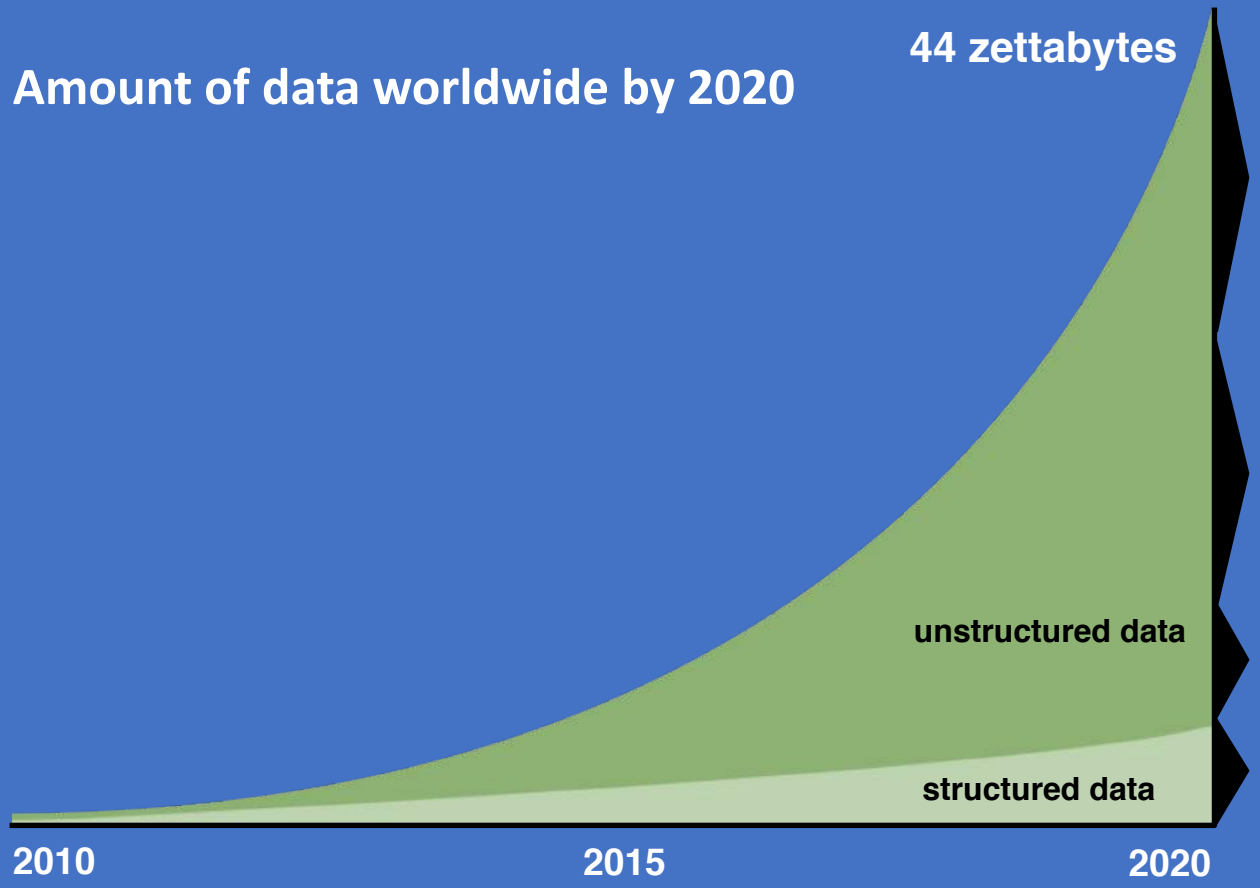
Deep Learning Insurgency

Data holds competitive value

What your data would say if it could talk…



- Internet Of Things
- Images & Multimedia
- Text
- Enterprise Data

- Mobile
- Medical
- Oil/Gas
- Weather

163 zettabytes by 2025 – Source: IDC

Today’s Challenges Demand Innovation

- Innovations from diverse ecosystem partners
- Moore’s Law fading
- Heterogeneous computing for today’s and tomorrow’s workloads
- Accelerator-assisted computing

IBM AI Systems Are Built With Optimized HW & SW

Dev Ecosystem

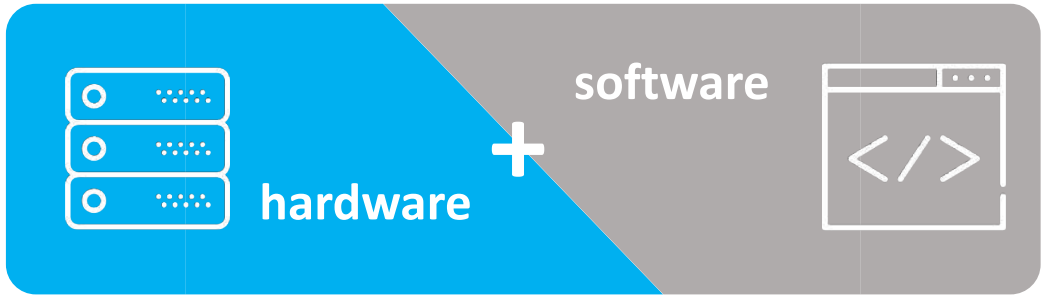
- Industry Alignment
- Partnerships
- Open Frameworks
- IBM Software
- Open Source Software

POWER8 → POWER9 → POWER10

- Open Accelerator Interfaces
- Accelerator Roadmaps

Not Just About Hardware Design

It's about co-optimized



IBM Was Awarded Two U.S. Department of Energy CORAL Contracts

CORAL:

