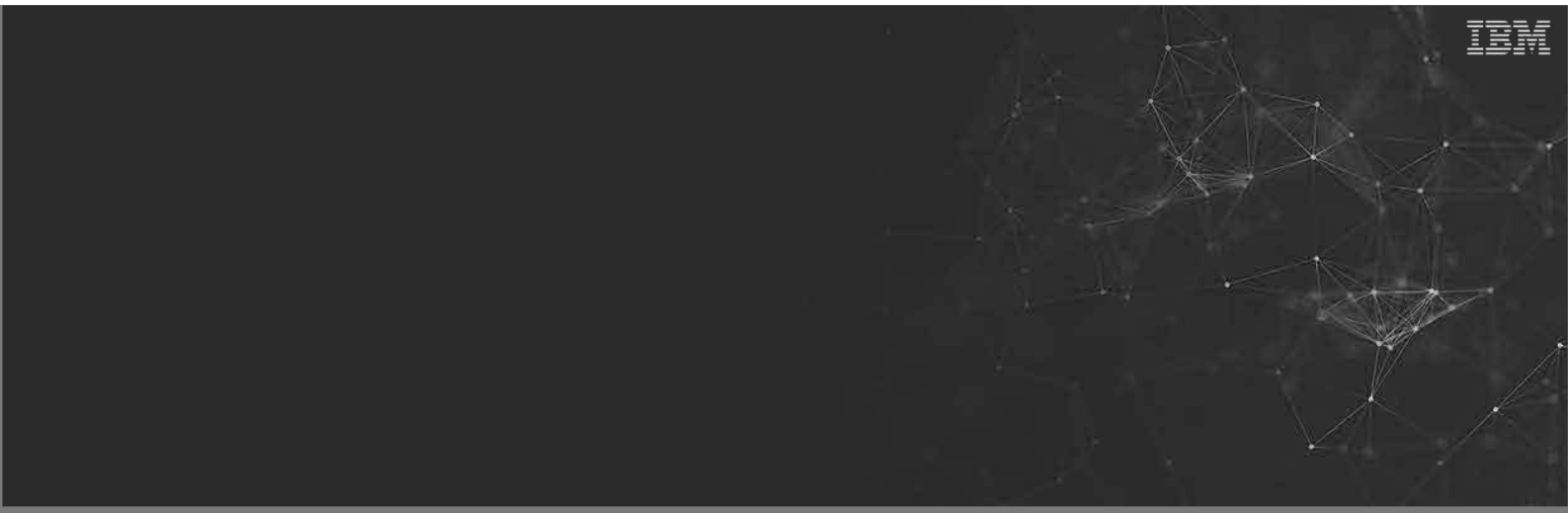


AI: why & why now

25 gigabytes of data per hour is generated by a connected car.

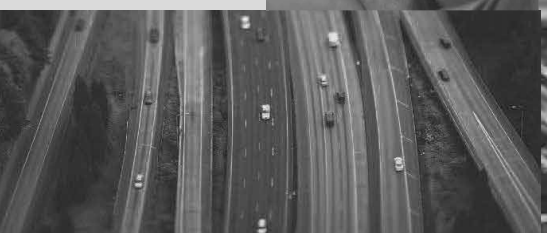
90% of cars will be connected by 2020.

80 million wearable health devices will be available by 2018.

153 exabytes of healthcare data generated by devices in 2013.

Increasing to **2,314 exabytes** in 2020.

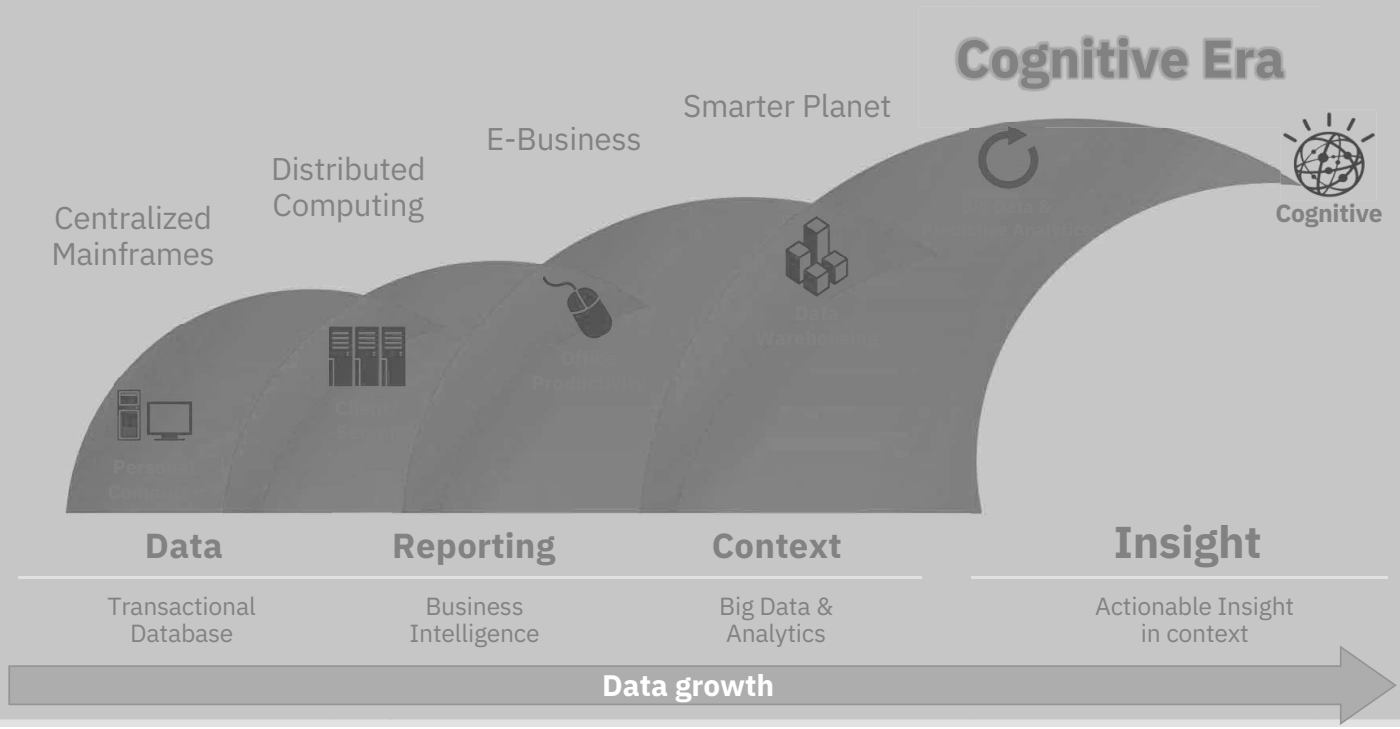
2.5 quintillion bytes of data generated daily by connected machines.

