Accelerated Hyperscale Compute for AI at the Edge
S9756

—

Brandon Jones
Systems Architect
b.jones@ibm.com

March 20, 2019 / © 2019 IBM Corporation

Introductions

Who am I?

Who are you?

Why should you care?

GOAL

- To gain insight into the changing landscape of AI for the Edge inside the Telecommunications Industry.
- How are we going to get there?
- Finally, why we will need more than traditional compute to accomplish this.

Definitions

What is 5G?

What is Edge?

5G value chain will invest an average of $200B per year to expand 5G capabilities

IHS research estimates that from 2020-2030, 5G value chain CAPEX spending will contribute **$2.4T** to the world economy.

- 5G value chain will invest an average of $200B annually to continually expand their 5G capabilities within network and business application infrastructure.

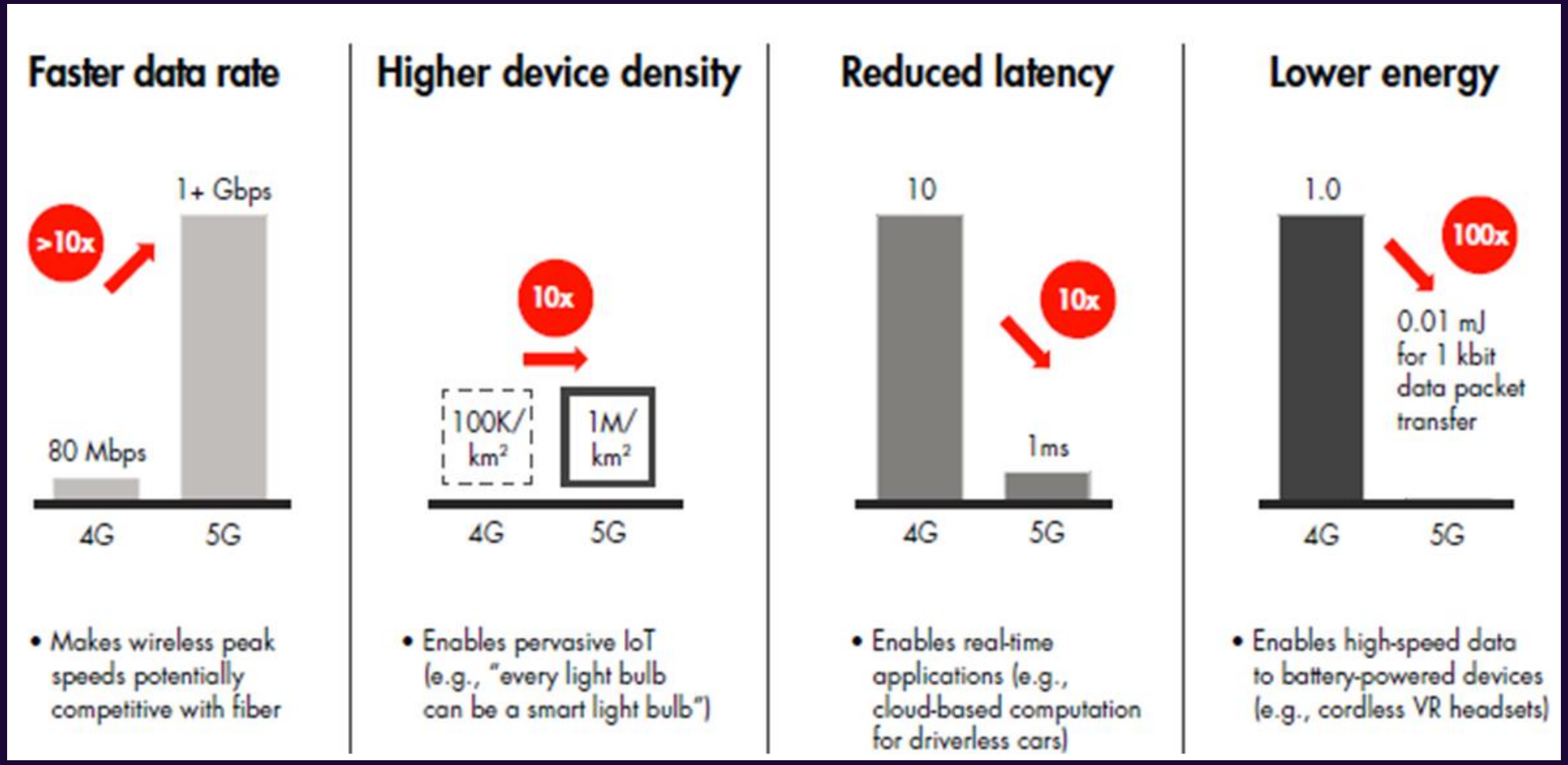
From 2020–2035, 5G will grow WW GDP equivalent to an economy size of India (the 7th largest economy in the world).

Examples of current services that will increase as 5G is adopted:

- Asset Tracking
- Smart Agriculture
- Smart Cities
- Energy/utility Monitoring
- Smart Homes
- Beacon & Connected Shoppers
- Remote monitoring
- Physical infrastructure

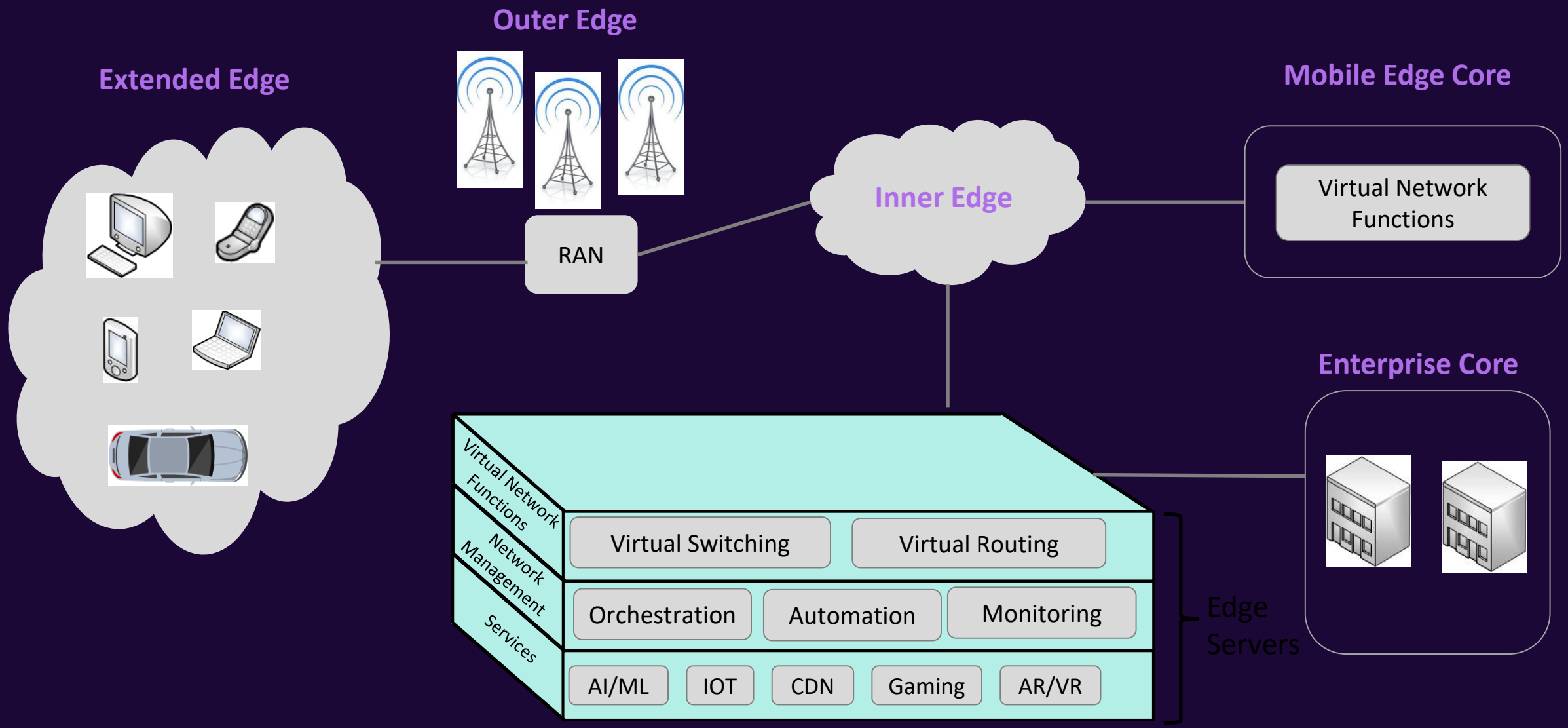
Source: IHS, The 5G economy: How 5G technology will contribute to the global economy, Qualcom: https://www.qualcomm.com/media/documents/files/ihs-5g-economic-impact-study.pdf
IHS, Connectivity technologies An in-depth view into the competition, applications and influencers driving the foundation of IoT, June 2018, https://cdn.ihs.com/www/pdf/IHS%20Markit%20-%20Connectivity%20Technology%20Competition%20Drivers%20and%20Influencers.pdf?utm_campaign=PC10706-1_JM_eT1_TMT_GLOBAL_TMT_IoT%20Ecosystem%20-%20Innovation%20-%20Client&utm_medium=email&utm_source=Eloqua