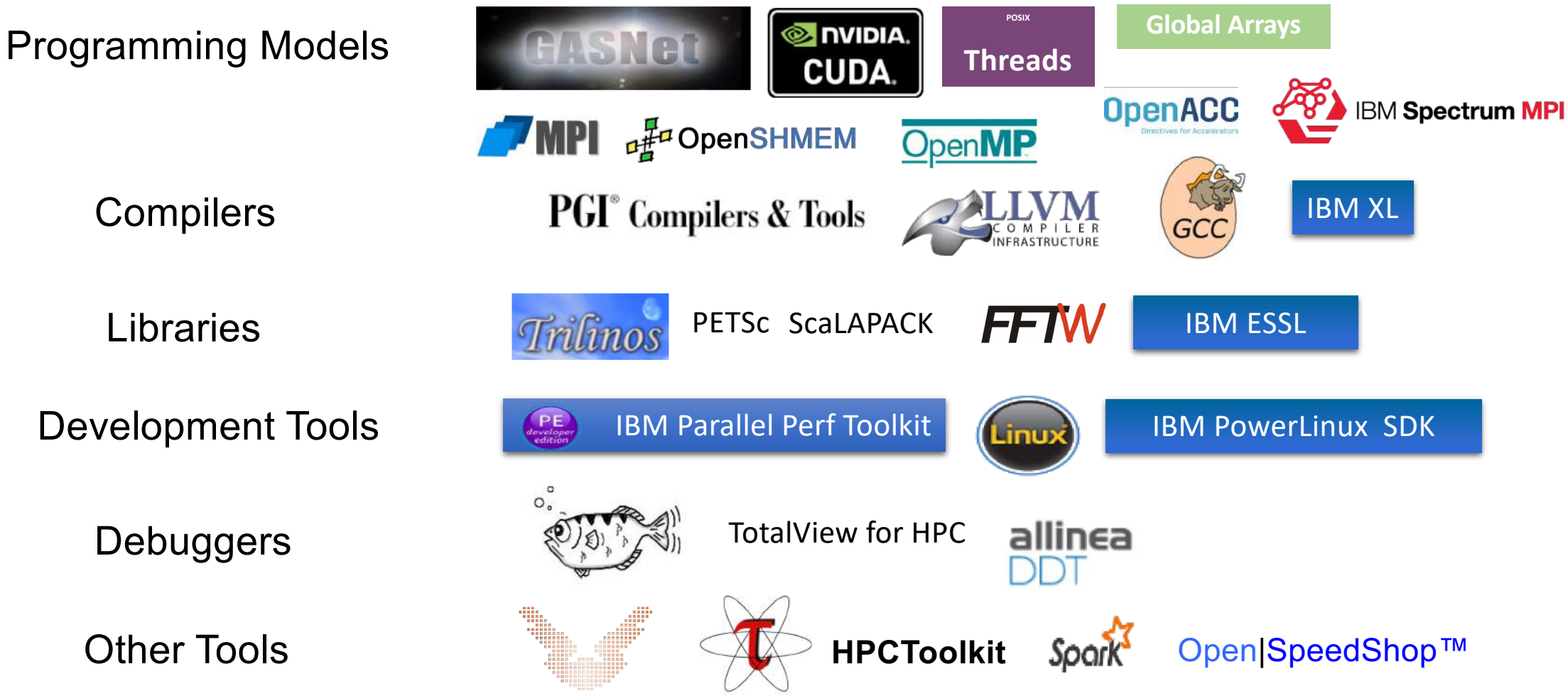
Collaboration of Oak Ridge, Argonne, and Livermore

Two supercomputers for Oak Ridge and Lawrence Livermore Labs in 2017/2018.

IBM System at Oak Ridge National Laboratory

- IBM AC922 servers
- Most powerful and smartest computer in the world
- 200 PFLOPS
- 250 PB usable storage
- 4608 nodes
- 9216 IBM POWER9 processors
  - 202752 cores
- 27648 Nvidia V100 GPUs
- Mellanox EDR

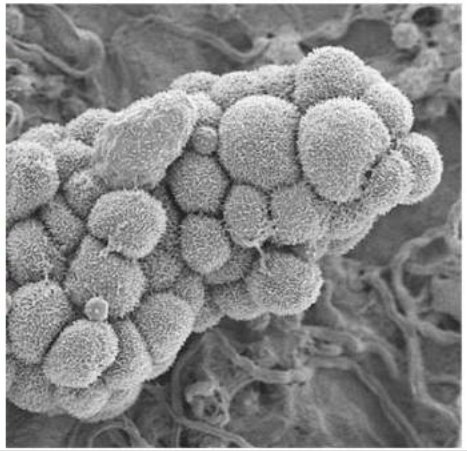
Various Tools Form a Key Part of an Open Source and IBM-Proprietary Software Ecosystem



AI Is Far and Wide

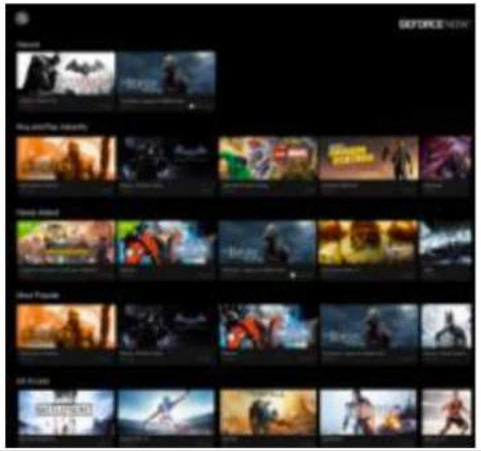
INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading Drug Discovery