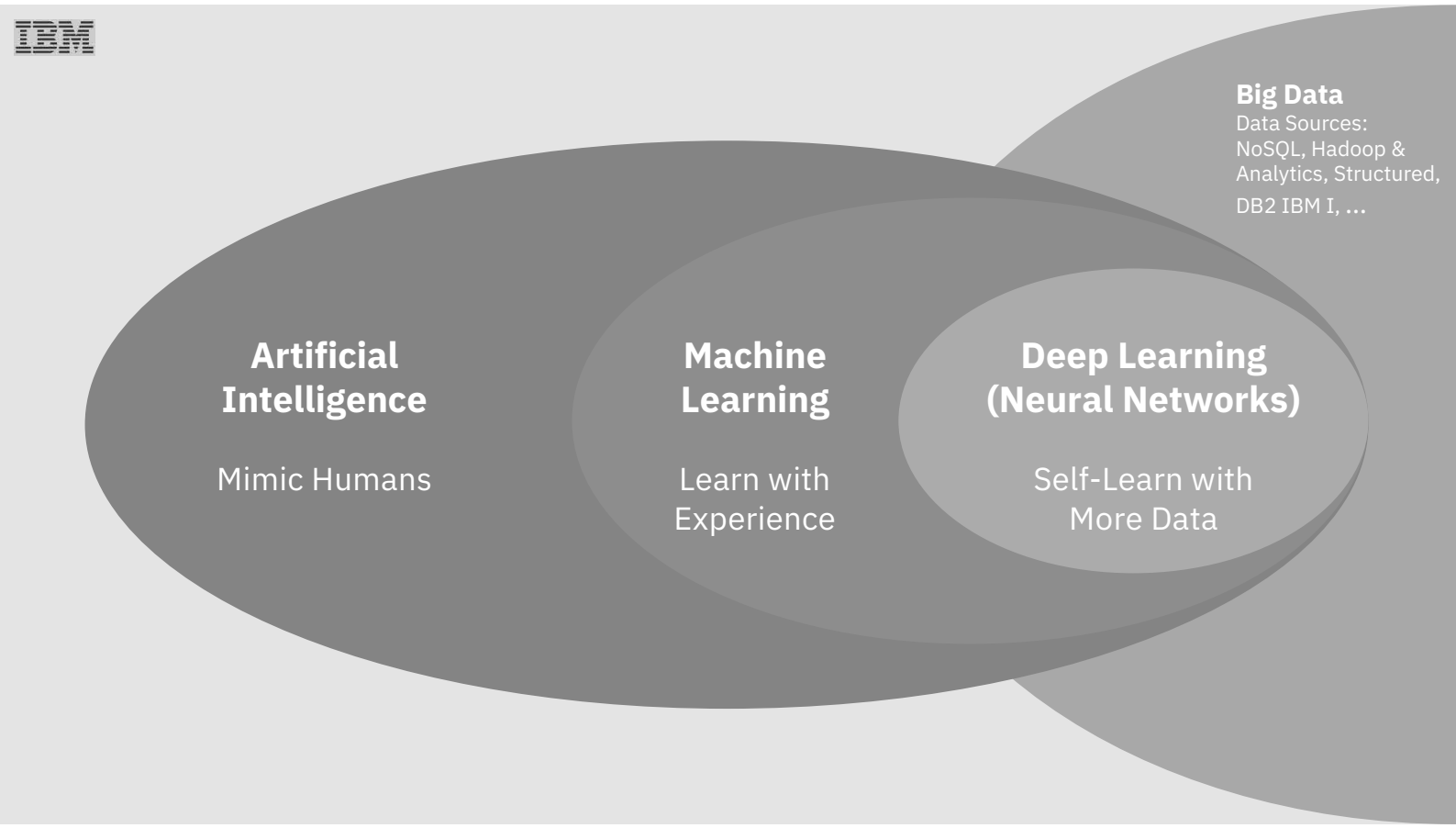


There will be **28 times more sensor-enabled devices than people** by the year 2020.

1.7 megabytes of data per second generated by every human being on the planet by 2020.

