

PowerAI : Accelerating Deep Learning Adoption in the Enterprise Space

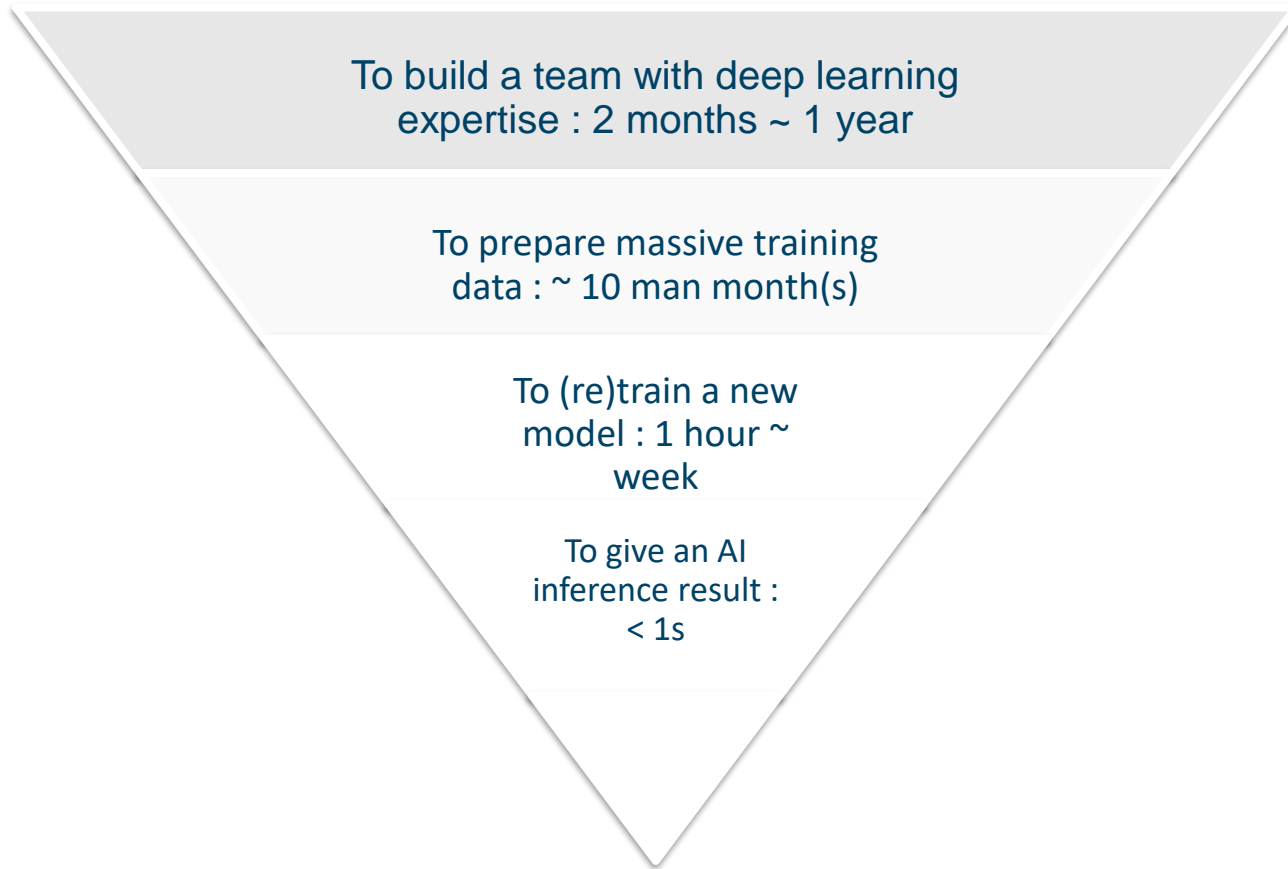
Gilbert Thomas
gilbert@sg.ibm.com



Agenda

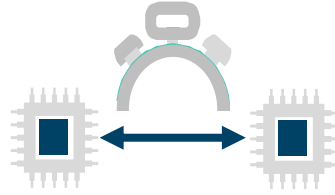
1. Challenges of enterprises in adopting Deep Learning
2. PowerAI – How it addresses those challenges
3. CognitiveClass.AI
4. Conclusion