What 5G Means



Use Cases for AI in Telco

- Log Analysis (SIEM tooling and Security)
  Event analysis and correlation combined with predictive modeling for future outcomes.
- Real-Time (or Near Real-Time) Network Analytics
  Demand, Usage, Capacity, Forecasting, Performance Management Optimization
- Customer Experience – NPS –Customer Satisfaction
- Customer Care
  Customer Satisfaction – Churn – 360 customer view
- Marketing
  Targeted campaigns, success rate, strategy
- IOT & Connected Vehicle Analytics
- Environmental Monitoring and Analytics
- Fraud Analytics
  Network - CDR - Billing - Web – Retail
- Visual Asset and Equipment Inspection
  Towers and Antenna including predictive or ongoing maintenance
- Security
  Physical and or asset security and safety
- Marketing
  Observe in store traffic flows and product placement for analysis recommendations



My current view of what Edge will look like

Why do we even need acceleration?

Traditional Throughput Limitations

Examples:

Can We Achieve Line Rate Speed with only CPU with DPDK?

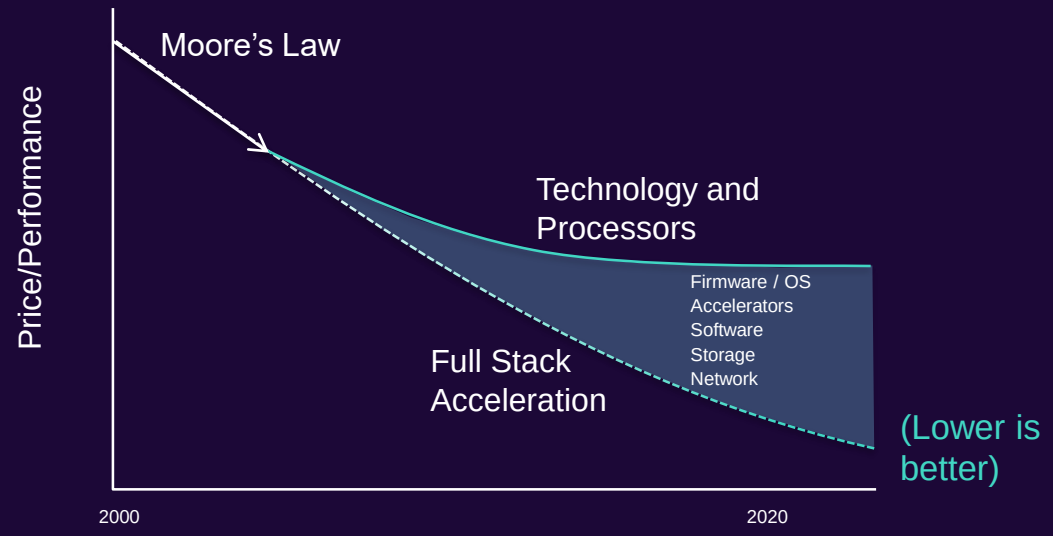
Latency concerns

Planar Throughput

Connectivity and Data Movement

Fundamental forces are accelerating change in our industry

IT innovation can no longer come from just the processor



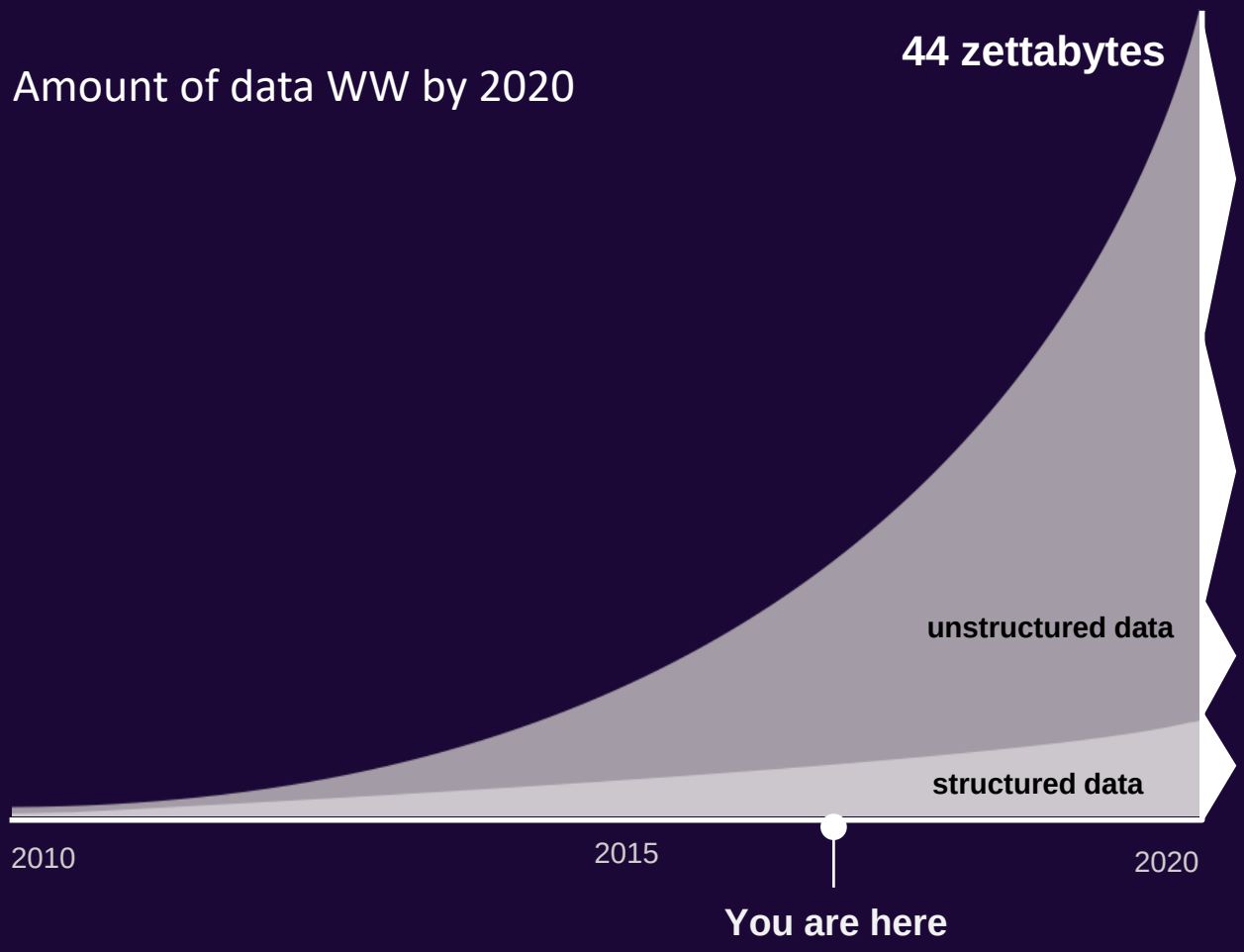
Full system stack innovation required

IT consumption models are expanding

- Cognitive
- Custom Hyperscale Data Centers
- Hybrid Cloud
- Open Solutions

Not only is Moore's Law "coming to an end in practical term, in that chip speeds can be expected to stall, but it is actually likely to roll back in terms of performance …" – William Holt, Intel Executive Vice President and General Manager

Data holds competitive value



Internet Of Things

Images & Multimedia

Text

Enterprise Data

Mobile

Medical

Oil/Gas

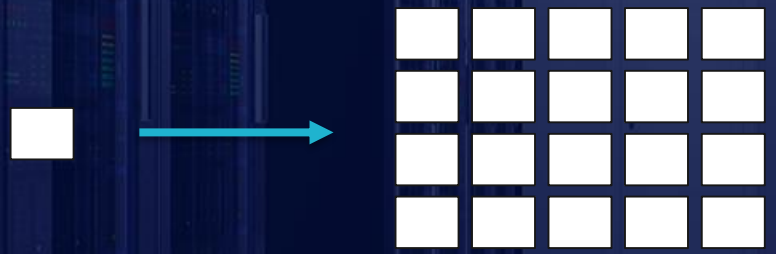
Weather

Homogenous was yesterday's approach

The AI era requires a new one

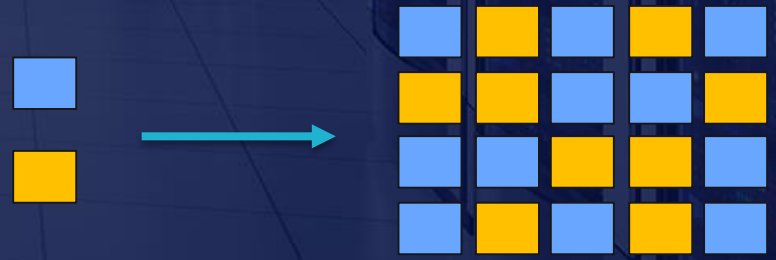
Legacy Approach

ONE SIZE FITS ALL - Approach all application requirements with a single non-optimized building block