Our industry continues to transform through waves of change

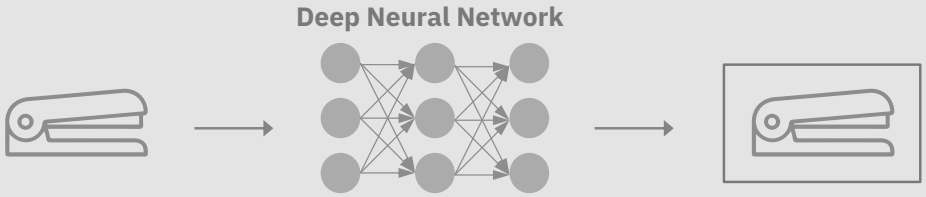



Big Data
Data Sources:
NoSQL, Hadoop &
Analytics, Structured,
DB2 IBM I, ...
Artificial Intelligence
Mimic Humans
Machine Learning
Learn with Experience
Deep Learning (Neural Networks)
Self-Learn with More Data

Machine Learning

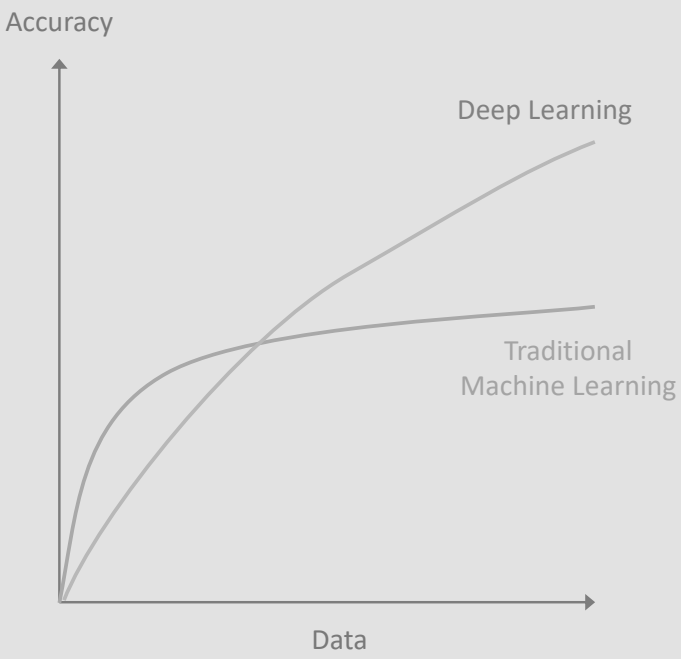
Input → Feature Extraction → Features → Classification → Output

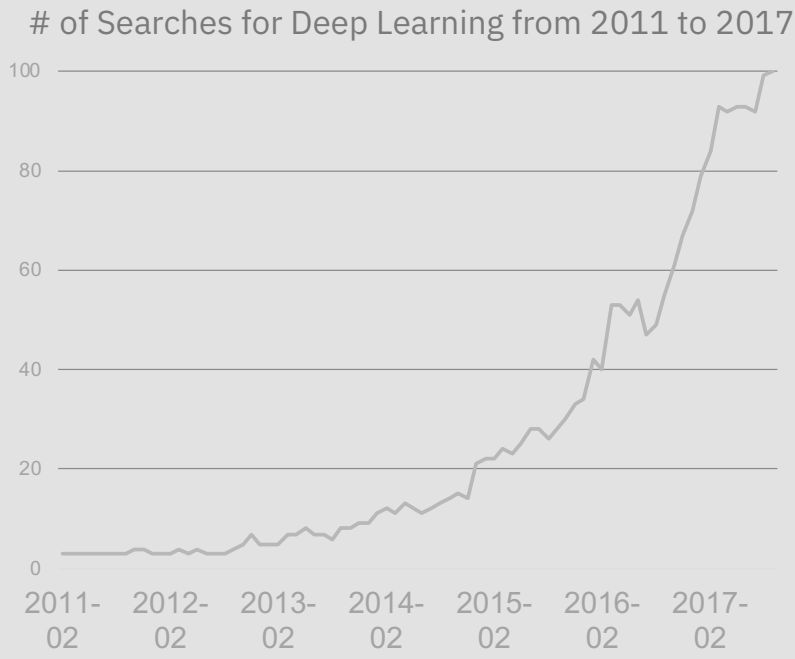
Deep Learning



Input → Feature Extraction & Classification → Output

Deep Learning Has Revolutionized Machine Learning





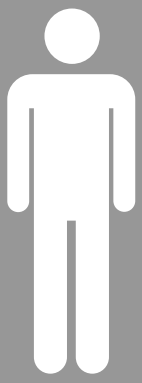
Source: Google Trends. Search term "Deep Learning"