MEDIA & ENTERTAINMENT

Video Captioning Video Search Real Time Translation

SECURITY & DEFENSE

Face Detection Video Surveillance Satellite Imagery

AUTONOMOUS MACHINES

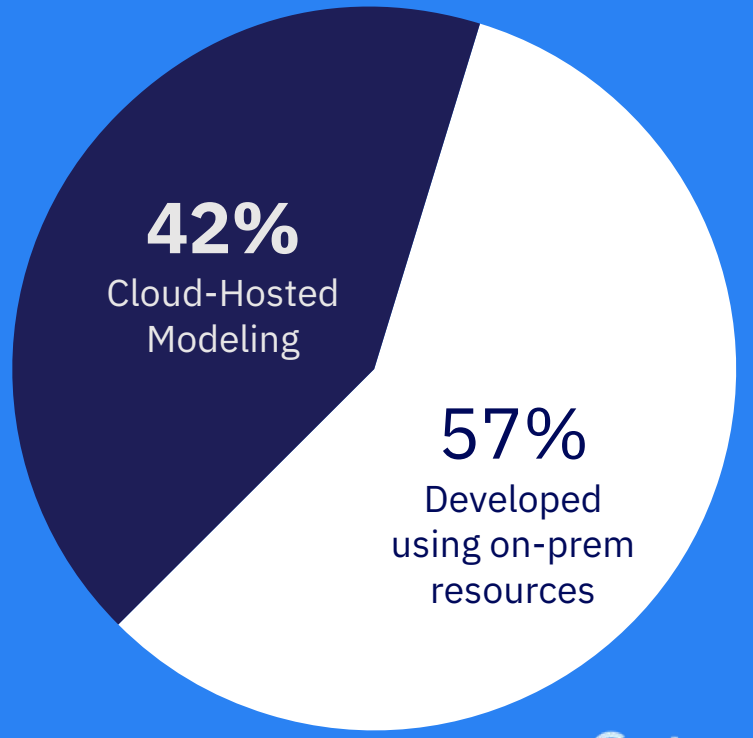
Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

Enterprises Embracing Open-Source AI Software

- Enterprises building Machine Learning teams
- Most using Open-Source software: TensorFlow is most popular
- IDC 2021 Market Size Projection: $14B for AI Servers

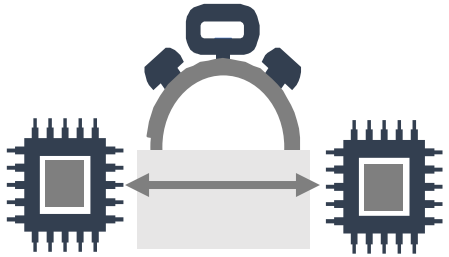
57%

AI Developed On-Premise

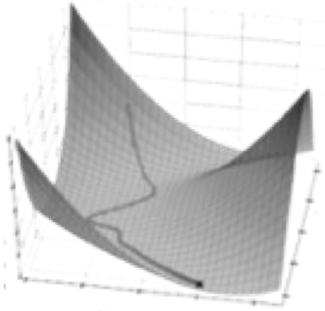


Gartner

enterprise-ready
software distribution
built on open source



performance:
faster training times
for data scientists



tools for
ease of
development

IBM PowerAI

IBM PowerAI Enterprise

Original design: Simplify the process of installing and running optimized Deep Learning on Power

Integrated & Supported AI Platform

TensorFlow Caffe

Keras PYTORCH

IBM PowerAI
Enterprise

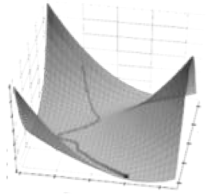
Faster Time to Results

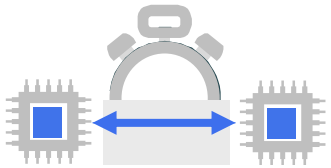
Increased Resource Utilization

Simplified Management

Enterprise Solution

IBM PowerAI Enterprise





Enable non-Data Scientists to use AI
(Tools for ease of use)

Integrated & Supported AI Platform

TensorFlow Caffe

Keras PYTORCH

Higher Productivity for Data Scientists
(Faster Training with Larger Models)

Bringing AI to Production

PowerAI Enterprise

Deep Learning Impact
Data Management and ETL
Training visualization and monitoring
Hyper-parameter optimization

Spectrum Conductor
Multi-tenancy support & security
User reporting & charge back
Dynamic resource allocation
External data connectors

Distributed Deep Learning (DDL)

Support Line L1-L3

PowerAI (Base)

Open Source Frameworks: Supported Distribution

TensorFlow™ Caffe TensorFlow Keras Chainer PYTORCH

Large Model Support

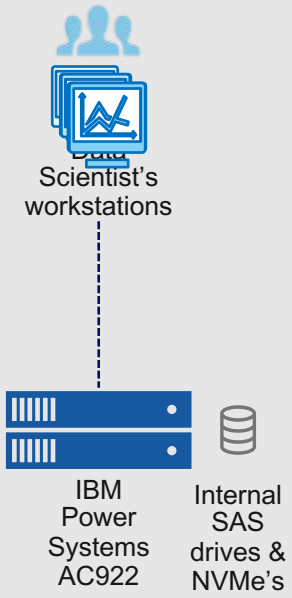
IBM

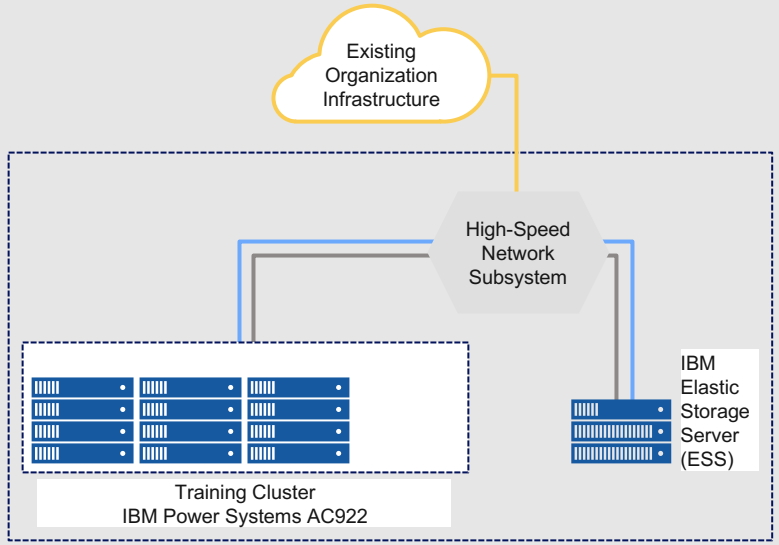
IBM AI Architecture from Experimentation to Expansion