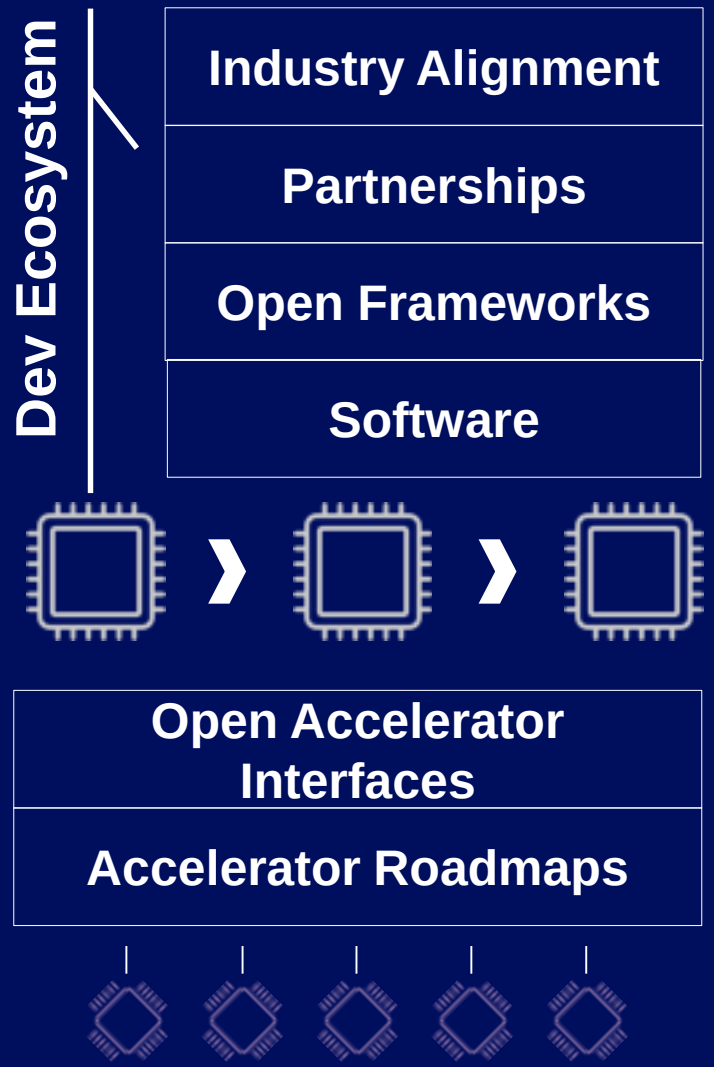


Modern Approach

Leverage optimized servers designed for the AI era and the vastly different requirements



Evolving from Compute Systems to **Cognitive Systems**



Not Just About Hardware Design

It's about co-optimization

which ***just works*** for ML, DL, and AI

IBM POWER SYSTEMS

AC922

An Acceleration Superhighway
Unleash state of the art IO and accelerated computing potential in the post “CPU-only” era

Designed for the AI Era
Architected for the modern analytics and AI workloads that fuel insights

Delivering Enterprise-Class AI
Flatten the time to AI value curve by accelerating the journey to build, train, and infer deep neural networks

Seamless CPU and Accelerator Interaction

coherent memory sharing

enhanced virtual address translation

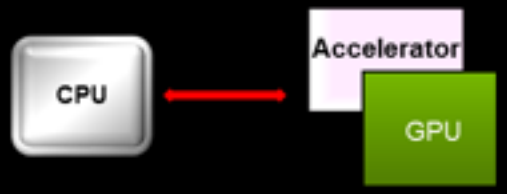
POWER9

Broader Application of Heterogeneous Compute

designed for efficient programming models

accelerate complex AI & analytic apps

Others



PCIe Gen3

extreme CPU and Accelerator bandwidth

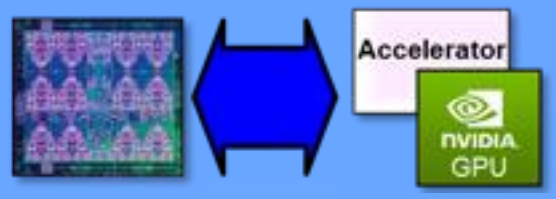
2x

PCIe Gen4

5x

POWER8
with NVLink 1.0

7-10x



POWER9 with 25G Link + NVLink 2.0

“SUMMIT” on POWER vs.“TITAN” on x86 >

5-10x FASTER vs. previous x86 system

75% LESS NODES for superior density

~29x PER NODE PERFORMANCE (>40TF)

~8x MORE STORAGE (250PB@2.5TB/s)

16x MORE MEMORY per node

"Summit, like Titan, will open a door to new ways to simulate and explore complex systems in the natural world. Our scientific community will see decreased time to solution, along with the ability to increase the complexity of their computational models, improving the simulation fidelity of a wide variety of important phenomena that are beyond the range of conventional experimental investigations."

— James J. Hack, Oak Ridge Leadership Computing Facility

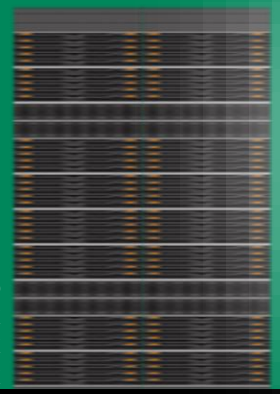
Compute Rack
18 Servers/rack
779 TF/s/rack
10.8 TB/rack
55 KW max

POWER9 2 Socket Server
Standard 2U 19" Rack Mount Chassis
2 P9 + 4/6 Volta GPU (@7 TF/s)
512 GB SMP Memory (32 GB DDR4 RDIMMs)
64/96GB GPU Memory (HBM stacks)

Mellanox TECHNOLOGIES
Scalable Active Network
IB4X EDR Switch

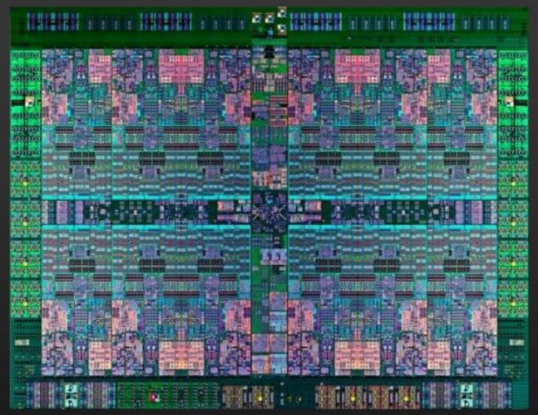
GSS-26
3 2U servers/rack
9 4U JBODs/rack
9 KW max/rack

ESS Building Block



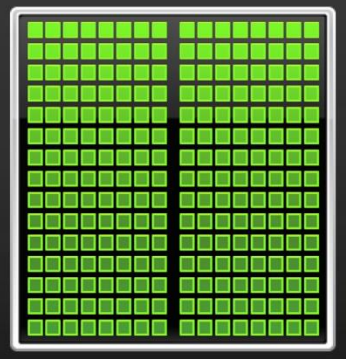
CORAL
Oak Ridge National Laboratory · Lawrence Livermore National Laboratory · Argonne National Laboratory
IBM · United States Department of Energy · NVIDIA

POWER9
22 cores
4 Threads/core, 0.65 DP TF/s



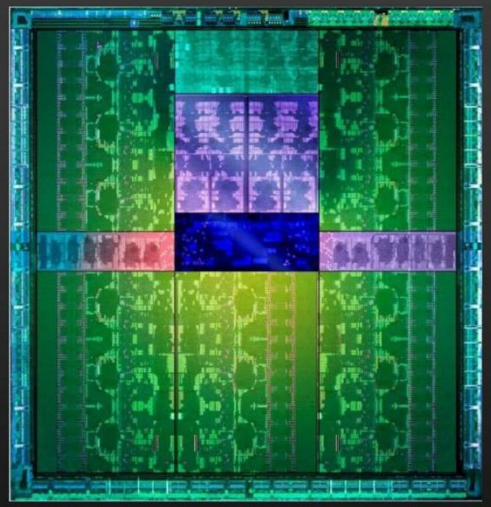
IBM POWER CPU
Most Powerful Serial Processor