Reality Check: To enable AI/deep learning capability for enterprises

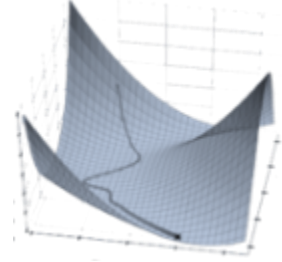




**enterprise-ready
software distribution
built on open source**



**performance
faster training times
for data scientists**

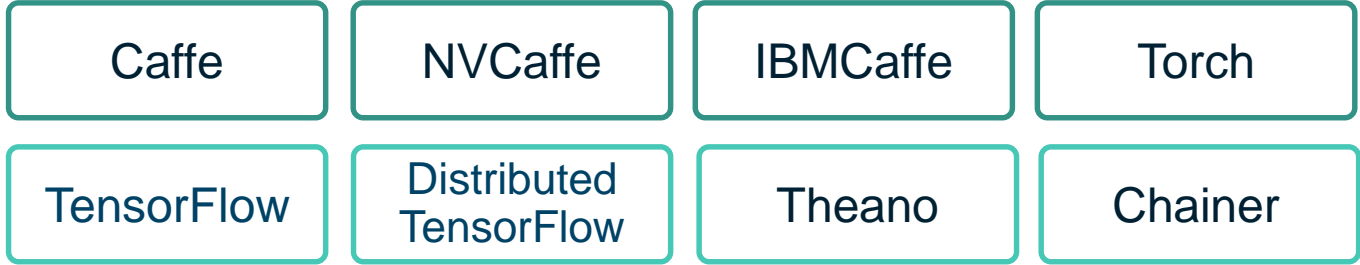


**tools for ease
of development**

IBM PowerAI

PowerAI Deep Learning Software Distribution

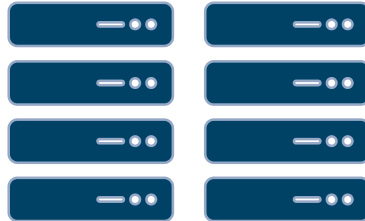
Deep Learning Frameworks



Supporting Libraries



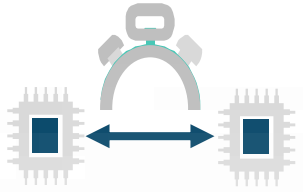
Cluster of NVLink Servers



Spectrum Scale:
High-Speed Parallel
File System



Accelerated Servers and Infrastructure for Scaling

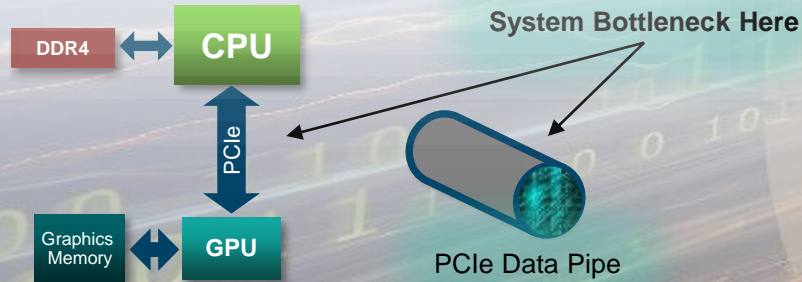


Performance...
Faster Training
and Inferencing

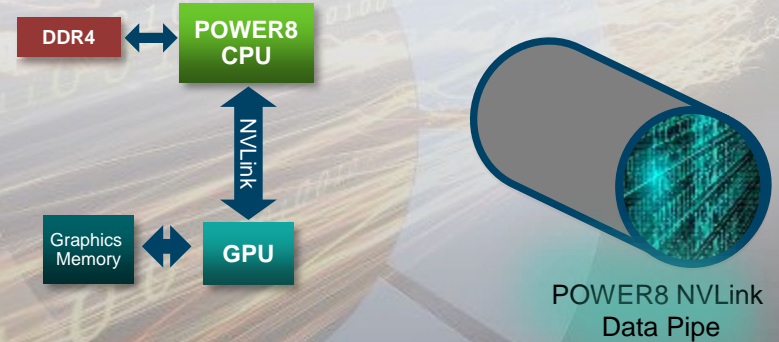
unique innovation through
OpenPower collaboration