2011

26% Errors

Machine Learning Based

Humans



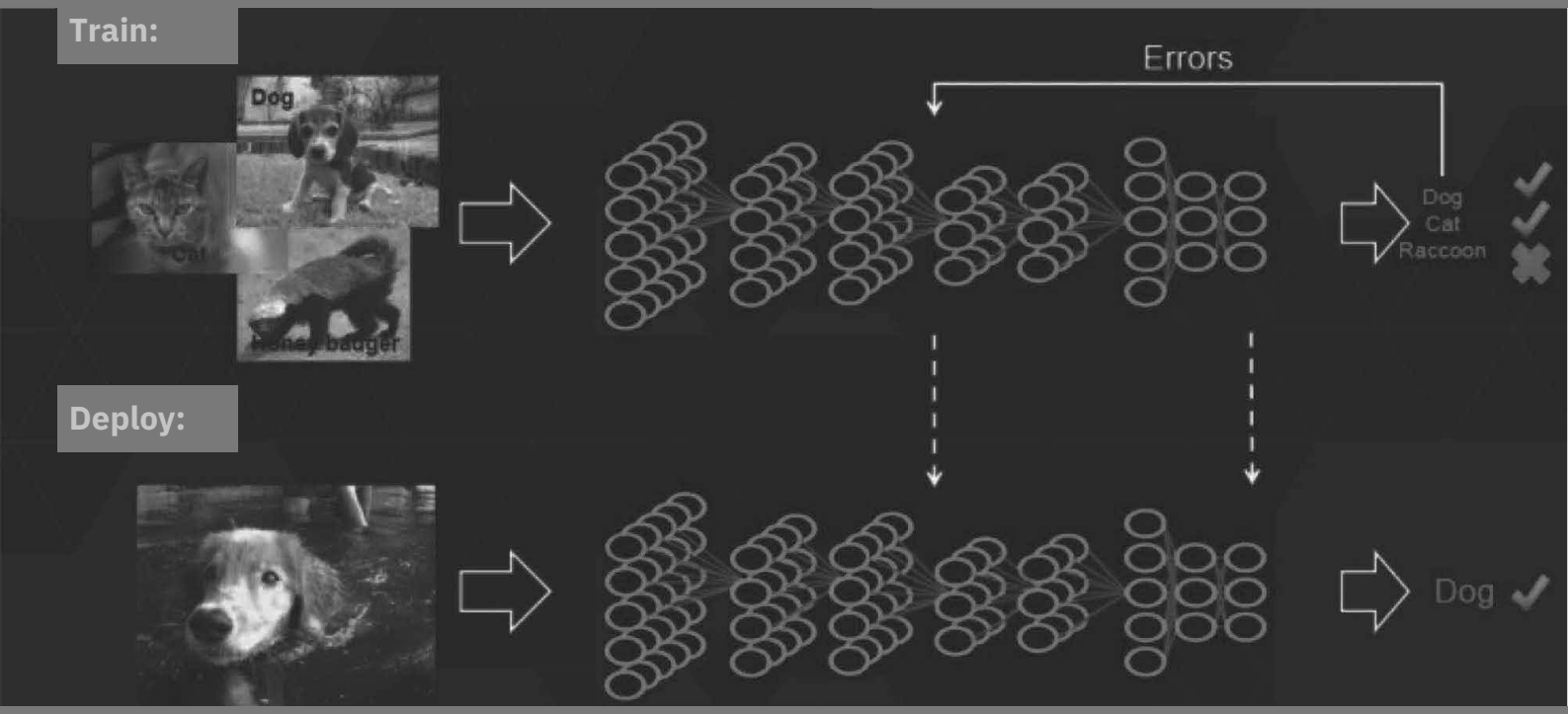
5% Error

2016

3% Errors

Deep Learning Based

Deep Learning Approach



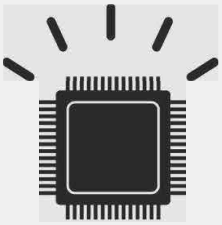
Why Deep Learning now?