100-150 GB/s



NVIDIA NVLink
Fastest CPU-GPU Interconnect

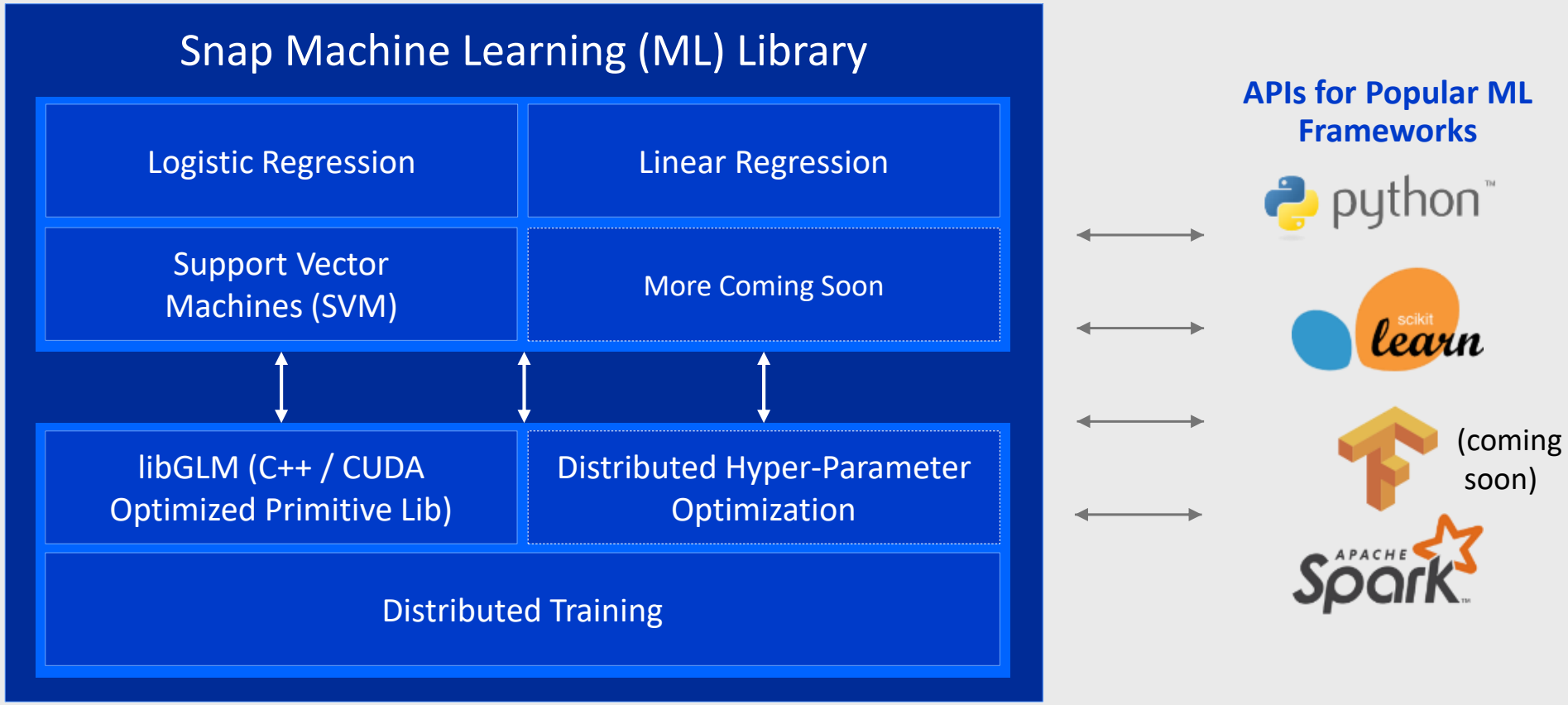
NVIDIA Volta 7 DP TF/s, 16GB @ 1.2TB/s



NVIDIA Volta GPU
Most Powerful Parallel Processor

Snap ML

Distributed GPU-Accelerated Machine Learning Library



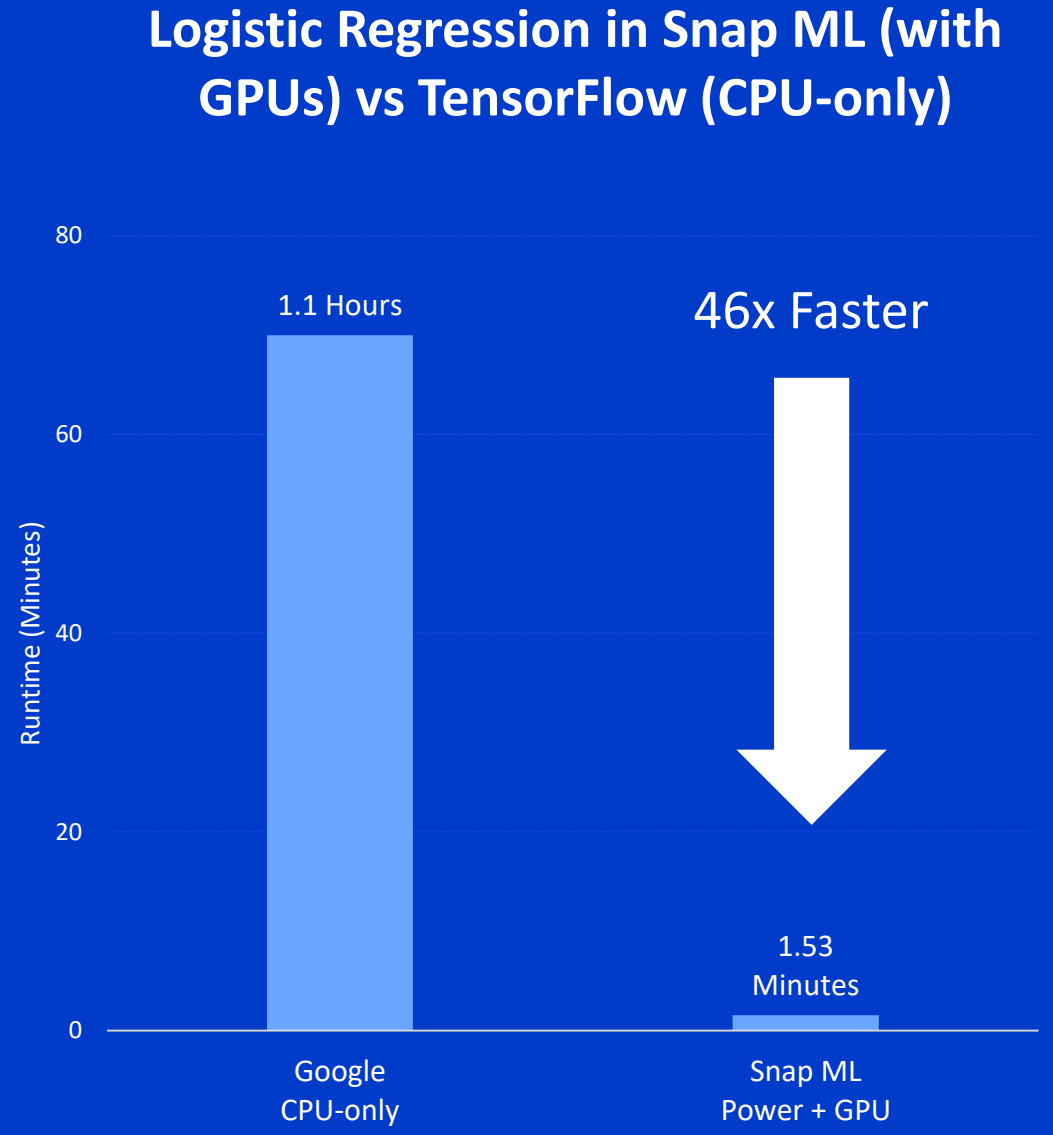
Snap ML: Training Time Goes From An Hour to Minutes

46x faster than previous record set by Google

Workload: Click-through rate prediction for advertising

Logistic Regression Classifier in Snap ML using GPUs vs TensorFlow using CPU-only