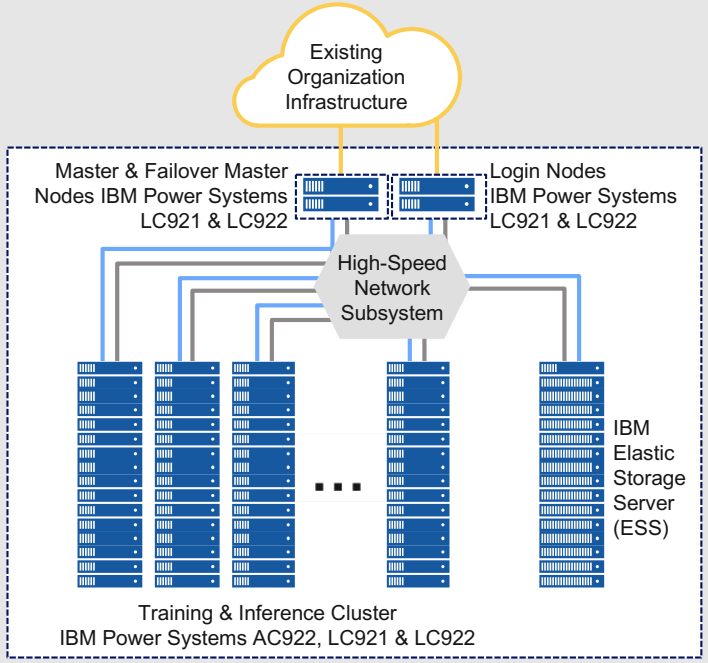
Experimentation
Single Tenant

Stabilization & Production
Secure Multitenant

Expansion
Enterprise Scale / Multiple Lines of Business





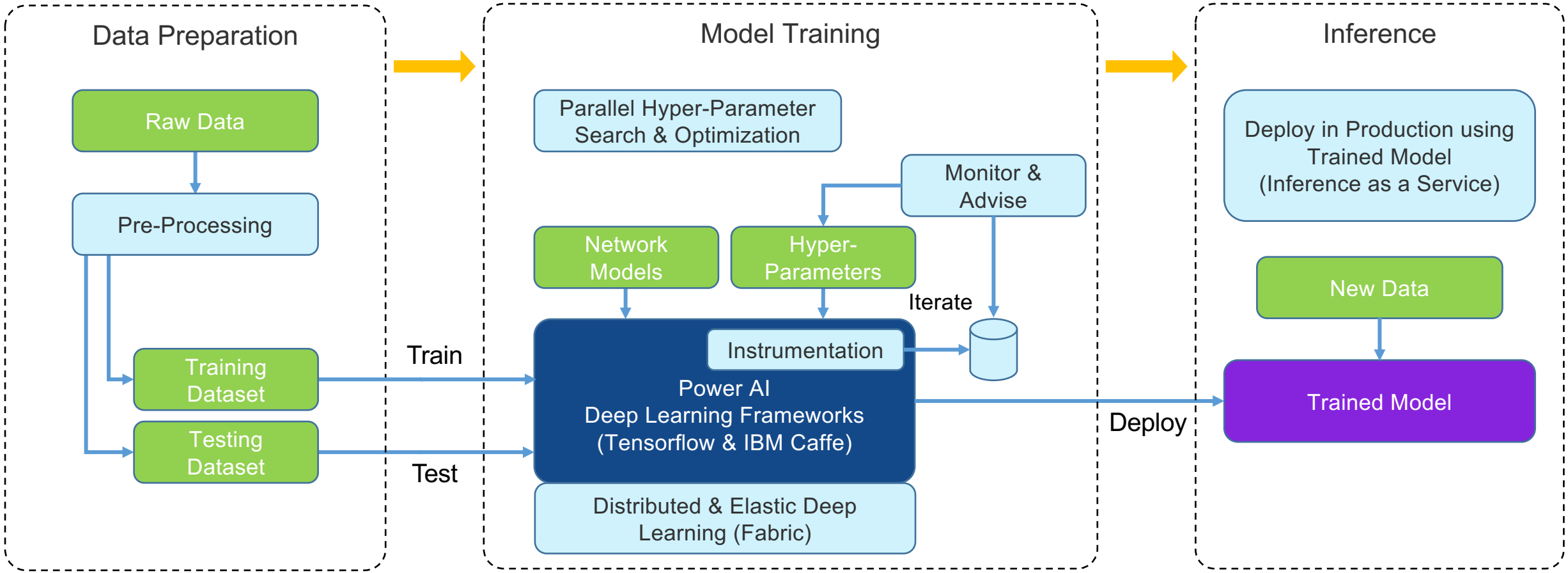


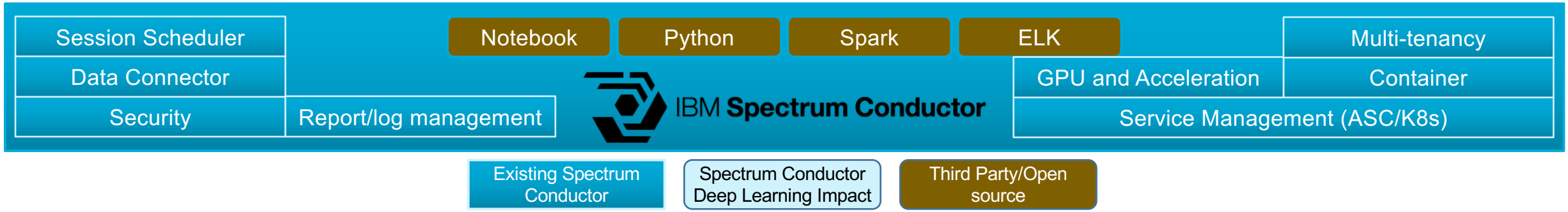
Services & Support

- IBM **PowerAI Enterprise**
- Red Hat Enterprise Linux (RHEL)
- IBM **Power System** & x86 Servers
- IBM **Spectrum Scale** / IBM **Elastic Storage Server** (ESS)