Big Data

Large-scale training data

AI

Algorithmic Innovations

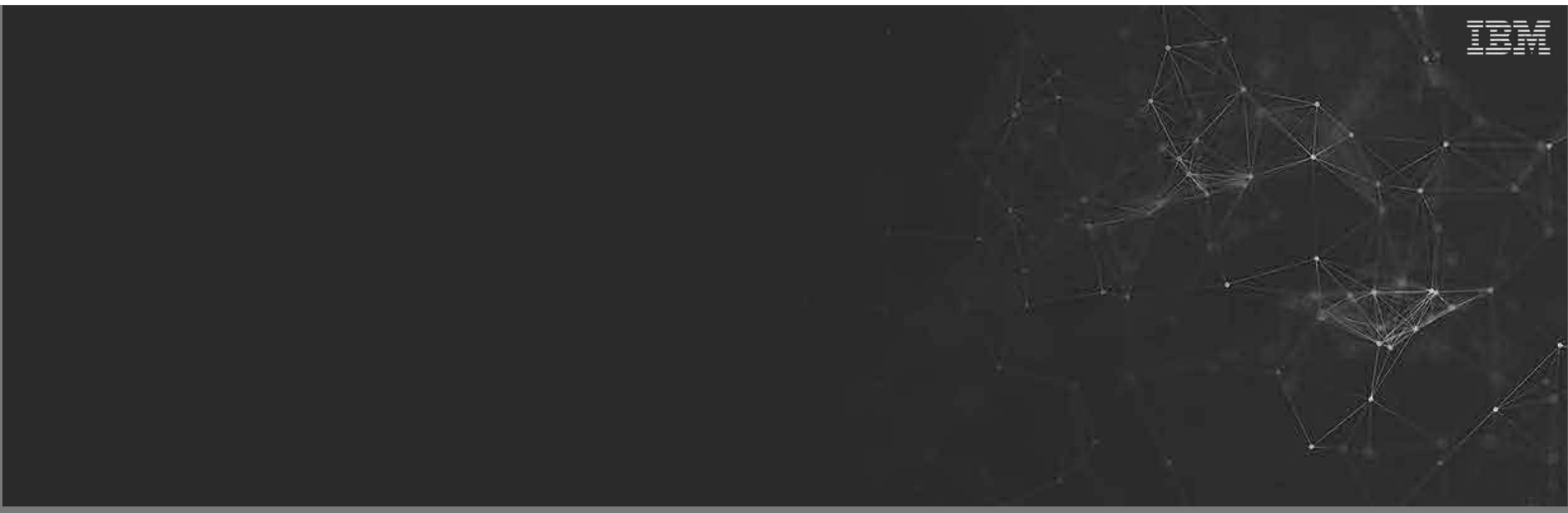


HPC

Highly parallel compute infrastructure

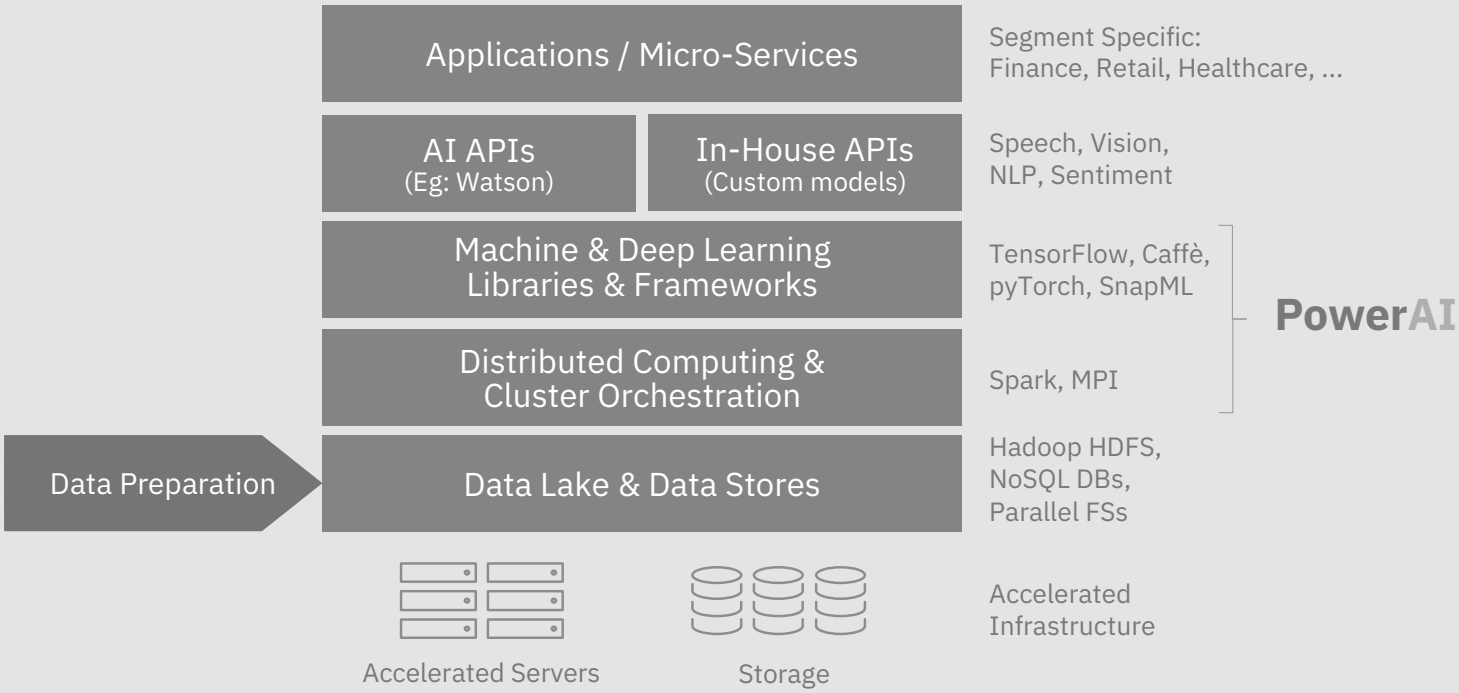
Core computation in training DNN: Dense matrix x Vector

A higly parallel workload	Experiment with different model for their dataset	Minimize training times



Cognitive Systems Strategy

IBM Infrastructure Stack



What's in the training of deep neural networks?

Data

Millions of images, sentences

Terabytes

Neural network model

Billions of parameters

Gigabytes

Computation

Iterative gradient based search

Millions of iterations
Mainly matrix operations

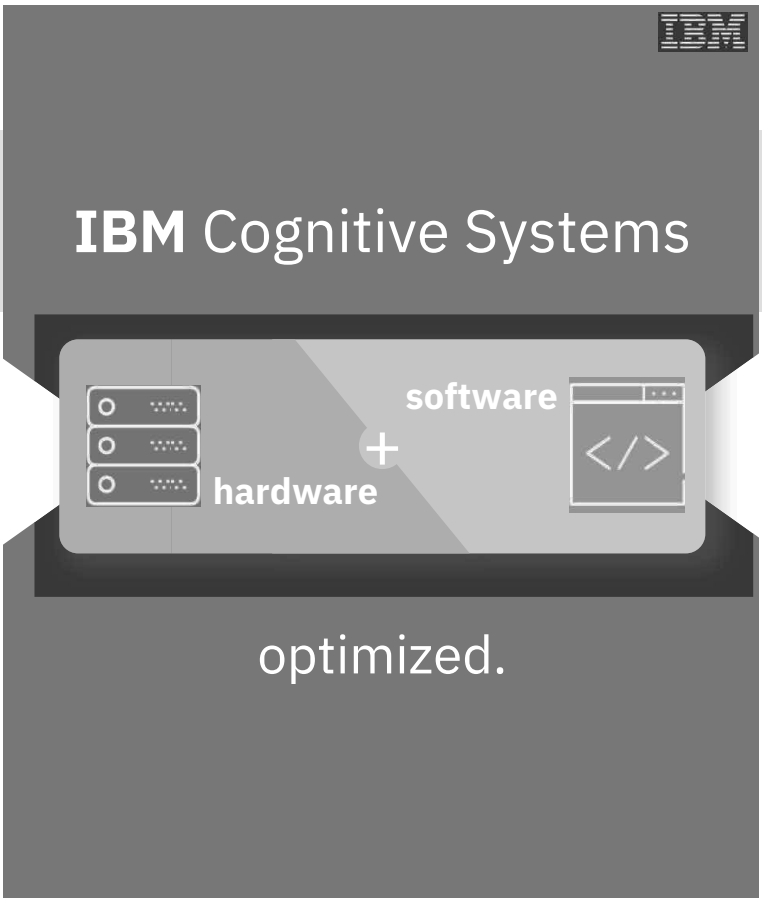
Workload characteristics: Both compute and data intensive!

POWER 9

"The only processor specifically designed for the AI era."

