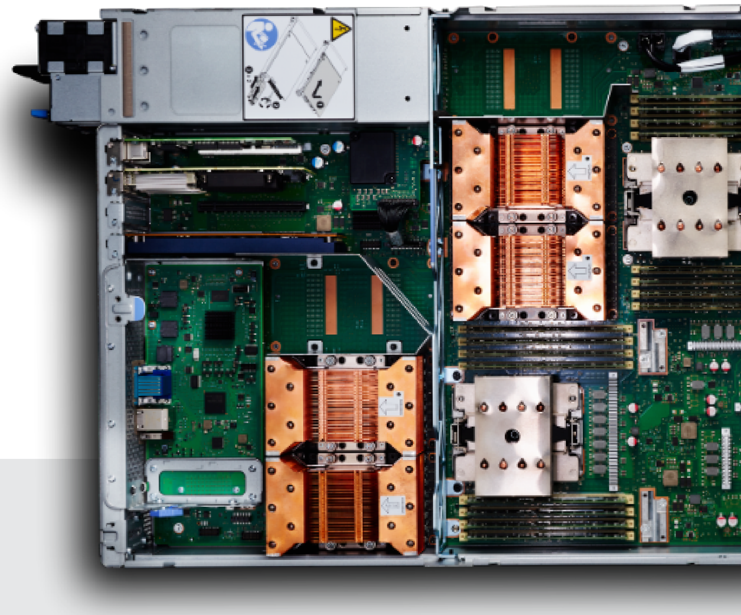
Train fast, learn faster

with the right infrastructure

Staying ahead of the competition requires enterprise AI. And infrastructure that breaks barriers.

Pair the Power AC922 with the IBM Watson Machine Learning Accelerator to help reduce model training times, accelerate iterations and improve insights.



Solve at the speed of Summit

The IBM® Power System AC922, which powers the world's fastest supercomputer, is purpose-built for AI training.

Power AC922 + IBM Watson Machine Learning Accelerator

3.7x **faster training for Caffe[1]**

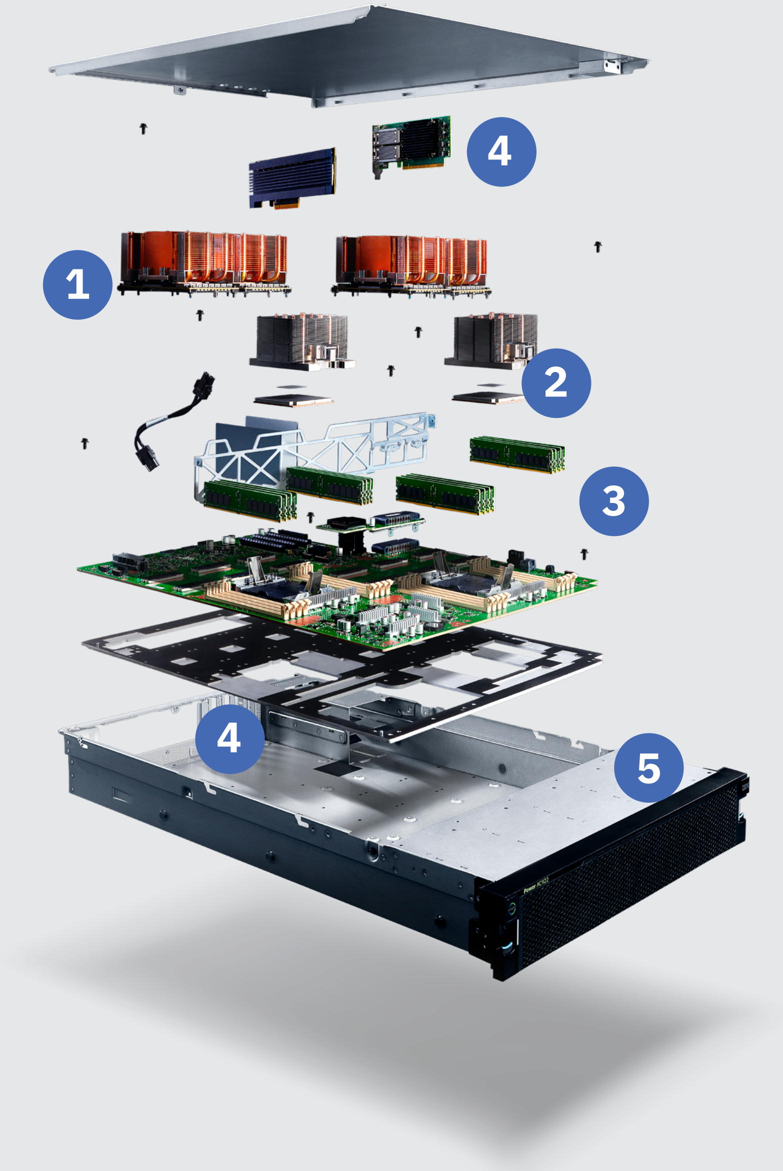
3.8x **faster training for Chainer[2]**

46x **faster Machine Learning iterations with SnapML[3]**

IBM POWER9™ + NVIDIA® NVLink™

5.6x **faster data throughput[4]**

- Up to 6 NVIDIA® Tesla® V100 GPUs + 2 IBM POWER9 processors
- State-of-the-art IO subsystems to handle massive data volume
- NVIDIA NVLink between CPUs and GPUs as well as GPUs



IBM Power System AC922

1. GPUs – Up to 6 NVIDIA Tesla V100 GPU processors
2. CPUs – 2 POWER9 processors with up to 44 cores
3. System memory – 2 TB max with 16 memory DIMM slots
4. 4x PCIe Gen 4 slots
5. Storage – 2 SFF (2.5") SATA drives, Max 4 TB (HDD) Max 7.68 TB (SSD)

Quickly build, train and retrain AI models using a server engineered to be the most powerful training platform

Get started now

https://www.ibm.com/it-infrastructure/power/enterprise-ai

[1] 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

[2] Results are based IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2560x2560). 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: Chainverv3 /LMS/Out of Core with patches found at https://github.com/cupy/cupy/pull/694 and https://github.com/chainer/chainer/pull/3762

[3] 46x SnapML (https://www.zurich.ibm.com/snapml/) In a newly published benchmark, using an online advertising dataset released by Criteo Labs (http://labs.criteo.com/2013/12/download-terabyte-click-logs/) with over 4 billion training examples, we train a logistic regression classifier in 91.5 seconds. This training time is 46x faster than the best result that has been previously reported (https://cloud.google.com/blog/products/gcp/using-google-cloud-machine-learning-to-predict-clicks-at-scale), which used TensorFlow on Google Cloud Platform to train the same model in 70 minutes.

[4] 5.6x I/O bandwidth claim based on CUDA H2D Bandwidth Test conducted on a Xeon E5-2640 V4 +P100 vs POWER9 + V100 (12 GB/s vs 68 GB/s rated)