IBM PowerAI Vision
with Xilinx Alveo U200 Accelerator Card

Presented By

Amartey Pearson
Sr. Technical Staff Member - IBM PowerAI Development
apearson@us.ibm.com
Oct 2\textsuperscript{nd}, 2018
Power Systems?

Summit by the numbers
200: Petaflops (quadrillion calculations per second)
9,216: Number of IBM POWER9 CPUs.
4,608: Number of nodes.
“With the IBM Power Systems AC922 server, IBM has already demonstrated that we have the best platform for enterprise AI training.

IBM sees inference as a key component of a complete, end-to-end AI platform, and POWER9's leadership I/O bandwidth for data movement makes it an ideal pairing with Xilinx’s new Alveo U200 accelerator card to bring inference to the enterprise.”

- Steve Sibley, Vice President of IBM Cognitive Systems.
Large AI Models Train ~4 Times Faster

POWER9 Servers with NVLink to GPUs vs x86 Servers with PCIe to GPUs

Caffe with LMS (Large Model Support) Runtime of 1000 Iterations

- Xeon x86 2640v4 w/ 4x V100 GPUs: 3.1 Hours
- Power AC922 w/ 4x V100 GPUs: 49 Mins

GoogLeNet model on Enlarged ImageNet Dataset (2240x2240)

Detailed Benchmark Information in Back
PowerAI
Vision

Auto-ML for Images & Video

- Label
- Train
- Deploy

PowerAI: Open Source ML Frameworks

- TensorFlow
- Caffe
- torch
- Chainer

Large Model Support (LMS)

PowerAI: Enterprise

Deep Learning Impact (DLI) Module

- Data & Model Management, ETL, Visualize, Advise

PowerAI

- Distributed Deep Learning (DDL)
- Auto ML

IBM Spectrum Conductor with Spark
Cluster Virtualization, Auto Hyper-Parameter Optimization

Accelerated Infrastructure

- Accelerated Servers
- Storage
PowerAI Vision: "Point-and-Click” AI for Images & Video

Label Image or Video Data

Auto-Train AI Model

Package & Deploy AI Model

![Image of label and AI model training process]

- **Data set**: Various types of images and videos
- **Objects**: Cars, people, crosswalk
- **Trained model / fasterrcnn-model**: Details of the trained model
  - Learning rate: 0.0001
  - Momentum: 0.9
  - Ratio: 0.8
  - Accuracy: 98%
- **Loss vs Iteration**: Graph showing the learning process

![Image of package and deployment]

- **Export model**: Options for deploying the model
- **IBM PowerAI Vision**: Platform for managing and deploying AI models

![Image of Xilinx product]

- **Xilinx**: Partner for hardware acceleration of AI models
Core Capabilities

> Manage and Label datasets - images and video

> Classification of Images
  o Identify object in an image and classify to one of the trained categories

> Detect Objects
  o Analyze Images and Videos and detect several trained objects

> Auto labeling for rapid processing of raw datasets
  o Partially trained models auto label raw datasets to create larger datasets

> Data Augmentation

> RESTful APIs to develop custom solutions and extend existing applications
Semi-Automatic Labeling using PowerAI Vision

1. **Manually Label**
   - Define Labels
   - Manually Label Some Images / Video Frames

2. **Train DL Model**
   - Run Trained DL Model on Entire Input Data to Generate Labels

3. **Use Trained DL Model**
   - Correct Labels on Some Data
   - Manually Correct Labels on Some Data

4. **Repeat Till Labels Achieve Desired Accuracy**
Computer vision at play

Safer Cities

Asset Management

Retail
Training vs Inference

Inferencing
- Single forward pass
- Some layers can be merged / pruned
- Weights can be quantized (FP32 → INT16/8/4/…)
- Network choice - speed vs accuracy
- Latency / Throughput
Deploying Trained Models

Data Center: Train model & Compile to Edge

PowerAI Vision

- Model Parser & Compiler
- Model Optimization (Layer Merge + Quantization)
- Output model + weights

Output model + weights

Map to Different Platforms

- Cloud or Edge
  - FPGAs, CPUs, GPUs
  - Xilinx Alveo U200
  - Embedded FPGA
  - Embedded GPU
  - Neural network processor

Trained DNN model

CPU + Accelerator
PowerAI Vision + Alveo U200
Model Choice - YOLO (You Only Look Once)