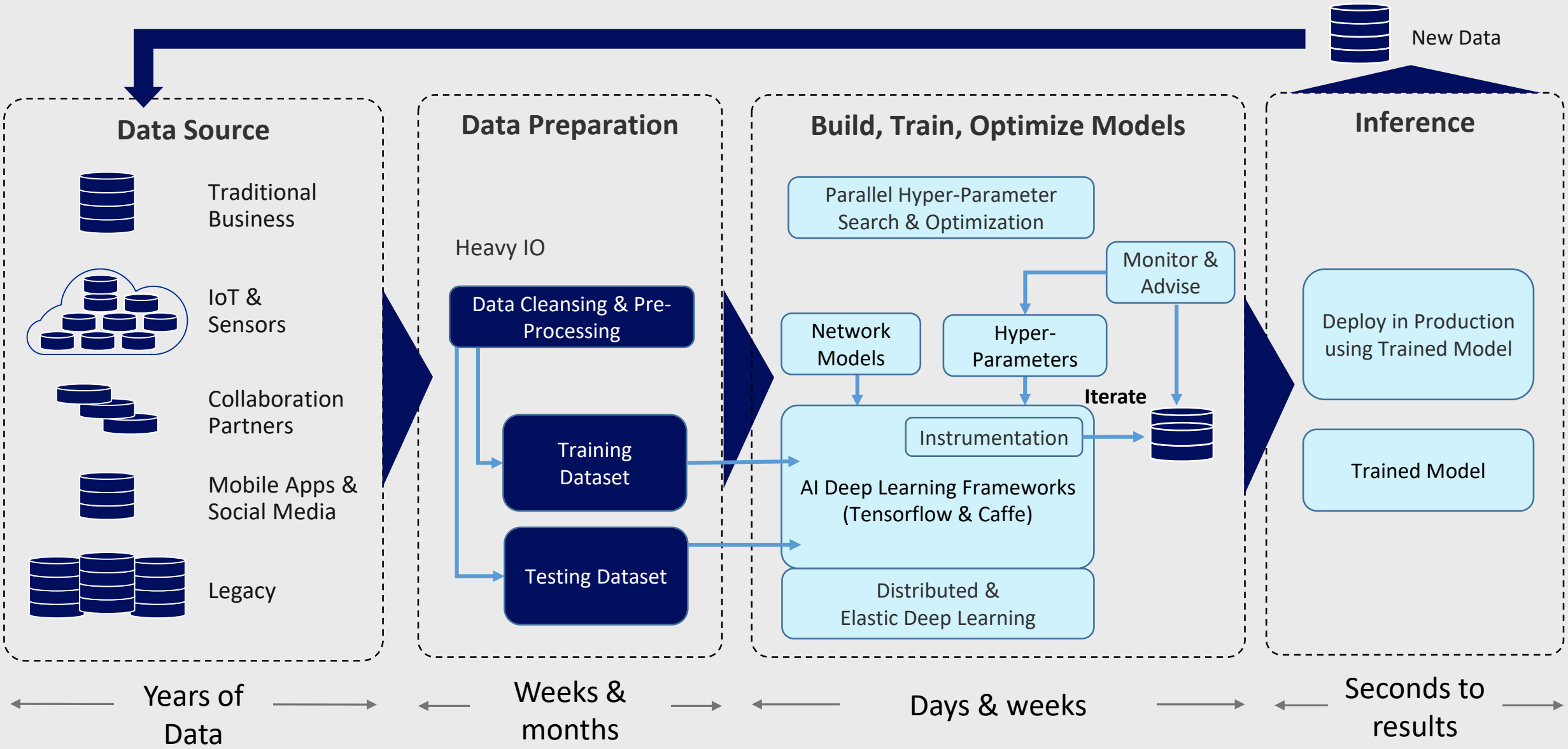
Dataset: Criteo Terabyte Click Logs (http://labs.criteo.com/2013/12/download-terabyte-click-logs/)
4 billion training examples, 1 million features
Model: Logistic Regression: TensorFlow vs Snap ML
Test LogLoss: 0.1293 (Google using Tensorflow), 0.1292 (Snap ML)
Platform: 89 CPU-only machines in Google using Tensorflow versus 4 AC922 servers (each 2 Power9 CPUs + 4 V100 GPUs) for Snap ML
Google data from this Google blog



Modeling will also need Inferencing in the Edge

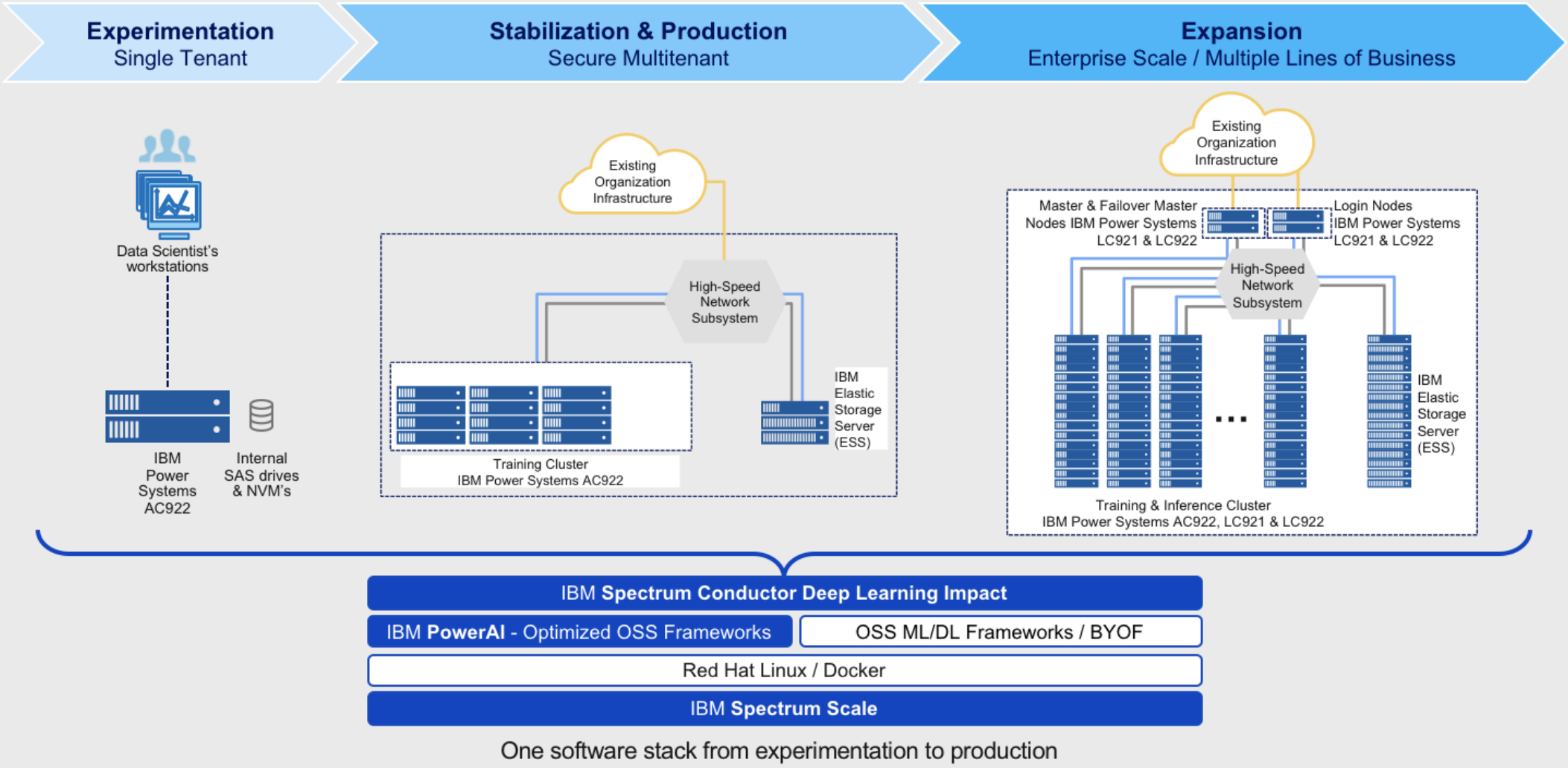
https://www.nvidia.com/en-us/data-center/tesla-t4/?ncid=pr-int-wnnc-58352