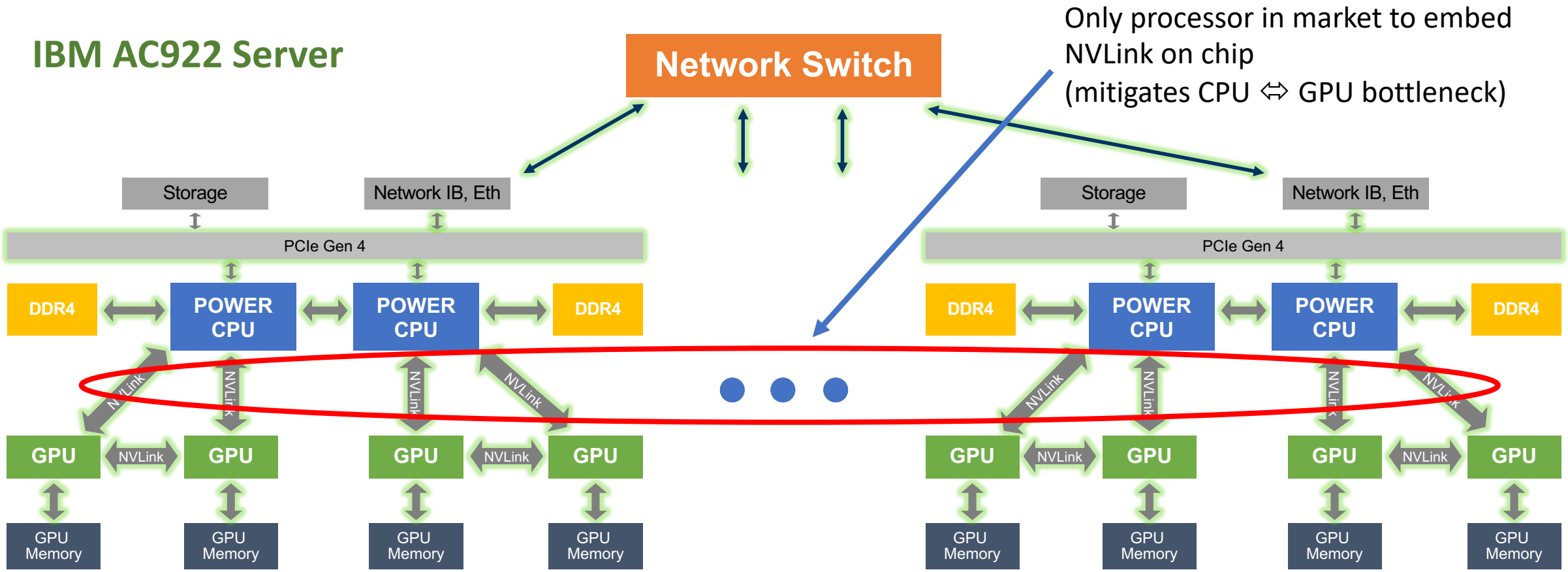
One software stack from experimentation to expansion

IBM Spectrum Conductor & Deep Learning Impact





COMMUNICATION PATHS

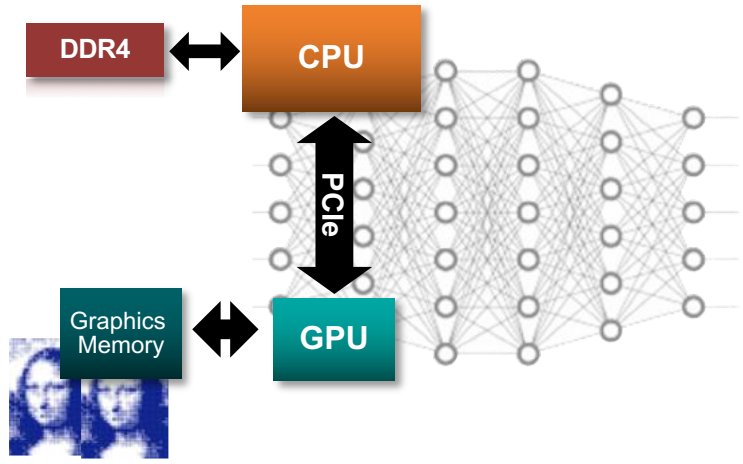


Power AI DDL: Fully utilize bandwidth for links within each node and across all nodes → Learners communicate as efficiently as possible

Train Larger More Complex Models

Traditional Model Support

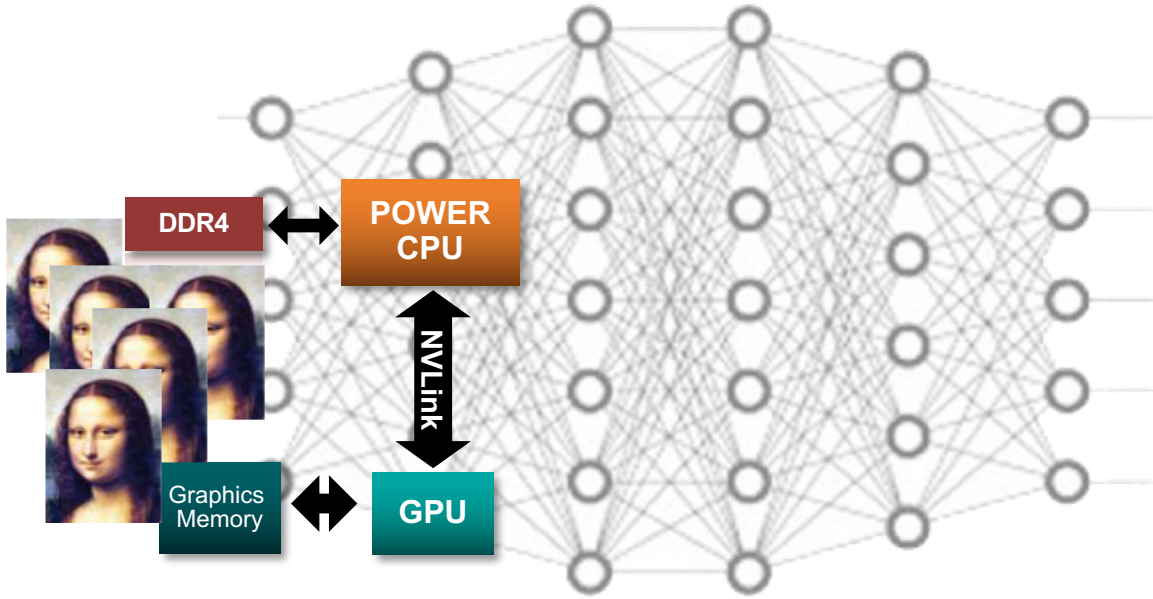
Limited memory on GPU forces trade-off in model size / data resolution