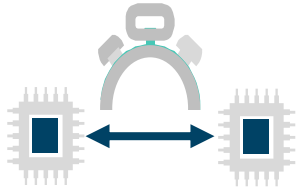
THE NEXT PLATFORM

THE SYSTEM BOTTLENECK SHIFTS TO PCI-EXPRESS



POWER8 with NVLink
delivers 2.8X the bandwidth





Performance... Faster Training and Inferencing

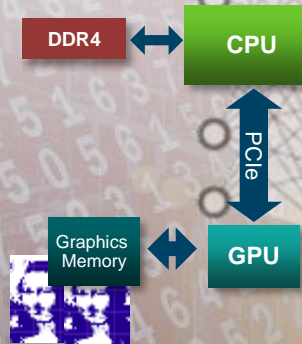
faster training times for data scientists

Distributed Deep Learning



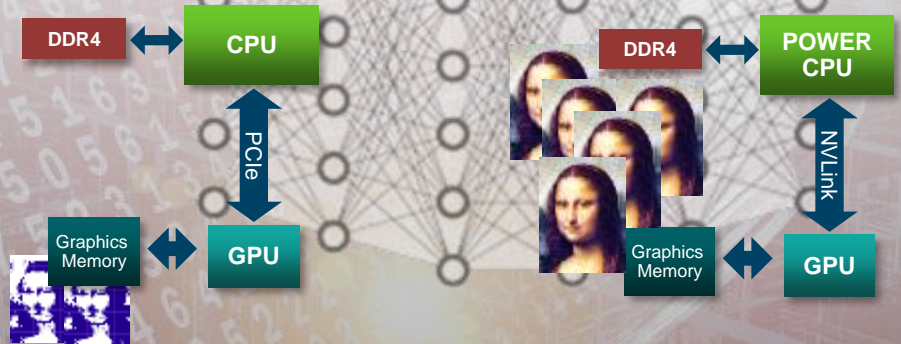
Traditional Model Support

(Competitors)
Limited memory on GPU forces
trade-off in model size / data
resolution



→ *Large Model Support*

(PowerAI)
Use system memory and GPU
to support more complex models
and higher resolution data



IBM Distributed Deep Learning Library fastest in the world

IBM Research

Blog Cognitive Computing Tr

August 8, 2017
Posted in: AI, Cognitive Computing

IBM Research achieves record deep learning performance with new software technology

Summary: IBM Research publishes in arXiv close to i which achieved record communication overhead and framework over 256 NVIDIA GPUs in 64 IBM Power s Facebook AI Research of 89% for a training run on C Research also beat Facebook's time by training the n Using this software, IBM Research achieved a new in trained on a very large data set (7.5M images). The p 29.8% accuracy.

A technical preview of this IBM Research Distributed 4.0 distribution for TensorFlow and Caffe.

Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them

- Home
- Technologies
- Sectors
- Exascale
- Specials
- Resource Library
- Events
- Job Bank
- About

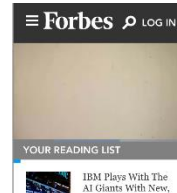
IBM Raises the Bar for Distributed Deep Learning

By Tiffany Trader

August 8, 2017

IBM is announcing today an enhancement to its PowerAI software platform aimed at facilitating the practical scaling of AI models on today's fastest GPUs. Scaling to 256 GPUs with its new distributed deep learning (DLL) library, IBM reports that it has bested previous records set by Google and Facebook on two well-known image recognition workloads.

"This is one of the bigger breakthroughs I have seen in a while in all of the deep learning industry announcements over the last six months," said Patrick Moorhead, president and principal analyst of Moor Insights & Strategy. "The interesting part is that it is from IBM, not one of the web giants like Google, which means it is available to enterprises from on-prem use using OpenPower hardware and PowerAI software or even through cloud provider Nimbix."