Work flow and data flow is complex



IBM AI Reference Infrastructure

*can be adapted to usage needs



https://www.ibm.com/downloads/cas/W1JQBNJV

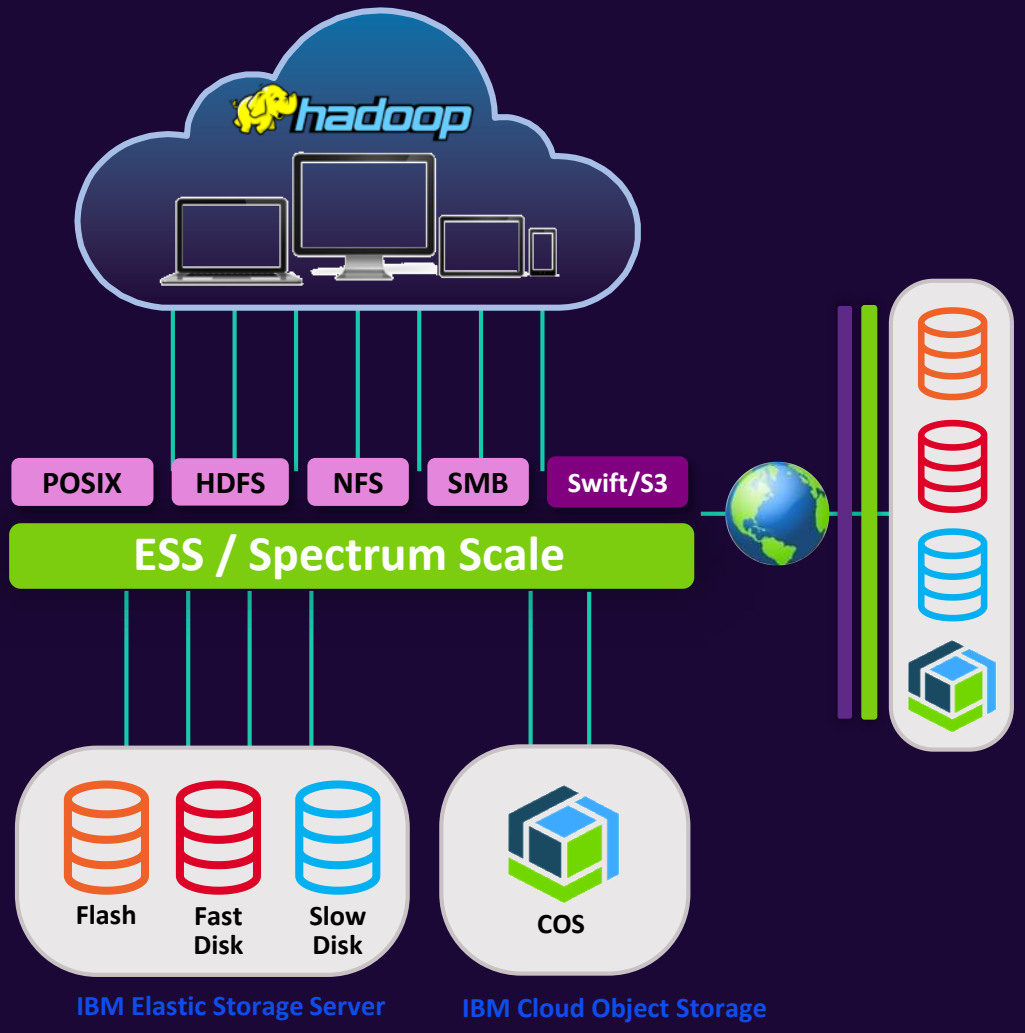
High-Performance Storage

IBM Spectrum Scale on ESS

Outperform
- Faster performance (40GB/s)
- More efficient - ~20% vs 200% overhead
- 92.5GB/s reads & 51.5GB/s write in CDO Benchmarks

Enterprise-Grade
- POSIX compliant
- Enterprise security, replication, reliability, etc.
- Scales to exabytes

Flexible
- Multi-protocol - HDFS, NFS, SMB, S3, Swift, and iSCSI



© IBM Corporation 2018

Spectrum Storage for AI with NVIDIA DGX for deep learning

IBM Storage & NVIDIA DGX Partnership

Announced Last Fall 2018

https://www.ibm.com/blogs/systems/introducing-spectrumai-with-nvidia-dgx/

https://blogs.nvidia.com/blog/2018/12/10/ibm-nvidia-ai-infrastructure/

Storage Reference Architecture

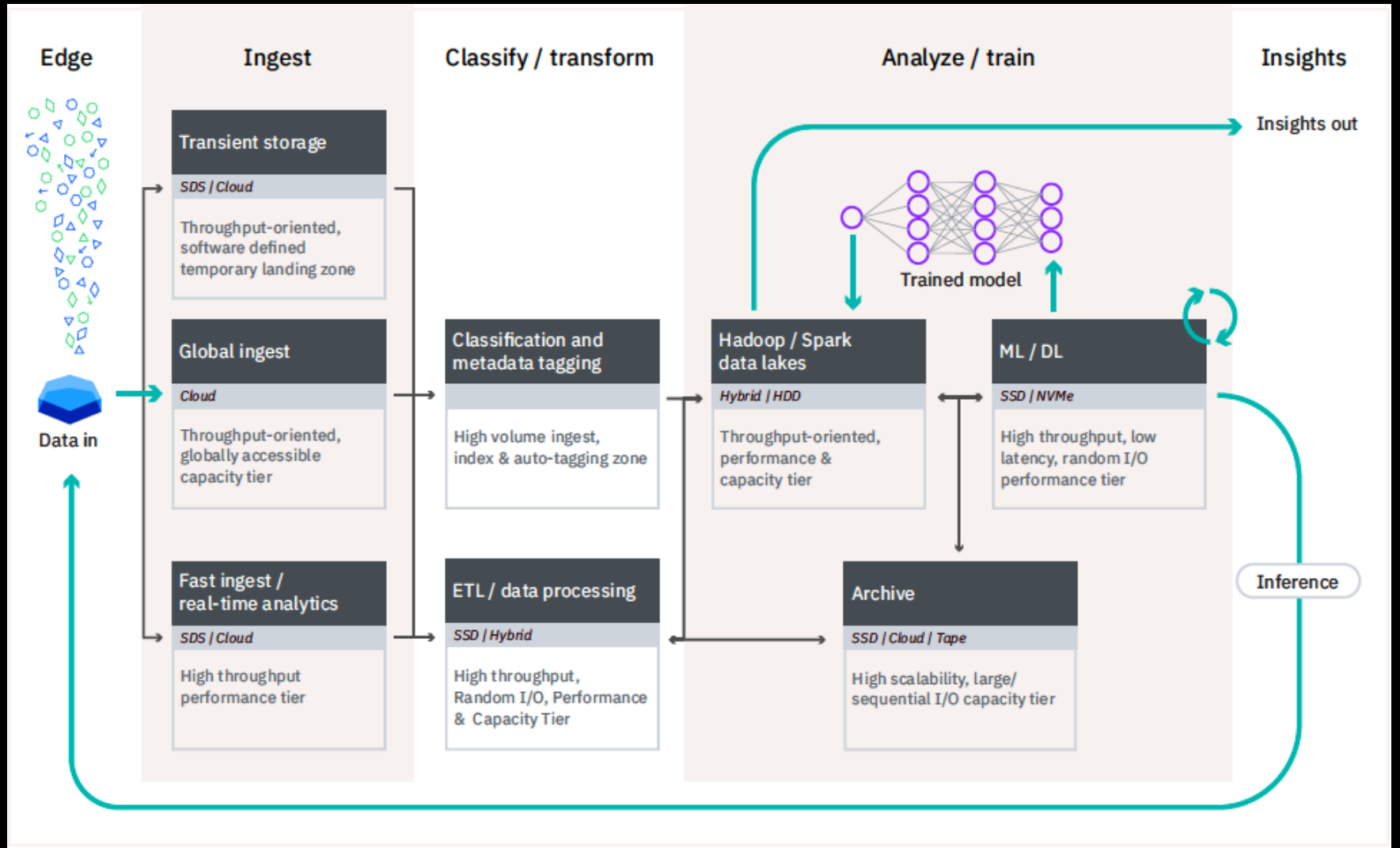
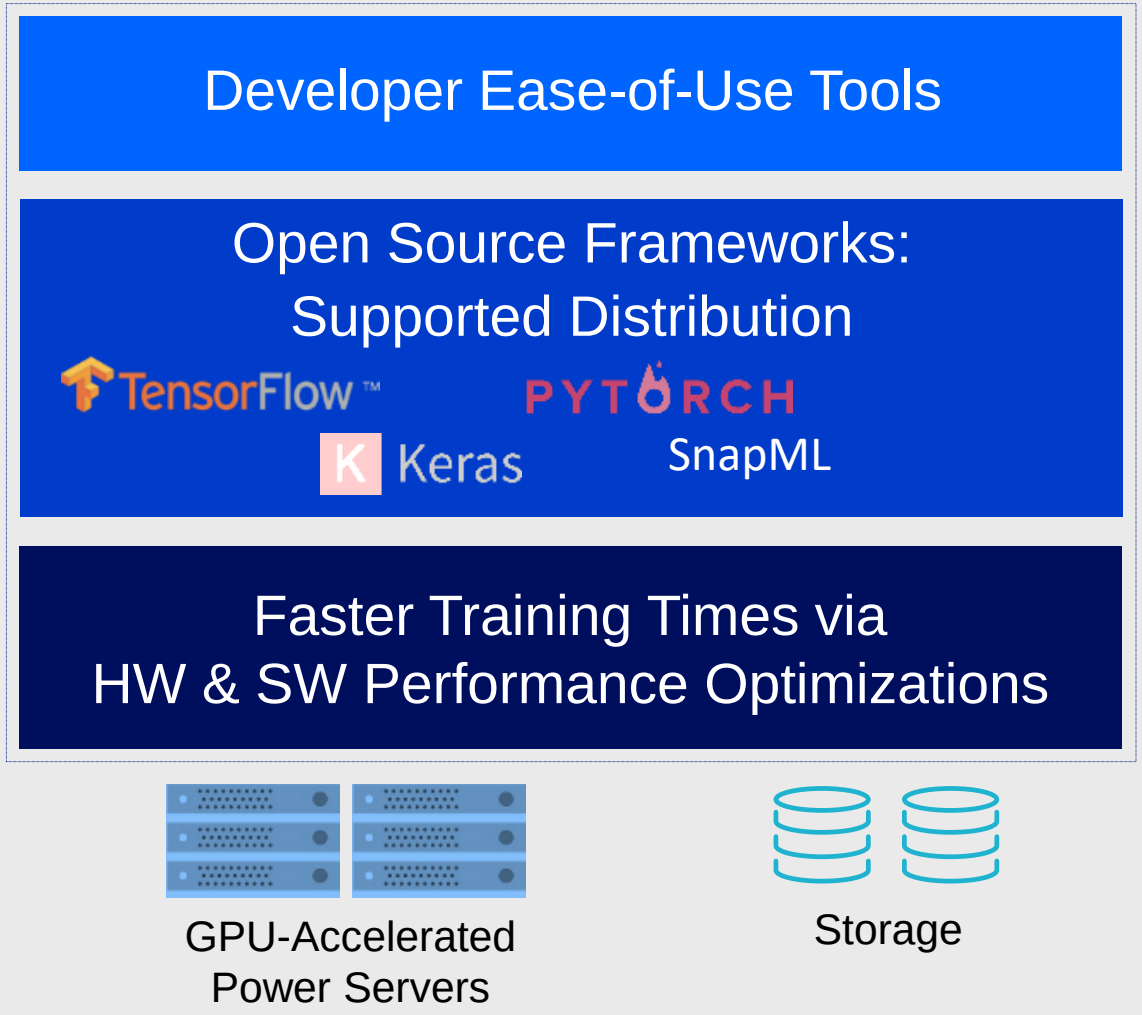