- Inference is fast (single network eval), and takes entire image into context
- Tiny-yolo is even faster, though accuracy may be compromised
- [https://youtu.be/VOC3huqHrss](https://youtu.be/VOC3huqHrss)

NOTE: PowerAI Vision supports multiple model types, but we focus here on YOLO
Import images and label objects
Train labeled model
Deploy Trained Model

Deploy Model

Deployed model name:
- rightsized_model

Deploy to accelerator:
- GPU
- GPU
- Xilinx FPGA - 8 bit
- Xilinx FPGA - 16 bit

Cancel  Deploy
Deploy Trained Model

- **Optimization:** Optimized for speed (tiny YOLO V2)
- **Accelerator type:** XILINX FPGA - 16 Bit
PowerAI Vision - Microservices on Kubernetes

> **Microservices:**
  > UI
  > Database
  > Auth
  > API

> **Training:**
  > Creates a new POD/Container
  > Allocates GPU
  > Trains
  > Exits

> **Inferencing:**
  > Creates a new POD/Container
  > Allocates Accelerator (GPU or FPGA)
  > Provides Inferencing API

```
# /opt/powerai-vision/bin/helm.sh status amartey
LAST DEPLOYED: Wed Aug 1 13:51:20 2018
NAMESPACE: default
STATUS: DEPLOYED

RESOURCES:
  ==> v1/Service
  NAME                      CLUSTER-IP EXTERNAL-IP PORT(S)     AGE
  powerai-vision-amartey-keycloak 10.10.0.151 <none>     8080/TCP,8443/TCP 58d
  powerai-vision-amartey-mongodb 10.10.0.93  <none>       27017/TCP    76d
  powerai-vision-amartey-portal   10.10.0.201 <none>       9080/TCP    76d
  powerai-vision-amartey-postgres 10.10.0.77  <none>       5432/TCP    76d
  powerai-vision-amartey-taskanaly 10.10.0.246 <none>       5000/TCP    76d
  powerai-vision-amartey-ui      10.10.0.18  <none>       80/TCP       76d

  ==> v1beta1/Deployment
  NAME                      DESIRED  CURRENT  UP-TO-DATE AVAILABLE  AGE
  powerai-vision-amartey-keycloak 1       1        1           1          58d
  powerai-vision-amartey-mongodb 1       1        1           1          76d
  powerai-vision-amartey-portal   1       1        1           1          76d
  powerai-vision-amartey-postgres 1       1        1           1          76d
  powerai-vision-amartey-taskanaly 1       1        1           1          76d
  powerai-vision-amartey-ui      1       1        1           1          76d

  ==> v1beta1/Ingress
  NAME                      HOSTS ADDRESS  PORTS AGE
  powerai-vision-amartey-ing  *      127.0.0.1 80    76d

  ==> v1/Secret
  NAME                      TYPE  DATA  AGE
  powerai-vision-amartey-secrets Opaque 6 76d

  ==> v1/ConfigMap
  NAME                      DATA  AGE
  powerai-vision-amartey-config 52 76d
```
Scheduling FPGAs on Kubernetes

> Device Plugin Framework
  >> Daemon on each node reports available FPGAs to master node
  >> Master node tracks allocation of FPGAs to pods that request it
  >> Master schedules pod to node
    - If there’s no FPGA available, pod waits in scheduling until one is
  >> Node’s daemon tells pod how to connect to FPGA

> Daemon
  >> Node’s daemon connects pod to required resources
    - PCI block devices (user and management PFs)
    - Mount paths (drivers)
    - Environment variables
  >> Deployed pod only given access to single allocated device
    - Does not need privileged access
Integration with Xilinx MLSuite

Train YOLO

Deployment

Schedule Available FPGA

Run Inference

Inference Container

{ REST API }

Deploy.py

GPU or FPGA?

Caffe Prototxt, Weights, & Anchors

Sample Training Images

 xfDNN Compile & Quantize (if needed)

 xfDNN Runtime

Sample Training Images

Sample Training Images
container Logs

Compile & Quantize

root : INFO INFO : unzipped the snapshot files to /opt/deploy

root : INFO INFO : root:

root : INFO INFO : --- Compiling ---

root : INFO INFO : WARNING: Logging before InitGoogleLogging() is written to STDERR


...
Come see us - TODO

PowerAI Vision

Accelerated Power Systems:
  ➢ https://www.ibm.com/it-infrastructure/power/accelerated-computing

PowerAI Enterprise