IBM Technology and Solutions for Artificial Intelligence and HPC

AI System Architecture

IBM Power9 and Beyond

Ulrich Walter
Ulrich.walter@de.ibm.com

Supercomputers Built for AI
The race is on…
Use cases for AI

**Automation**
- Process Automation
- Natural Language Creation
- Autonomous Driving
- Robotics

**Prediction**
- Predictive Analytics
- Prescriptive Analytics
- Recommendation Engines

**Security**
- Facial Recognition
- Fraud detection
- Pattern Recognition
- Compliance Check

**Discovery**
- Social Engineering and profiling
- NLP/Text Mining & Analysis
- Extraction Rules
AI DEMANDS A DIFFERENT TYPE OF SYSTEM

IBM Power Systems provides the cutting-edge advances in AI that data scientists demand, and the critical reliability that IT needs.

IBM has reimagined infrastructure for the journey to AI
And so it began

Mov SI, NumBytes
Mov CX, [SI]
Les DI, SegAdr1
Lds SI, SegAdr2
Mov AX, 0
Jne
Dec AX

RAM

Von Neumann Bottleneck
New workloads require different architectures

- **CPU = Scalar**
  - Designed for office apps
  - Has evolved for the web

- **GPU = Vector**
  - Designed for graphics
  - Has evolved for HPC

- **TPU2 = Tensor Processing Unit**
  - Accelerates neural network computations

- **IPU = Graph**
  - Designed for intelligence
  - The future of computing

- **IBM TrueNorth**
  - Neuromorphic CMOS integrated circuit

- **Quantum Computing**
New Kids on the Block
GPU powered AI, Analytics, Databases and Storage

Languages
- Go
- Python

Storage
- NYRIAD
- Excelero

Analytics Tool, Frameworks and Libraries
- Julia computing
- Pandas
- Iguazio
- NumPy
- H2O.ai
- Spark
- HORTONWORKS

GPU optimized databases
- MAPD
- Julia DB
- brytlyt
- SQREAM TECHNOLOGIES
- Kinetica

AI Frameworks
- PyTorch
- TensorFlow
- Caffe2
- Chainer

OpenPOWER
OpenCAPI
The OpenPower Ecosystem of Innovators Creates True Differentiation in Performance and Cost

IBM Power Systems LC Portfolio

With IBM warranty and enterprise support structure

Member-contributed Innovation

Ecosystem-Driven Customer Choice
The five phases of AI/DL - From data collection to action -

From data to information

From information to knowledge

From knowledge to action/reaction

Time to Collect and filter

Big Data, PB

Time to Analyse and prepare

Large Data, TB

Time to learn

Large Data, TB

Time to reason & conclude

Small Data, GB, MB
Challenges on storage for analytics and AI/DL data

Attributes and Considerations

<table>
<thead>
<tr>
<th>Volume</th>
<th>TB</th>
<th>PB</th>
<th>EB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity</td>
<td>$10^3$</td>
<td>$10^6$</td>
<td>$10^9$</td>
</tr>
<tr>
<td>xB/s IO/s</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TRUST

Data Phases: Collect, Prepare, Analyse, Learn, Inference

Intervall

IBM Systems
The five phases of DL/AI in the data Supply Chain – technical view

**Detect and Collect**
- Image&Video
- Text
- Voice&Sound
- Sensor
- ComInt, ELInt, SigInt

**Data preprocessing**
- Data selection
- Analyse
- Compress/Map Reduce
- Tag/Aggregate

**Learn**
- Distributed Deep Learning
  - Comparison and interpretation
  - Combine
  - Conclude/Reason
  - Control, loopback and update
  - Knowledge Base

**Inference**
- Applied Knowledge
  - Platforms
  - FPGA
  - Applications
  - Appliances

**Continuous Learning**

**Data Creation**
- Device type
- Latency
- Volume
- Velocity
- Intervall
- Interface
- Data source
- Data model
- Data Type
- API
- Originator
- Veracity

**Data preparation**
- Storage
- Bandwidth
- Latency
- Capacity
- I/O Blocksize
- Filesystem
- Storage Type

**Transformation and validation**
- Verified data
- Volume
- Storage Type
- Filesystem

**Compute (Training & Inference)**
- Bandwidth
- Storage
- RAM
- CPU
- GPU
- Compute Performance
- Latency
- I/O throughput
- Data Locality
- Memory size

**IBM Power 9 AC922**

**IBM Spectrum Family and other storage options**
- Flash + NVMe
- Disk
- Object Store
- Storage Rich Servers
- Tape & LTFS

**Shared PCIe-4 For IB**
- NVME Flash
- High I/O Storage
- Data locality and low latency

**CAPI Coherent Accelerator Processor Interface for Data**
- CPU-GPU
- NVLINK 2.0
- 150 GB

IBM Systems
IBM AI Reference Architecture from Experimentation to Expansion

**Experimentation**
Single Tenant

- Data Scientist's workstations
- IBM Power Systems AC922
- Internal SAS drives & NVM's

**Stabilization & Production**
Secure Multitenant

- Training Cluster IBM Power Systems AC922
- Existing Organization Infrastructure
- High-Speed Network Subsystem
- IBM Elastic Storage Server (ESS)

**Expansion**
Enterprise Scale / Multiple Lines of Business

- Master & Failover Master Nodes IBM Power Systems LC921 & LC922
- Login Nodes IBM Power Systems LC921 & LC922
- High-Speed Network Subsystem

**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
Performance = important, but bandwidth matters

76% of users said there is still a moderate to large amount of parallelism to extract in their code, but...

85% of respondents rated the difficulty level of expressing that parallelism as moderate or difficult – underlining the need for more support, training, and assistance in this area.

For the first time, FLOPS was NOT the #1 ranked requirement.