Figure 1: AI data pipeline with storage requirements

https://www.ibm.com/downloads/cas/MNEQGQVP

PowerAI: Open-Source Based Enterprise AI Offering



Offering

- Integrated & Supported Hardware-Software AI offering, with distribution of open-source AI software (TensorFlow, PyTorch, Keras ...)
- Open-source Enhanced for Ease of Use & Faster AI Training Times
- 3-4x Faster Training on Power-GPU Servers

Customers

- Enterprise, Academia, Autonomous Vehicle companies, Emerging Startups
- Focus on Production vs Experimentation: notion of SCALE as a key challenge
- Simplified approach: pull the drudgery out of developing AI

PowerAI Vision

Auto-ML for Images & Video

Label | Train | Deploy

PowerAI

PowerAI: Open Source ML Frameworks

TensorFlow™ PYTORCH Chainer SnapML

Large Model Support (LMS)

PowerAI Enterprise

Distributed Deep Learning (DDL)

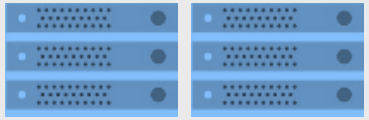
Auto ML (future)

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

Deep Learning Impact (DLI) Module

Data & Model Management, ETL, Visualize, Advise

Accelerated Infrastructure



Accelerated Servers

Storage

Summary and Conclusions

Special Note!

Using Tensor Swapping and NVLink to Overcome GPU Memory Limits with TensorFlow by Sam Matzek

S9426 Talk

- Wednesday, 3/20/19 | 16:00 - 16:50 - SJCC Room 210E (Concourse Level)

Thank You!

Backup

Distributed Deep Learning (DDL)

Deep learning training takes days to weeks

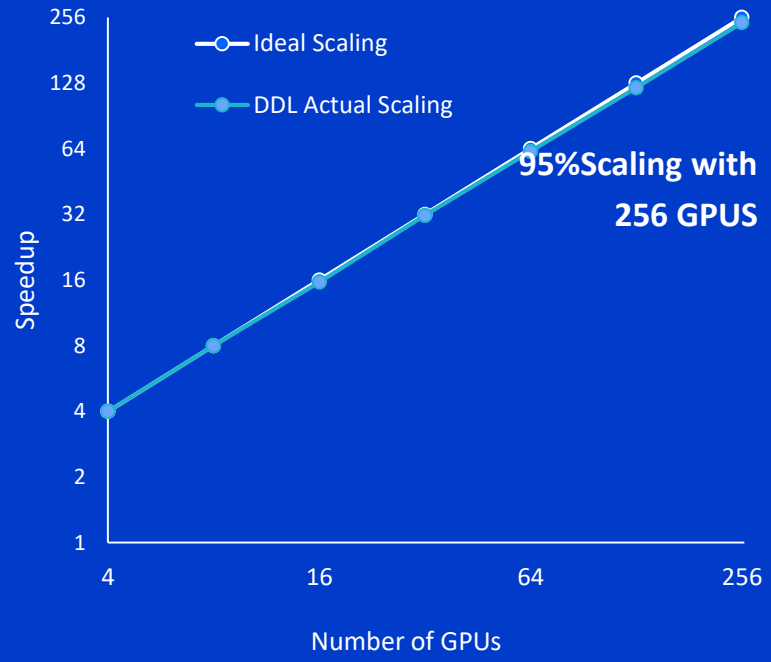
Limited scaling to multiple x86 servers

PowerAI with DDL enables scaling to 100s of servers

16 Days Down to 7 Hours
58x Faster

ResNet-101, ImageNet-22K

Near Ideal Scaling to 256 GPUs

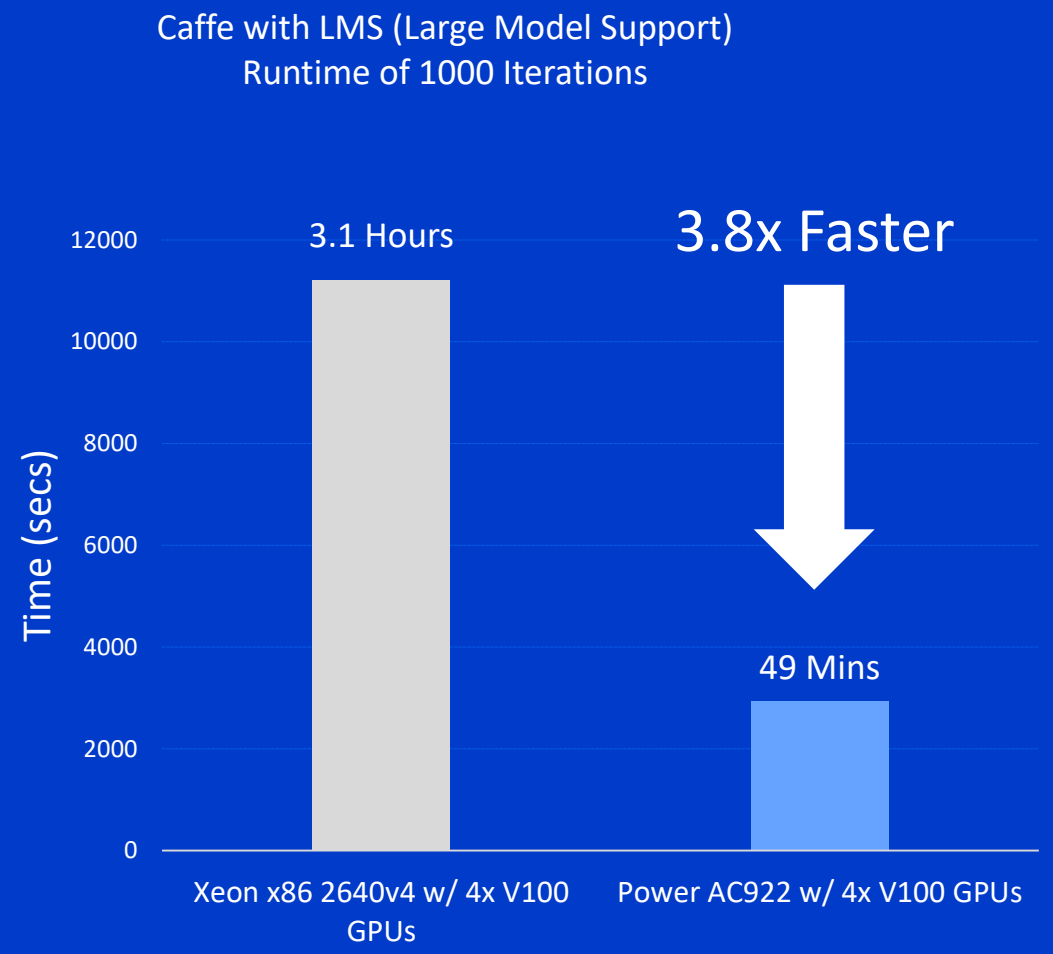


ResNet-50, ImageNet-1K

Caffe with PowerAI DDL, Running on Minsky (S822Lc) Power System

Large AI Models Train ~4 Times Faster

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



GoogleNet model on Enlarged ImageNet Dataset (2240x2240)