At a glance:

4x	more threads for high performance vs. x86
9.5x	more I/O bandwidth than x86
2.6x	more RAM possible vs. x86 (up to 2 TB)
1st	first CPU to deliver PCIe Gen 4



PowerAI

"IBM PowerAI: the complete environment for data science as a service"

Features:

TF Keras Caffe	Optimize the most popular DL opensource frameworks for this specific hardware
DDL	**D**istributed **D**eep **L**earning: scale training across multiple nodes
LMS	**L**arge **M**odel **S**upport: exploit the system memory (up to 2 TB) with GPUs
Snap**ML**	IBM library that enables GPU acceleration and scaling for ML algorithms

OpenPOWER

References

Google

*«Google announced that their IBM POWER9-based server, **Zaius**, is deployed and in the process of scaling up in their Data Center. Google's Maire Mahoney declared Zaius "Google Strong" and they are actively adding new production workloads onto Zaius and POWER9.»*

PayPal

«PayPal used IBM's OpenPOWER Systems and PowerAI to accelerate deep learning research for fraud prevention by unlocking the computation power on extra large datasets with the Power architecture.»

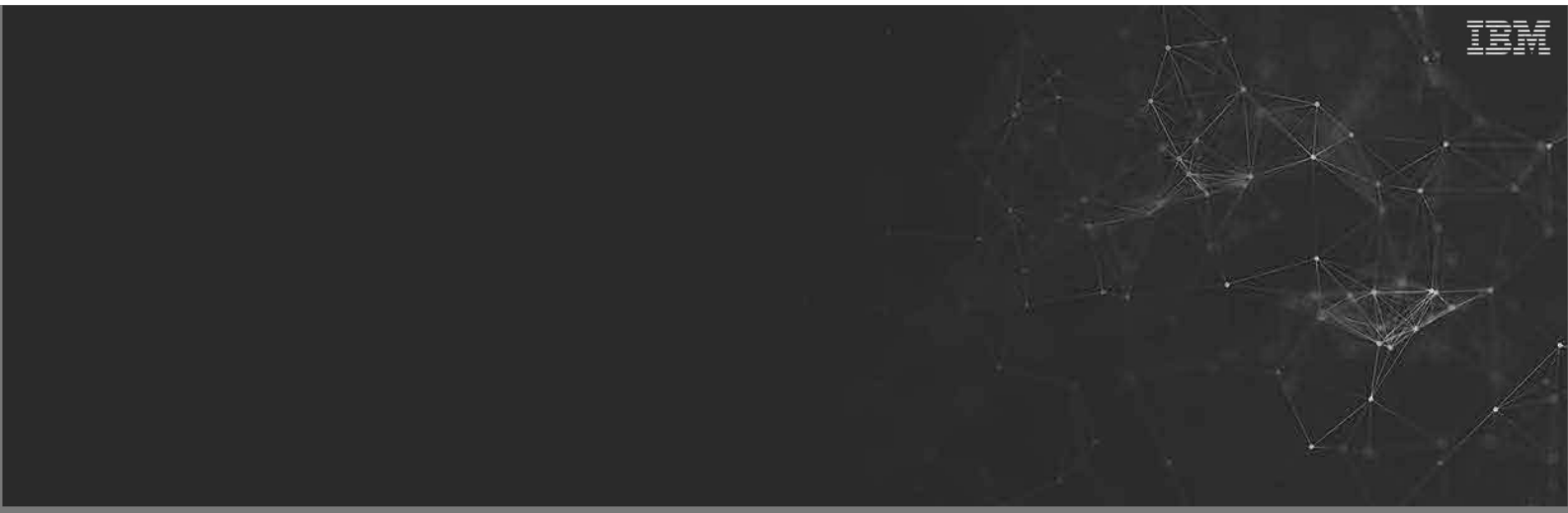
Tencent

«Tencent recently purchased a number of OpenPOWER-based systems to add to its growing enterprise data center. With adoption of OpenPOWER technology, Tencent's overall efficiency has improved by more than 30%, and with savings of 30% on rack and server resources.»

335+
Members

33
Countries

70+
ISVs



Deep Learning Hardware Stack

IBM Power Systems AC922

Technical specifications

POWER 9 processor

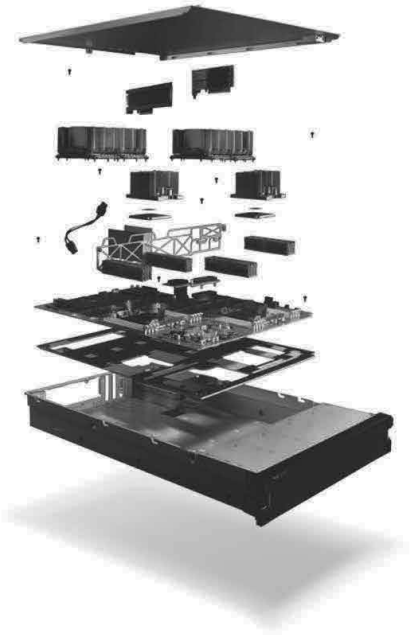
- 14nm lithography
- 20cores (x2 sockets)
- 2.00 GHz (2.87 GHz Turbo)
- Up to 2TB RAM

