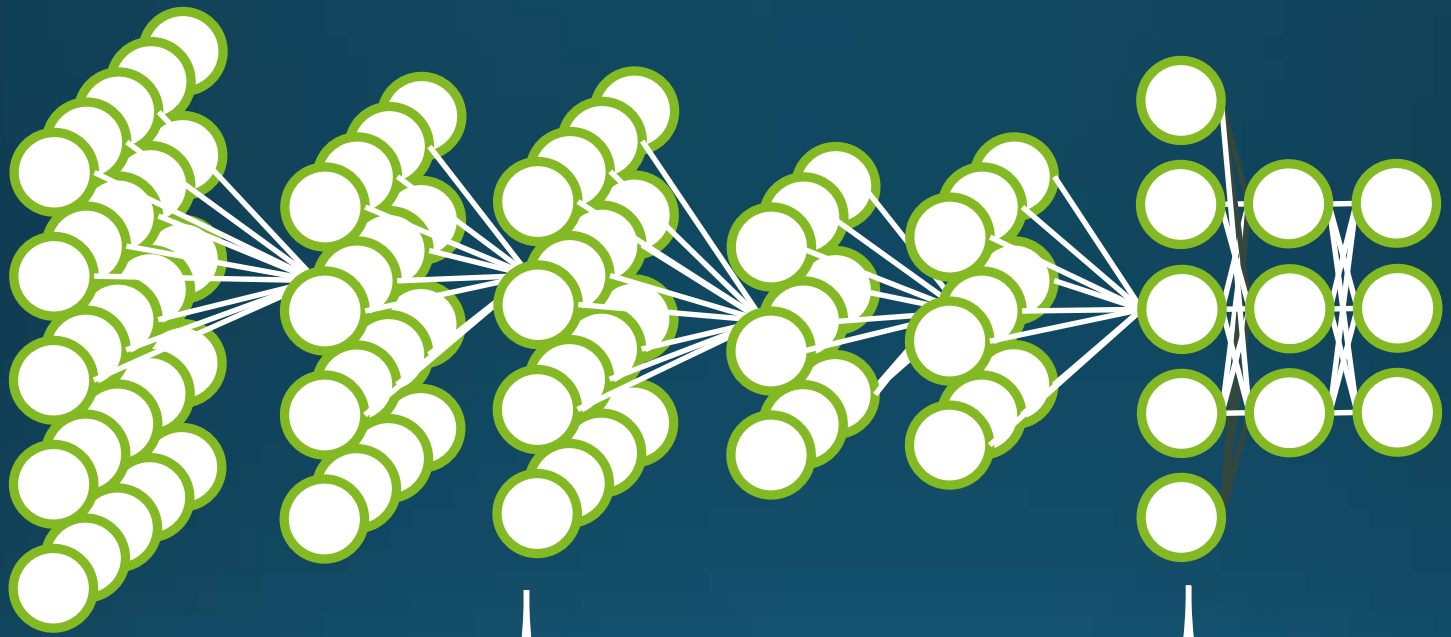


Feature Extraction & Classification

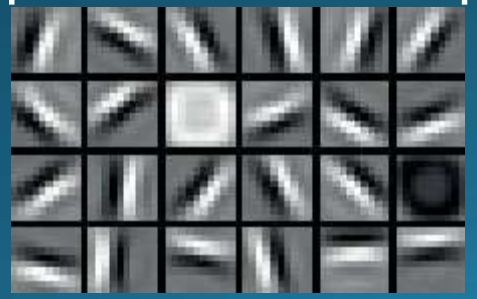
Output

Deep Learning Automatically Figures Out Important Features

Input



Result



Earlier Layers
Detect Edges

Later Layers
Detect Features
Eyes, Nose, Mouth