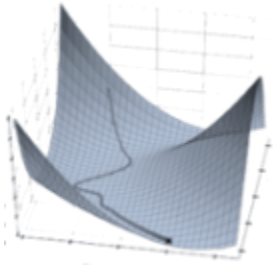
IBM Plays With The AI Giants With New, Scalable And Distributed Deep Learning Software



Patrick Moorhead, CONTRIBUTOR
I write about disruptive companies, technologies and usage models. [FULL BIO](#)

Opinions expressed by Forbes Contributors are their own.





**Tools for Ease
of Development**

**rich advisory and building
toolsets to flatten
time to value**



AI Vision
rich toolset image
recognition neural
networks



automated deep learning
toolkit data preparation

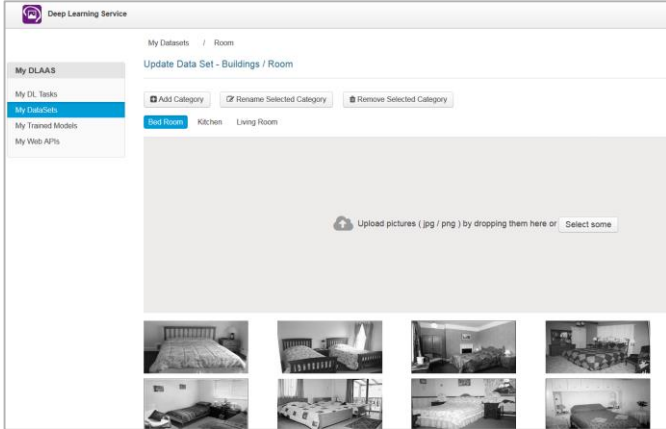


DL Insight toolkit supports
auto-training runs for
hyper parameter tuning

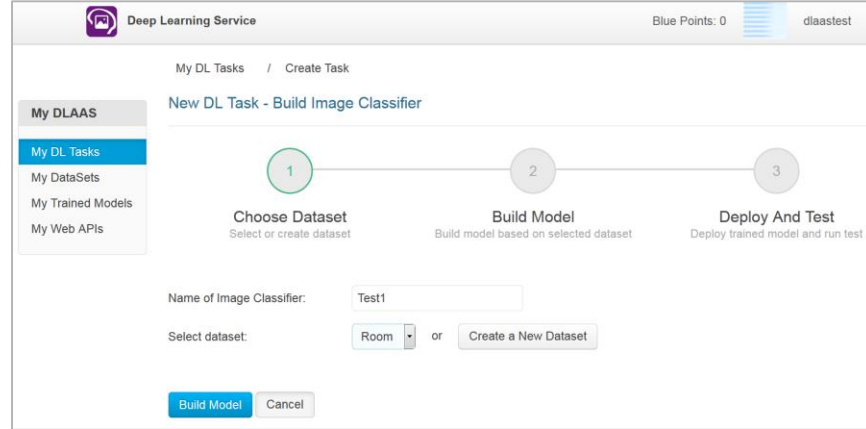
+++

AI Vision – Computer Vision without writing a single line of code!

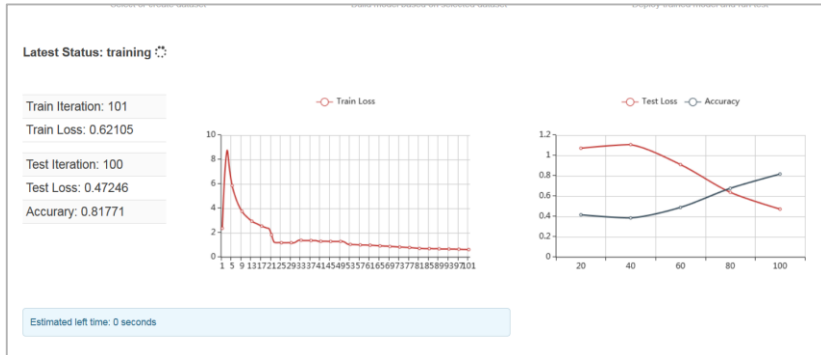
1. User will define the categories and upload data set for new model training