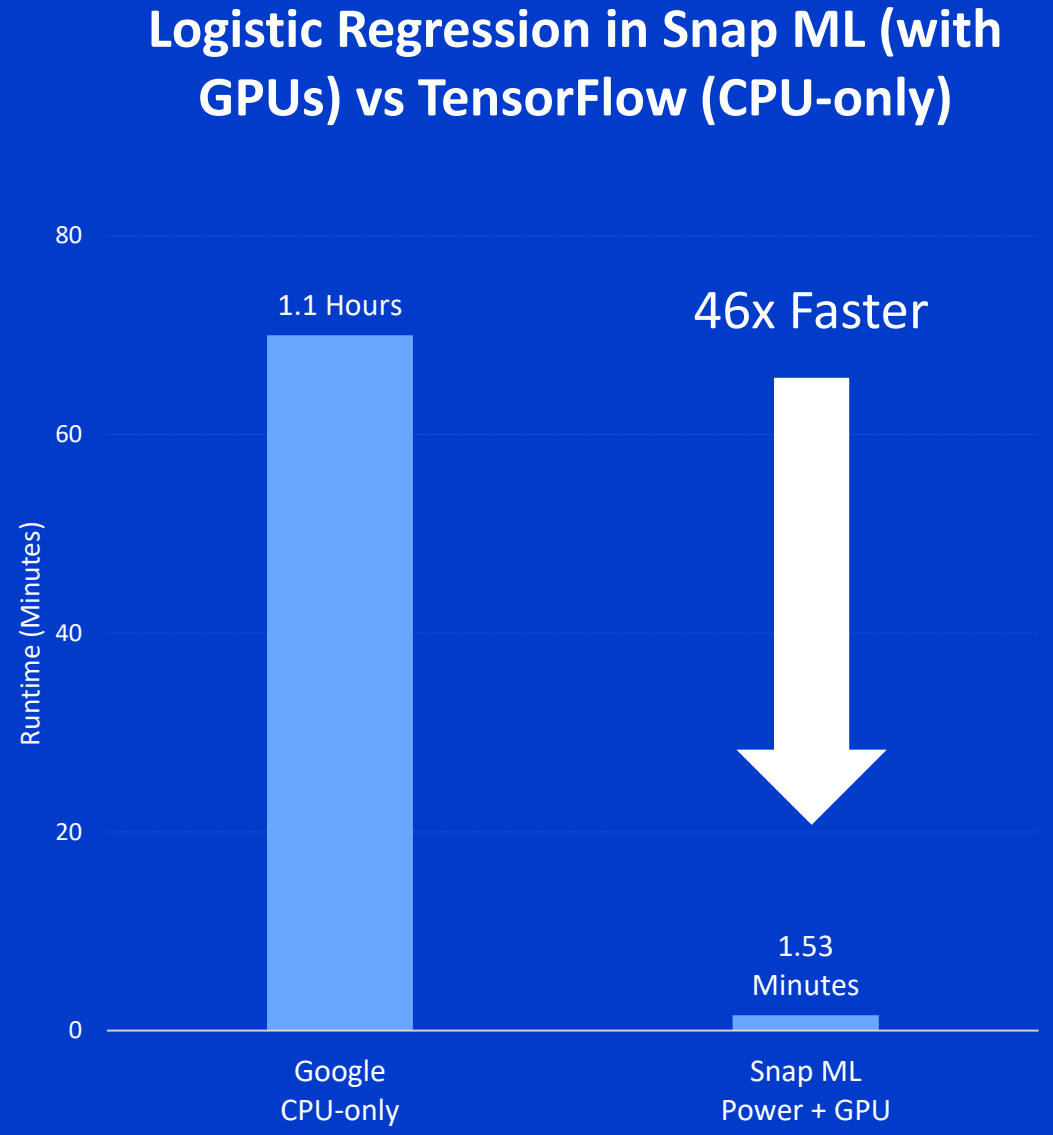
Snap ML: Training Time Goes From An Hour to Minutes

46x faster than previous record set by Google

Workload: Click-through rate prediction for advertising

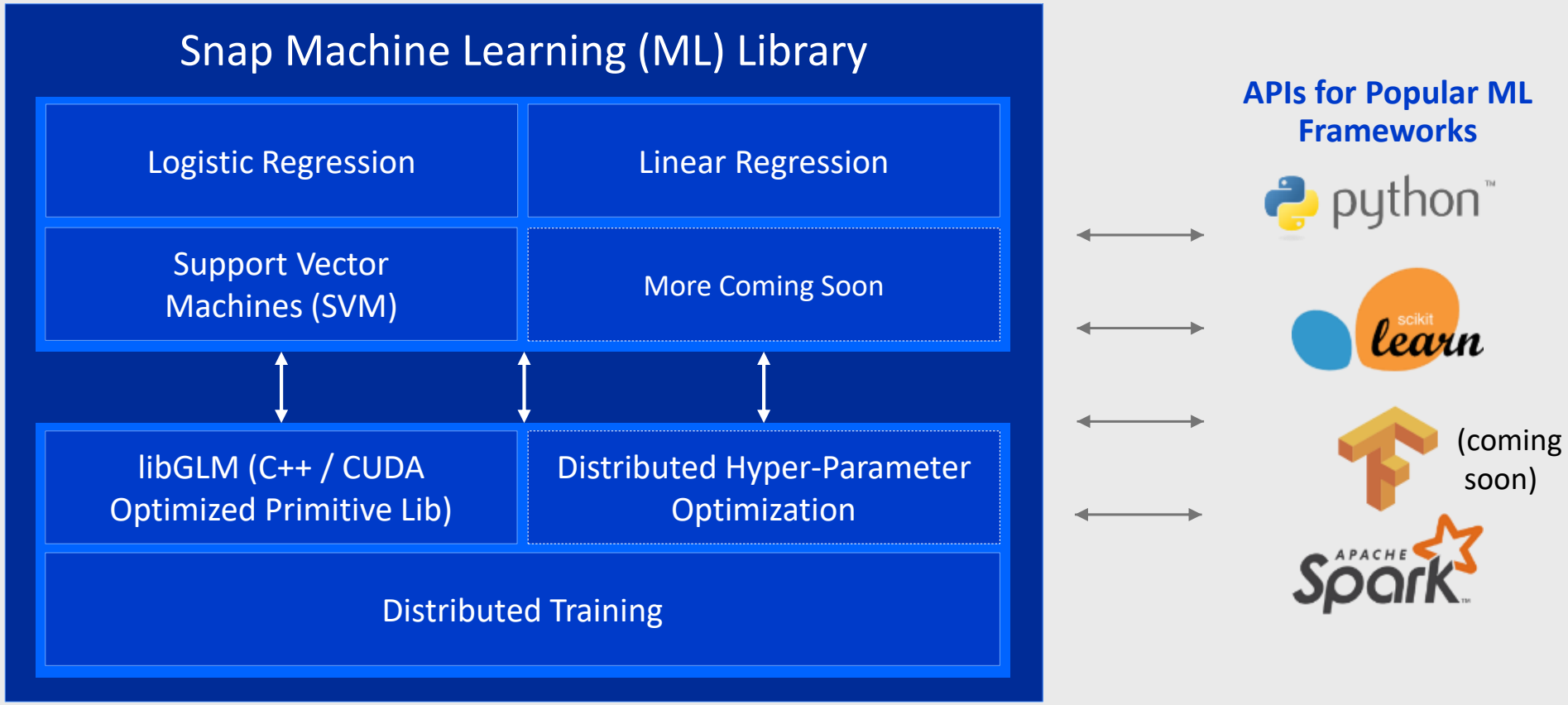
Logistic Regression Classifier in Snap ML using GPUs vs TensorFlow using CPU-only

Dataset: Criteo Terabyte Click Logs (http://labs.criteo.com/2013/12/download-terabyte-click-logs/)
4 billion training examples, 1 million features
Model: Logistic Regression: TensorFlow vs Snap ML
Test LogLoss: 0.1293 (Google using Tensorflow), 0.1292 (Snap ML)
Platform: 89 CPU-only machines in Google using Tensorflow versus 4 AC922 servers (each 2 Power9 CPUs + 4 V100 GPUs) for Snap ML
Google data from this Google blog



Snap ML

Distributed GPU-Accelerated Machine Learning Library



POWER9 processor

Others

PCIe Gen3

POWER9

2x faster

PCIe Gen4

State of the Art I/O and Acceleration Attachment Signaling

PCIe Gen 4 x 48 lanes
192 GB/s duplex bandwidth