Local memory capacity was not a driver for most users, perhaps in recognition of cost trends.

Data sharing, long-term archival storage of 100s of PB, data analytics, and faster WAN connections also showed up as important requirements.

IBM AC922

An Acceleration Superhighway
Designed for the AI Era
Delivering Enterprise-Class AI

The Best Server for Enterprise AI
POWER9 Acceleration Server (AC922)

- 2U server - 19” Rack enclosure
- 2 POWER9 Scale Out SMT4 Processors (Up to 22-cores per socket)
- 4 or 6 NVidia Tesla GPU processors (**NVLink 2.0 attached**)
  - SXM2 Slot Form Factor
  - 300W V100 GPU  16/32 GB NVRAM
- Up to 1TB Total DDR4 Industry Standard memory RDIMMs
  - 16 IS RDIMM slots
  - 341 GB/s total system memory peak bandwidth
- 4 PCIe Gen4 Low Profile slots
- 2 SFF (2.5”) SATA bays (HDD or SSD)
- 2 Ethernet 1Gb ports
- Air and Water cooled versions
- Linux (RHEL) only platform with BMC service processor
Resolve the PCI-E bottleneck for your code with POWER9 and NVLink 2.0:
Transfer data **5.6X faster than the CUDA Host-Device Bandwidth of tested x86 platforms**

**THE** Platform for developers requiring CPU-GPU Bandwidth

POWER9 is the **only** processor with NVLink 2.0 from CPU to GPU

– Delivering **5.6X Host-Device bandwidth vs Xeon E5-2640 v4** based systems with CUDA H2D Bandwidth Test
– No code changes are required to leverage NVLink capability
  • Application performance could be further increased with application code optimization

- Results are based on IBM Internal Measurements running the CUDA H2D/D2H Bandwidth Test
- Hardware: Power AC922; 3240 cores (2 x 20c chips), POWER9 with NVLink 2.0: 2.25 GHz, 1024 GB memory, 4xTesla V100 GPU; RHEL 7.4 for Power LE (POWER9); S822LC for HPC; 20 cores (2 x 10c chips), POWER8 with NVLink 1.0: 2.86 GHz, 1024 GB memory, Tesla P100 GPU, RHEL 7.3
- Competitive HW: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 512 GB memory, 4xTesla P100 GPU, Ubuntu 16.04

© 2017 IBM Corporation
NVLink Evolution in POWER HPC

P100 GPU with NVLink 1.0
- NVLink between CPUs and GPUs enables fast memory access to large data sets in system memory
- Two NVLink connections between each GPU and CPU-GPU leads to faster data exchange
- Time saving +300% vs x86
- Add. CAPI feature for fast IO to storage and network

Power8 Chip with NVLink

IBM POWER9 bandwidth optimized
V100 GPU with NVLink 2.0
- +87% Bandwidth Vs NVLINK 1.0
- 468% vs x86 PCIe3
- Tensor Cores
- V100 Technology

IBM POWER9 scalability optimized
- +20% Bandwidth
- Tensor Cores
- V100 Technology
- 3 GPUs/CPU
- 33% more GPU RAM
- + watercooled

x86
- NVLink only between GPUs
- Long lasting ramp-up times due to PCIe Bottleneck
- Reduced efficiency

Pcie x16
The NVLINK 2.0 effect

Train Large AI Models
~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
Power AC922 w/ 4x V100 GPUs

GoogleNet model on Enlarged ImageNet Dataset (2240x2240)
NVLink and P100 Advantage:
reducing communication time, incorporating the fastest GPU for deep learning

- NVLink reduces communication time and overhead
- Data gets from GPU-GPU, Memory-GPU faster, for shorter training times

Power advantage: data communication and GPU performance

POWER8 + Tesla P100+NVLink

x86 based GPU system

ImageNet / Alexnet: Minibatch size = 128
Industry initiatives put SCM where it’s easy
- Easier from hardware development perspective
- Very limiting for end users
- Allows interfaces to be used as control point

SCM accessed via NVMe:
- Latency advantage reduced by software stack
- Bandwidth limited by PCIe infrastructure
- IOPs limited by CPUs consumed to run NVMe stack

SCM accessed over DDR memory buses:
- Load/Store access model improves over NVMe
  - Eliminates CPU cycles spent on NVMe stack
  - Reduces latency by removing SW pathlength
- Creates DRAM vs SCM configuration tradeoffs
  - Capacity and bandwidth spread across 2 tiers
  - Both tiers are sub-optimized
Auto Hyper-Parameter Tuning

Hyper-parameters
- Learning rate
- Decay rate
- Batch size
- Optimizer
  - Gradient Decedent, Adadelta, Momentum, RMSProp ...
- Momentum (for some optimizers)
- LSTM hidden unit size

Random

Tree-based Parzen Estimator (TPE)

Bayesian

Spark search jobs are generated dynamically and executed in parallel

Multi-tenant Spark Cluster
IBM Spectrum Conductor with Spark
Snap ML
Distributed GPU-Accelerated Machine Learning Library

Snap Machine Learning (ML) Library

- Logistic Regression
- Linear Regression
- Support Vector Machines (SVM)
- More Coming Soon
- libGLM (C++ / CUDA Optimized Primitive Lib)
- Distributed Hyper-Parameter Optimization
- Distributed Training

APIs for Popular ML Frameworks

(coming soon)
Resources

Power AI Developer Portal

OpenPower on Youtube
https://www.youtube.com/channel/UCYLtbUp0AH0ZAv5mNut1Kcg

TSRG Web form Link For IBMers:
http://w3.ibm.com/support/techline

TSRG Web form Link for Business Partners  Partners:
http://www.ibm.com/partnerworld/techline

Power AI Vision