2. Start the model training



3. During the training, AI Vision will monitor the performance data



4. After training, AI Vision will deploy the model as API. User could also do the test with web page.

| URL | Owner | Deployed At | Operation |
|---|----------|------------------|--|
| /webapis/6c9b0242-d9dd-4389-bae9-229c049ecd9 | dlaatest | 2016-05-31 20:32 | Run Test Actions |
| /webapis/3b9261e8-89c7-4659-ace3-261d3138b8a4 | dlaatest | 2016-05-31 11:30 | Run Test Actions |

Assisted Data Preparation for Deep Learning with Spark

Deep Learning

Datasets Model Templates Models Monitor

+ Import Remove

| | | |
|----------------------------------|---|-----------------|
| Lmdb | ↑ | DBbackend |
| TFRrecords | | LMDB |
| Image for Classification | | Rawdata |
| Raw data | | Rawdata |
| CSV | | CSV |
| Image for Object Detection | | CSV |
| <input type="radio"/> mitoesdata | | ObjectDetection |
| <input type="radio"/> qq1 | | LMDB |
| <input type="radio"/> testqq | | CSV |

Import data from different formats

Create New Dataset

* Name tumor-test01

* Training Object Detection Folder /gpfs/dataset/tumor/train

* Validation Percentage 10

* Testing Percentage 10

Resize Object Detection

* Length Percentage 50

* Width Percentage 50

* Split Algorithm hold-out

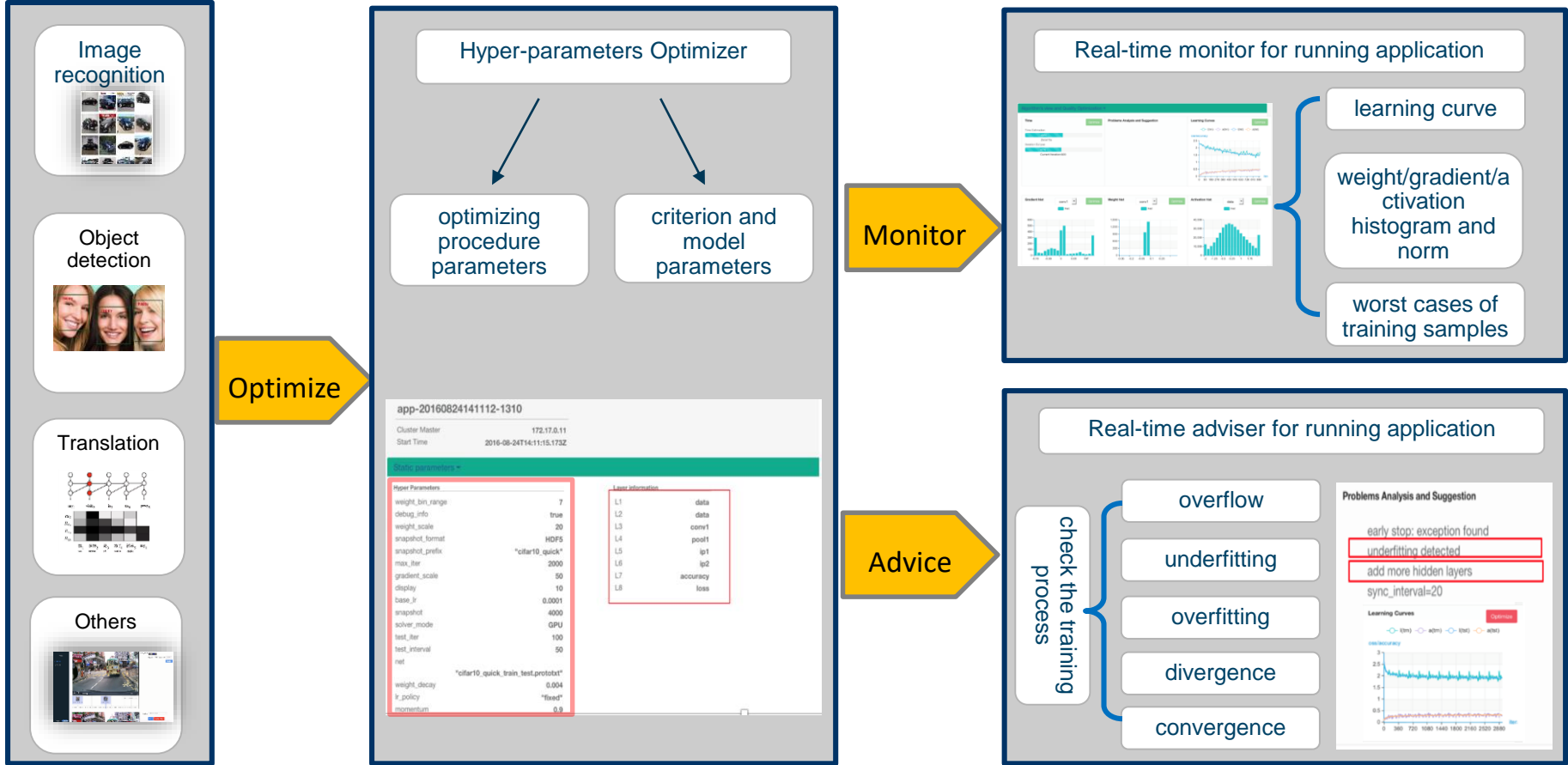
* Spark Master swtest

Add Cancel

Transform, split and shuffle data

DL Insight: Monitor, Adviser and Optimizer Tools for DL Workloads

Deep Learning Applications



➔

Advice

Real-time adviser for running application

check the training process

overflow

underfitting

overfitting

divergence

convergence

Problems Analysis and Suggestion

early stop: exception found

underfitting detected

add more hidden layers

sync_interval=20

Financial Application - Data Workflow



"Data Scientist Playground"

ORACLE®

18 months of Transactions



ETL/Curation/Feature Engineering



Training Data Set

TERADATA

48 months of Account History / Behavior



ETL/Curation/Feature Engineering

Training Data Set



S822LC for Big Data