GPU Accelerated

- Up to 6x NVIDIA Tesla V100 (32gb) per server

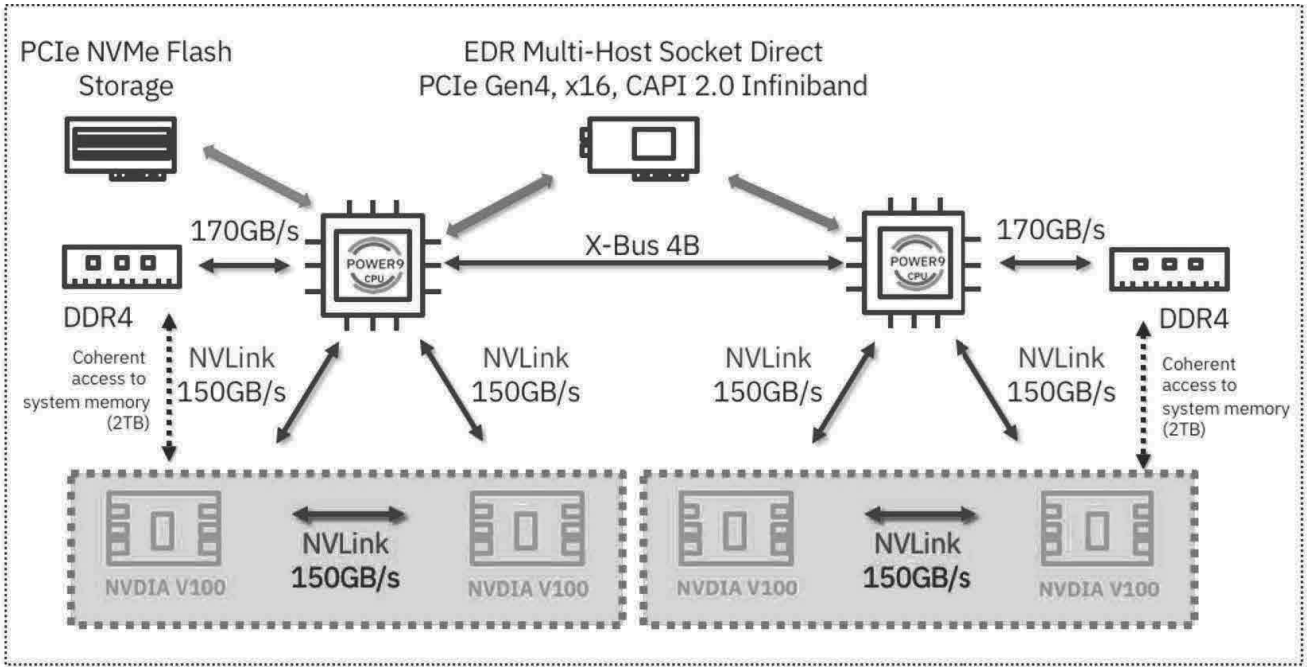
I/O protocols

- OpenCAPI
- NVLink 2.0
- PCIe Gen4



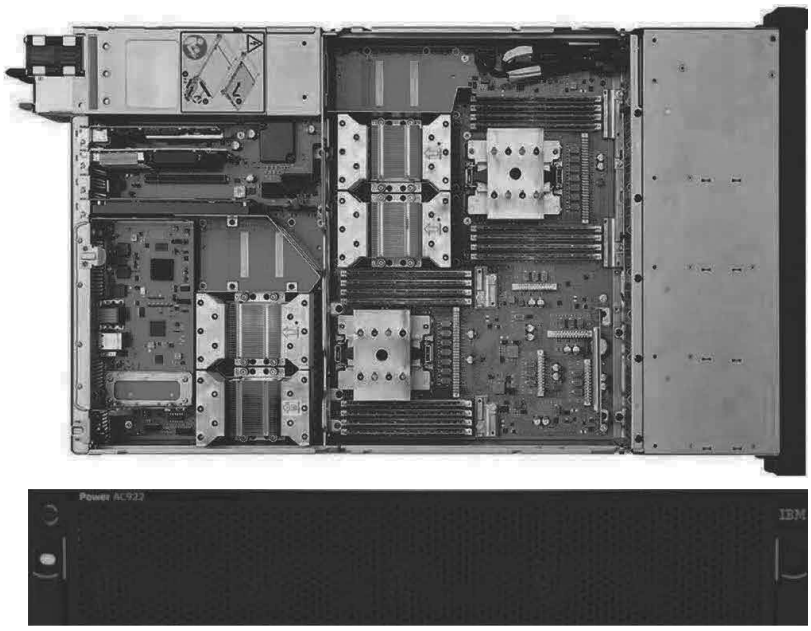
IBM AC922 Deep Learning System Architecture

AC922-GTG



IBM AC922 Deep Learning System

AC922-GTG



Processor Features

- **32 Cores** Processor Modules
  190W – 250W (2.25GHZ - 3.12GHZ)
- **40 Cores** Processor Modules
  190W – 250W (2.25GHZ - 2.80GHZ)

Memory Options

- 128GB, 256GB, 512GB DDR4
- 1TB, 2TB DDR4

Storage Features

- SSD 960GB 2.5” SATA
- SSD 1.92TB 2.5” SATA
- SSD 3.84TB 2.5” SATA
- 1.6TB NVMe PCIe Flash Adapter
- 3.2TB NVMe PCIe Flash Adapter
- 6.4TB NVMe PCIe Flash Adapter