Large Model Support

IBMPowerAI

Use system memory and GPU to support more complex and higher resolution data



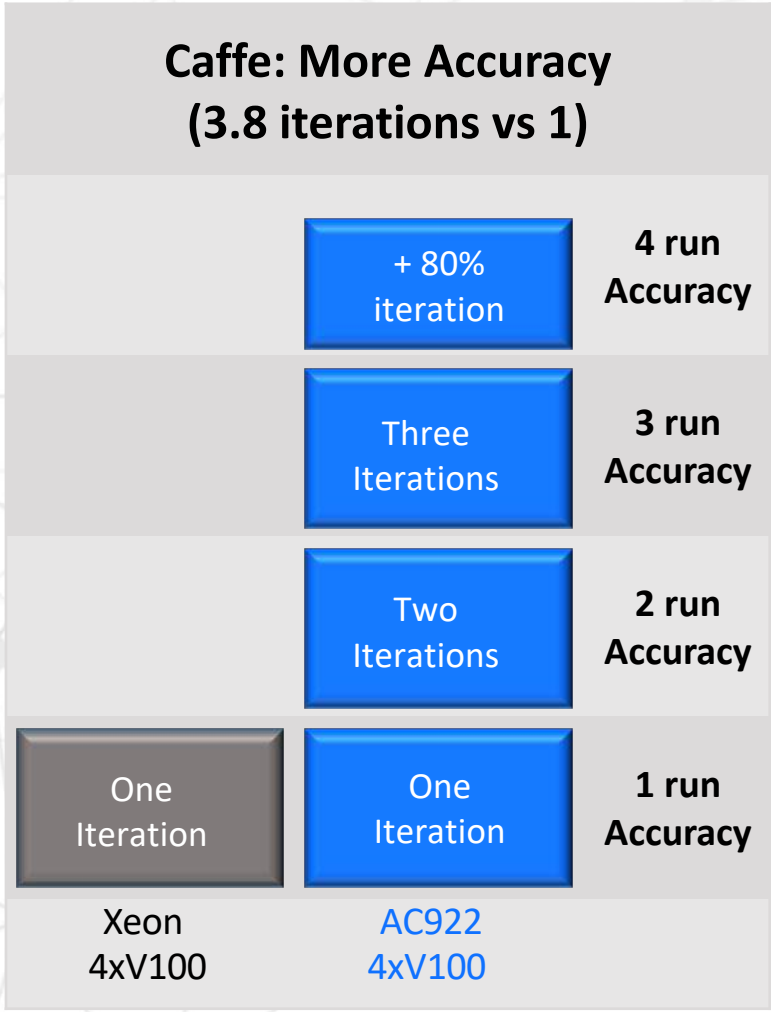
IBM

Designed for the AI era: Caffe provides a 3.8X reduction in AI model training vs tested x86 systems

Caffe

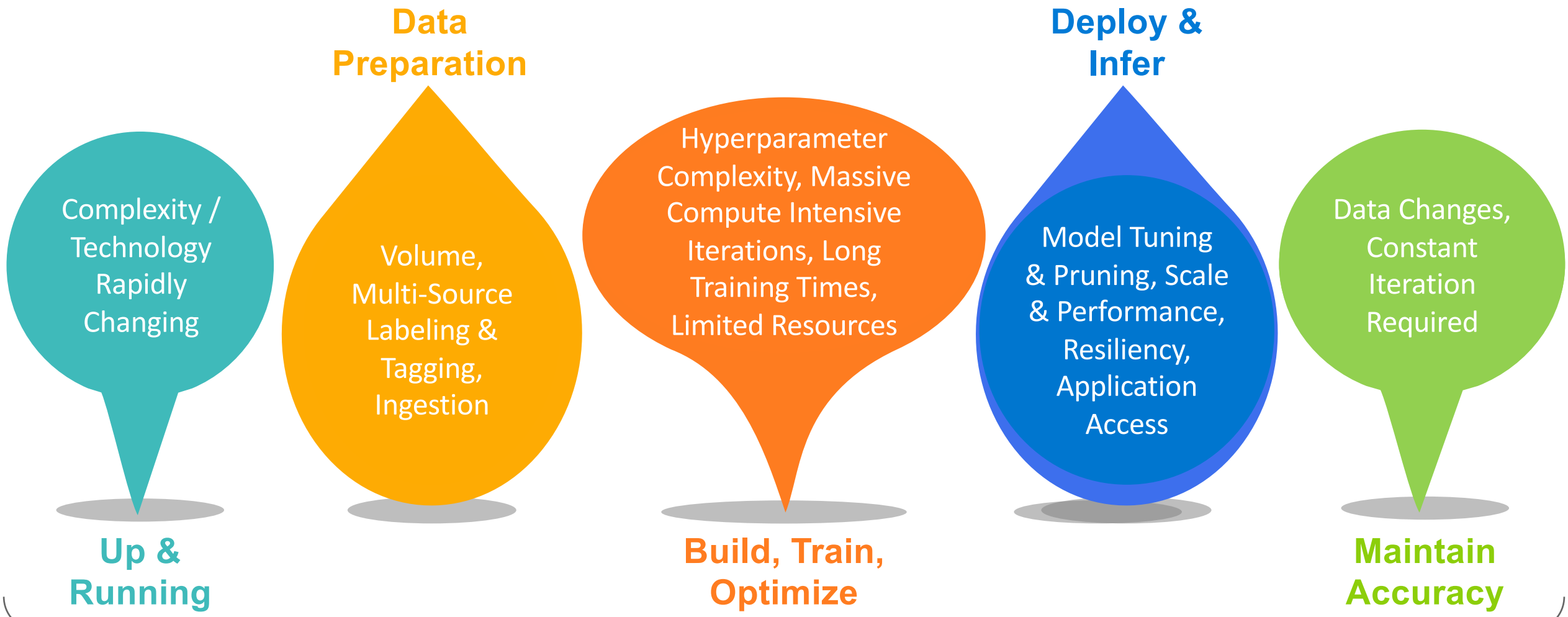
Maximize research productivity running training for medical/satellite images with Caffe with the AC922