"Deep Learning training cluster"

Training Data Set

Model Training



Training Data Set

Model Training



Fraud and Risk Classification each Billing Cycle

Ad-Tech / Promotions / Cross-sell / Up-Sell

S822LC for HPC
"Minsky" cluster



PowerAI in the Cloud

Sign up for a Nimbix trial account and get 60 free GPU hours (P100)

<https://www.nimbix.net/powerai>

NIMBIX

Compute

Dog Classification Demos for POWER8

Nimbix, Inc.

From \$5.00/hr

 Machine Learning


inference demonstration software, based on ...



IBM PowerAI: ML/DL environment for POWER8

Nimbix, Inc.

From \$5.00/hr


 Ubuntu Linux ephemeral environment on POWER8 system with IBM...



Kinetica for POWER8

Kinetica

From \$19.35/hr


 A database designed from the ground up to deliver truly real...



NVIDIA® DIGITS 5 for IBM POWER8

NVIDIA

From \$5.00/hr


 The NVIDIA Deep Learning GPU Training System (DIGITS) puts t...



Torch/cutorch with OpenMPI for POWER8

Nimbix, Inc.

From \$5.00/hr


 Ubuntu Linux environment on POWER8 system with Torch/cutorch...



Ubuntu Linux for POWER8

Nimbix, Inc.

From \$5.00/hr

 Ubuntu Linux ephemeral environment on POWER8 system with NVI...



<https://cognitiveclass.ai/courses/>

Analytics, Big Data, and Data Science Courses

Your awesome career in Data Science and Data Engineering starts here.

SIGN UP



Takeaway: Why IBM PowerAI?

- Very fast time to market , server comes ready for deep learning
- Leverages unique Power8 CPU-GPU NVLink communications on “Minsky” and P100 GPUs for faster training
- Overcome GPU memory limits to run larger, more complex models
- Best performing distributed deep learning performance in the market with near linear scalability
- Tools to lower barrier to entry for novice data scientists trying to get into deep learning
- Free “A.I” courses to address the skill gap in the market.

Thank You

© 2010 IBM Corporation
All rights reserved. IBM, the IBM logo, and
"Think" are trademarks of International Business
Machines Corporation. All other trademarks are
the property of their respective owners.