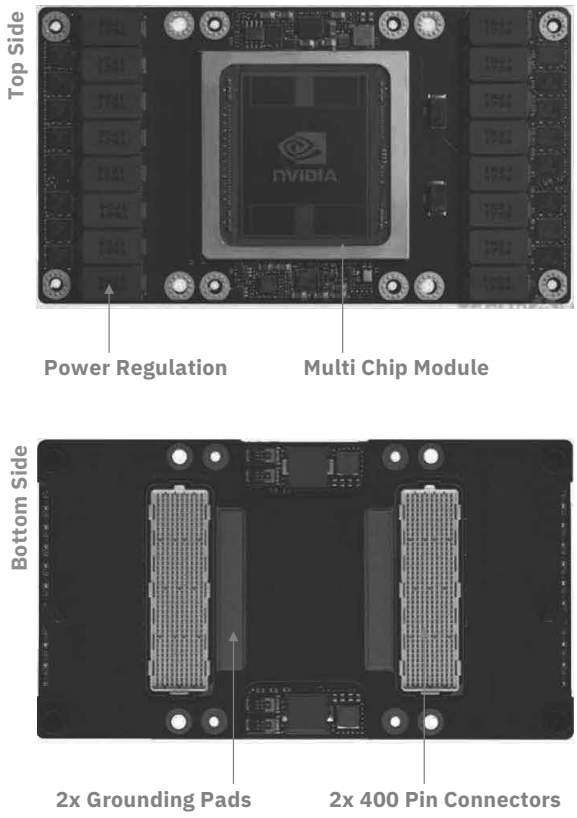
Accelerators Features

- NVIDIA V100 SMX2 16GB HBM2
- NVIDIA V100 SMX2 32GB HBM2
- Xilinix ADM-PCIE-9V3 FPGA

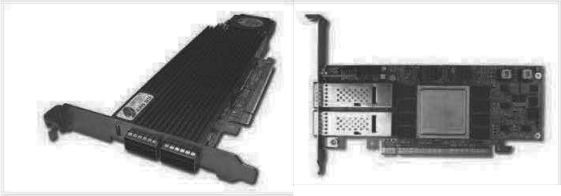
NVIDIA GPU Details

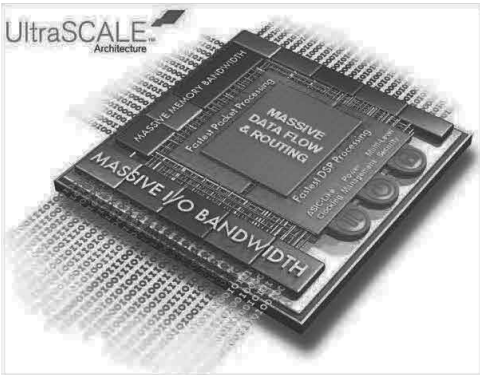
Volta SMX2 GPU Accelerator

NVIDIA Volta Specifications	
NVIDIA Tensor Cores	640
NVIDIA CUDA Cores	5120
Peak Double-Precision Performance	7.8 TFLOPS
Single-Precision Performance	15.7 TFLOPS
Tensor Performance	125 TFLOPS
Memory Bandwidth	900 GB/sec
GPU Memory Size	16 GB or 32GB HBM2
NVLink “Bricks” (8 lane interface)	6
NVLink Interconnect Bi-Directional	300 GB/sec
Maximum Power	300W



Server based **FPGA**: ie. ADM-PCIE-9V3





Features

- **Board Format :** Half-Length, low profile x16 PCIe form factor
- **Host I/F :** PCI Express Gen3 x8
- **Target Device :** Xilinx Virtex Ultrascale : XCVU095-2 - FFVC1517
- **SDRAM :** 2x banks of 1G x 72, DDR4-2400 (16GiB total), upgradable to 16GiB, DDR4-1866 (dual bank devices), per bank (32 GiB total)
- **FLASH :** On-board re-programmable flash memory for embedded configuration
- Optional integrated Board Support Package (BSP) including extensive FPGA example designs, plug and play drivers, and a mature Application Programming Interface (API)

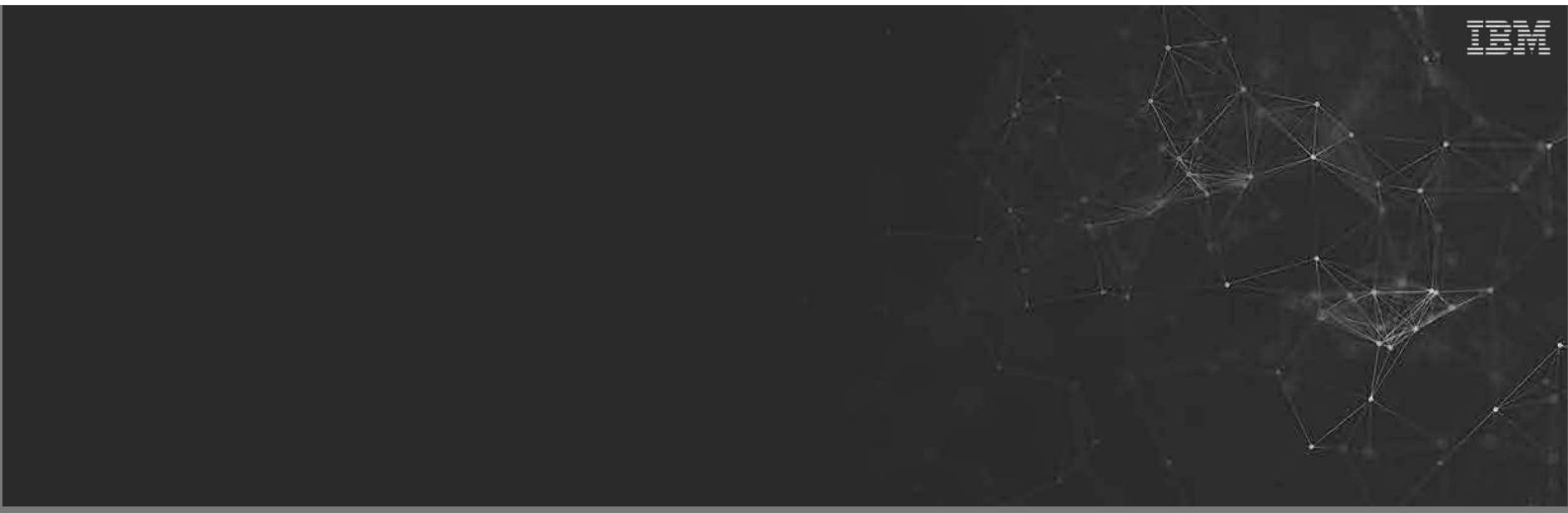
Summit

#1 of Top500: The World's Most Powerful and World's Smartest Supercomputer for Science

OAK RIDGE National Laboratory

Summit's base unit: **IBM AC922**

9.216	25	27.648	200	250
IBM POWER9 CPUs	gigabytes per second between nodes	NVIDIA Tesla GPUs	quadrillion calculation per second	petabytes storage capacity



Deep Learning Software Stack

PowerAI

Open-Source Based
Enterprise AI Platform

- Integrated & Supported AI Platform
- 3-4x Speedup for AI Training
- Ease of Use Tools for Data Scientists