- **3.8X reduction vs tested x86 systems** 1000 iterations running on competing systems to train on 2k x 2k images
- Critical machine learning (ML) capabilities such as regression, nearest neighbor, recommendation systems, clustering, etc. operate on more than just the GPU memory
  - NVLink 2.0 enables enhanced Host to GPU communication
- **Large Model Support** - use system memory and GPU memory to support more complex and higher resolution data



- Results are based IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240x2240) .
- 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) with CUDA 9.1/ CUDNN 7;. 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. with CUDA .9.0/ CUDNN 7 .
- Software: IBM Caffe with LMS Source code https://github.com/ibmsoe/caffe/tree/master-lms

Pain Points – Deep Learning Pipeline



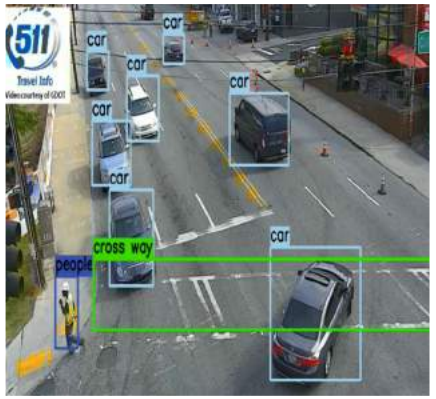
Share valuable resources across multiple users, lines of business & applications with security & resiliency at scale

PowerAI Vision

- Provides a complete ecosystem to label raw data sets for training, creating, and deploying deep learning-based models
- Designed to empower Subject Matter Experts with no skills in deep learning technologies to train models for AI applications
- Quickly train highly accurate models to classify images and detect objects in images and videos

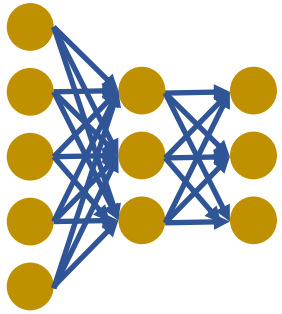
Semi-Automatic Labeling using PowerAI Vision