Drones/Deep Learning Use Case with IBM

Research Institute

- Research division of Telco
- Develop future technologies to enhance their Competence

Use Case

- **Drones** for visual inspection of **42,372** transmission towers
- To save **6.59 million** a year

Requirement & Challenge

- To process vast amount of images from drones
- Based on **Open Standard and end-to-end solutions**
- Superior and Highly RAS GPU-Based NVLink solutions
- High speed concurrent filesystems Storage and network

IBM Cognitive Systems Solutions

- IBM PowerAI Software
- IBM Spectrum Scale Software
- IBM OpenPower S822LC (Minsky Server)
- IBM Elastic Storage Server
- IBM Cluster Switch 100 Gbps EDR Infiniband (Mellanox)

Benefit Objective



822LC Power System for HPC

First Custom-Built GPU Accelerator Server with NVLink

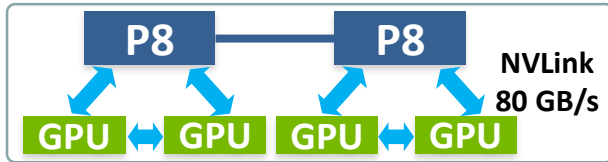


- Custom-built GPU Accelerator Server
- High-Speed NVLink Connections between CPUs & GPUs and among GPUs
- Features NVIDIA P100 Pascal GPU accelerators
- 2.5x faster



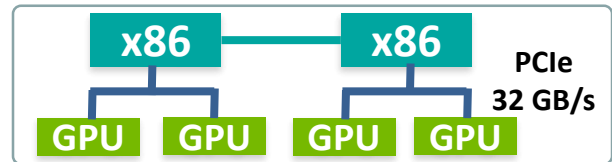
NVIDIA P100 Pascal GPU

2.5x Faster CPU-GPU Data Communication via NVLink



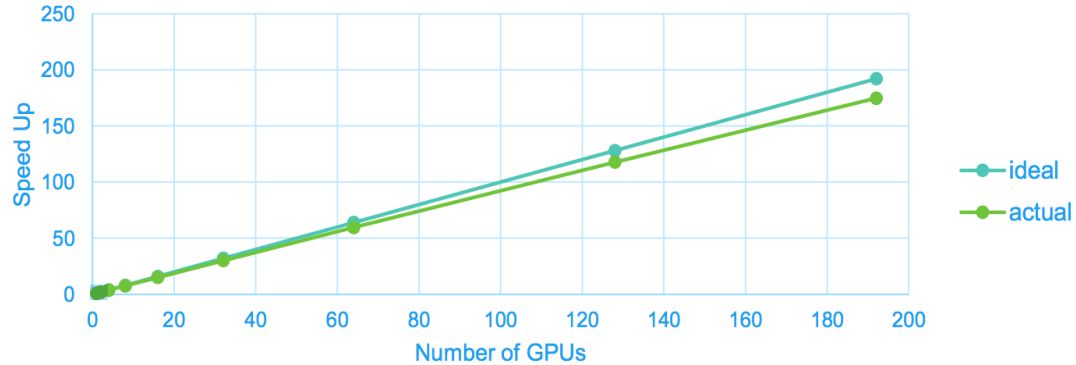
POWER8 NVLink Server

No NVLink between CPU & GPU for x86 Servers: PCIe Bottleneck



x86 Servers with PCIe

Near Linear Scaling for Caffe across 192 CPUs



Seconds per 100 iterations for 32 images →

| GPUs | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 192 |
|--------------------|-------|-------|-------|-------|-------|-------|-------|--------|--------|
| IBMCaffe-prc | 33.93 | 34.77 | 35.54 | 35.54 | 35.89 | 36.22 | 36.45 | 36.89 | 37.29 |
| Speedup | 1.00 | 1.95 | 3.82 | 7.64 | 15.13 | 29.98 | 59.58 | 117.73 | 174.70 |
| Scaling efficiency | 1 | 0.98 | 0.95 | 0.95 | 0.95 | 0.94 | 0.93 | 0.92 | 0.91 |

Legal Notices

Copyright © 2017 by International Business Machines Corporation. All rights reserved.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-785
U.S.A.