Announcements

IBM Unveils Industry’s Most Advanced Server Designed for Artificial Intelligence

New Power Systems deliver nearly 4x deep-learning framework performance over x86 POWER9 processor designed for AI workloads Google, Department of Energy’s "Summit” and "Sierra" tap POWER9

Bengaluru, KARNATAKA, India - 06 Dec 2017: IBM (NYSE: IBM) today unveiled its next-generation Power Systems Servers incorporating its newly designed POWER9 processor. Built specifically for compute-intensive AI workloads, the new POWER9 systems are capable of improving the training times of deep learning frameworks by nearly 4x[i] allowing enterprises to build more accurate AI applications, faster.

The new POWER9-based AC922 Power Systems are the first to embed PCI-Express 4.0, next-generation NVIDIA NVLink and OpenCAPI, which combined can accelerate data movement, calculated at 9.5x[i][ii] faster than PCI-E 3.0 based x86 systems.

The system was designed to drive demonstrable performance improvements across popular AI frameworks such as Chainer, TensorFlow and Caffe, as well as accelerated databases such as Kinetica.

As a result, data scientists can build applications faster, ranging from deep learning insights in scientific research, real-time fraud detection and credit risk analysis.

POWER9 is at the heart of the soon-to-be most powerful data-intensive supercomputers in the world, the U.S. Department of Energy’s “Summit” and “Sierra” supercomputers, and has been tapped by Google.

“Google is excited about IBM's progress in the development of the latest POWER technology," said Bart Sano, VP of Google Platforms "The POWER9 OpenCAPI Bus and large memory capabilities allow for further opportunities for innovation in Google data centers."

Viswanath Ramaswamy, Director - Systems (India/South Asia) said, “The new IBM Power Systems Servers with POWER9 processor will be a game-changer for AI and deep learning workloads. The new processor delivers on unprecedented cognitive capabilities and can help Indian enterprises across all verticals to transform and up-scale on their AI and machine learning journey.

The new server (AC22) improves training time of deep learning frameworks by 4x (vs x86), in turn delivering 10x faster performance bandwidth acceleration. This is the only platform with NVIDIA NVLink, PCI Express 4.0 and OpenCapi giving breakthrough acceleration for modern AI, high performance computing and accelerated database workloads.”

Accelerating the Future with POWER9

Deep learning is a fast growing machine learning method that extracts information by crunching through millions of processes and data to detect and rank the most important aspects of the data.

To meet these growing industry demands, four years ago IBM set out to design the POWER9 chip on a blank sheet to build a new architecture to manage free-flowing data, streaming sensors and algorithms for data-intensive AI and deep learning workloads on Linux.

IBM is the only vendor that can provide enterprises with an infrastructure that incorporates cutting-edge hardware and software with the latest open-source innovations.

With PowerAI, IBM has optimized and simplified the deployment of deep learning frameworks and libraries on the Power architecture with acceleration, allowing data scientists to be up and running in minutes.

IBM Research is developing a wide array of technologies for the Power architecture. IBM researchers have already cut deep learning times from days to hours with the PowerAI Distributed Deep Learning toolkit.

Building an Open Ecosystem to Fuel Innovation

The era of AI demands more than tremendous processing power and unprecedented speed; it also demands an open ecosystem of innovative companies delivering technologies and tools. IBM serves as a catalyst for innovation to thrive, fueling an open, fast-growing community of more than 300 OpenPOWER Foundation and OpenCAPI Consortium members.

Learn more about POWER9 and the AC922: http://ibm.biz/BdjCQQ

Read more from Bob Picciano, Senior Vice President, IBM Cognitive Systems: https://www.ibm.com/blogs/think/2017/12/accelerating-ai/

---

[i] x86 PCI Express 3.0 (x16) peak transfer rate is 15.75 GB/sec = 16 lanes X 1GB/sec/lane x 128 bit/130 bit encoding.

[ii] POWER9 and next-generation NVIDIA NVLink peak transfer rate is 150 GB/sec = 48 lanes x 3.2265625 GB/sec x 64 bit/66 bit encoding.

1 Results of 3.7X are based IBM Internal Measurements running 1000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2560x2560). Hardware: 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

[1] Results of 3.8X 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

[1] x86 PCI Express 3.0 (x16) peak transfer rate is 15.75 GB/sec = 16 lanes X 1GB/sec/lane x 128 bit/130 bit encoding.

[1] POWER9 and next-generation NVIDIA NVLink peak transfer rate is 150 GB/sec = 48 lanes x 3.2265625 GB/sec x 64 bit/66 bit encoding.