Manually Label



Define Labels Manually Label Some Images / Video Frames

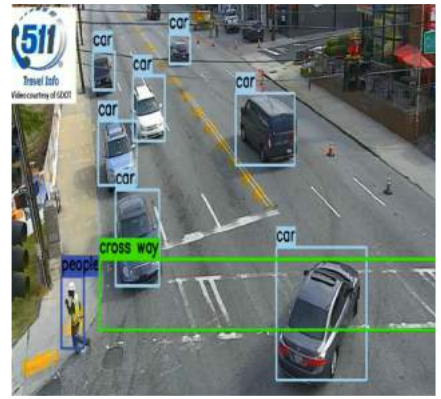
Train DL Model



Use Trained DL Model

Run Trained DL Model on Entire Input Data to Generate Labels

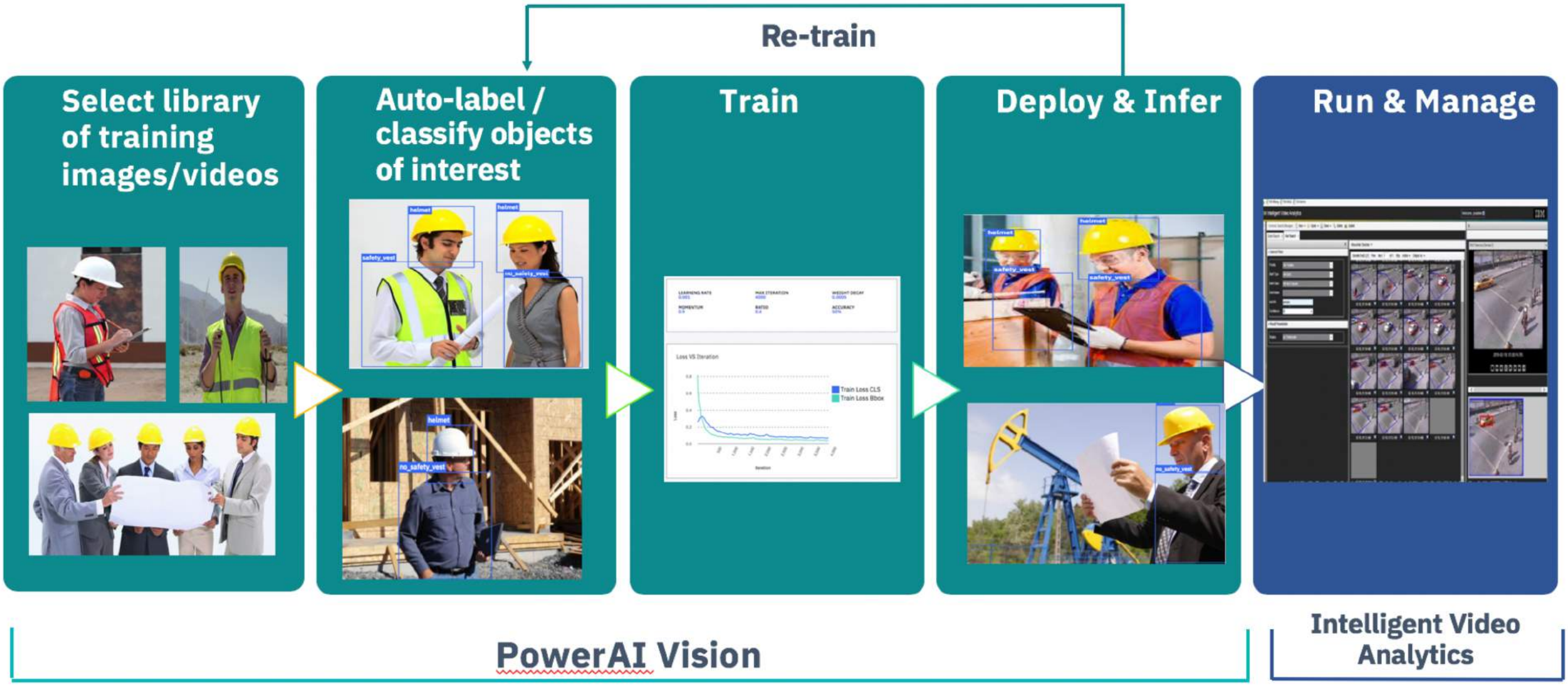
Correct Labels on Some Data



Manually Correct Labels on Some Data

Repeat Until Labels Achieve Desired Accuracy

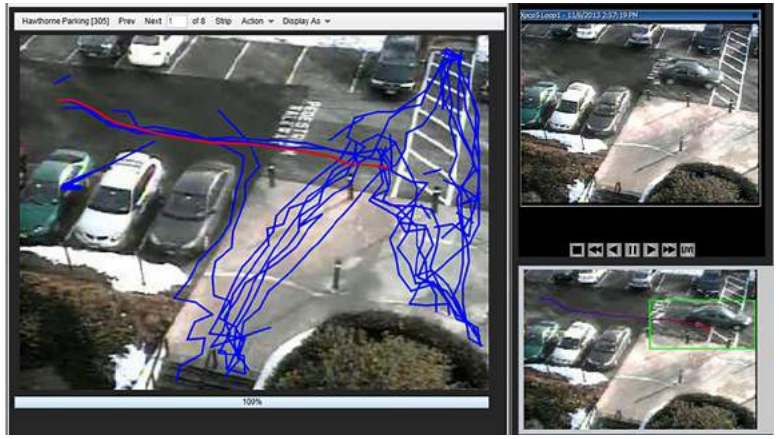
Customizing Video Analytics with PowerAI Vision



IBM Intelligent Video Analytics (IVA)

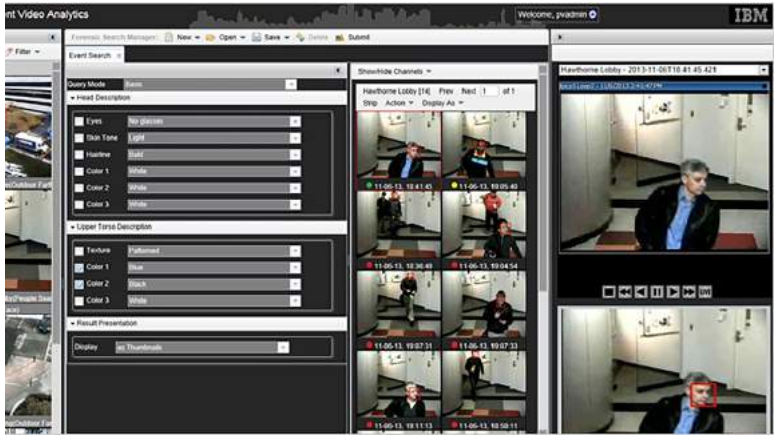
- Video Analytics Software with Pre-Trained AI Models
- Complex Event Monitoring with GUI-based Configuration
- Targeted at Public Safety, Remote Monitoring, etc

Detect Changes to Patterns

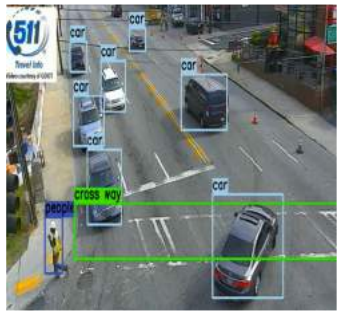


Redaction of Faces

Facial Recognition & People Search



PowerAI Vision + IVA Integration



New Video Training Data

PowerAI Vision

- Label
- Train
- Deploy

GPU-Accelerated Power Servers for Model Training

New AI Model

Live Video

IBM IVA

- GUI
- AI Models
- Event Detection

GPU-Accelerated Power Servers for Inference

Top Reasons to Choose PowerAI

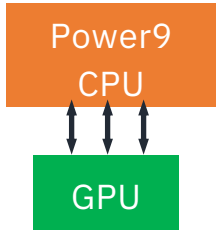
Simplicity: Integrated Platform that Just Works

Curate, Test, and Support Fast Moving Open Source

Provide Enterprise Distribution on RedHat

Easy to deploy Enterprise AI Platform

Ease of Use, Unique Capabilities



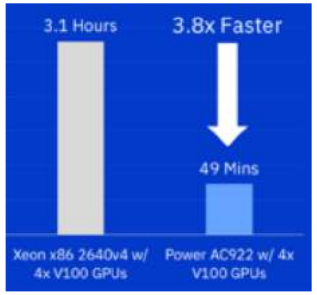
Large data & model support due to NVLink

Acceleration of Analytics & ML

AutoML: PowerAI Vision

Elastic Training: Scale GPUs as Required

Faster Model Training Time

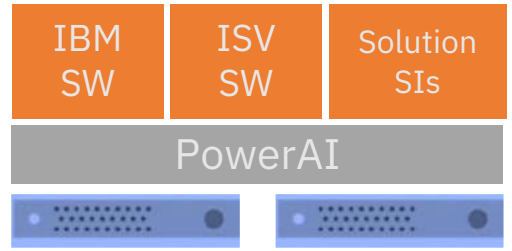


Faster Training Times in Single Server

Scalability to 100s of Servers (Cluster level Integration)

Leads to Faster Insights and Better Economics

Open AI Platform w/ Ecosystem Partners



Platform that Partners can build on

Software Partners: H2O, IBM, Anaconda

SIs, Solution Vendors & Accelerator Partners

Summary

- Deep Learning is driving a new class of workloads
  - Driving new architectures within existing IT environments
- IBM Power is an OPEN platform
  - OpenPOWER Foundation (founded in 2013) – openpowerfoundation.org
  - PowerAI
    - open-source Deep Learning frameworks highly optimized
- IBM is driving value across the stack
  - Hardware Platforms
  - Software Platforms
  - Deep Learning Frameworks / Algorithms / Apps
  - Services / Consulting