MACHINE LEARNING AT SUPERCOMPUTER SPEED

WEI LI PH.D.
Vice President, CVCG
GM, Machine Learning and Translation

May 2018
DATA GROWTH IS EXPLODING

1.5 GB¹
3,000 GB²
4,000 GB²
40 K GB²
1 M GB³

Avg. Internet User | Smart Hospital | Autonomous Driving | Airplane | Smart Factory

1M GB³

All numbers are approximated.

SUPERCOMPUTER IN A BOX

Paragon
#1 in 1993
143 GFLOPS¹

ASCI Red
#1 in 1997
1 TFLOPS²

One Xeon Server
2018
8 TFLOPS³

Cloud
2018
PFLOPS-EFLOPS?

¹ https://www.top500.org/featured/systems/intel-xps-140-paragon-sandia-national-labs/
³ https://ark.intel.com/products/120496/Intel-Xeon-Platinum-8180-Processor-38.5M-Cache-2.50-GHz
AI PERFORMANCE = HARDWARE + SOFTWARE

Deliver significant AI performance with hardware and software optimizations on Intel® Xeon® Scalable Family

**INFERENCETHROUGHPUT**

- **Up to 198x**
- Intel® Xeon® Platinum 8180 Processor
- higher Intel optimized Caffe GoogleNet v1 with Intel® MKL
- inference throughput compared to
- Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe
- Inference and training throughput uses FP32 instructions

**TRAININGTHROUGHPUT**

- **Up to 127x**
- Intel® Xeon® Platinum 8180 Processor
- higher Intel Optimized Caffe AlexNet with Intel® MKL
- training throughput compared to
- Intel® Xeon® Processor E5-2699 v3 with BVLC-Caffe

Optimized
Frameworks

Optimized Intel®
MKL Libraries

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system. The results may need to be revisited to account for these updates. Results depend on the specific platform configurations and workloads utilized in testing. The results and comparisons to any particular user system or application are not guarantees or depictions of expected performance. The results and comparisons to other benchmarks are not used for any purpose of comparison or endorsement of any specific system or processor. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on non-Intel microprocessors. Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on non-Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.
Moreover, the leading entries were split across TPUs, GPUs and CPUs: no single type of hardware dominated in all cases. Deep learning systems are both diverse and rapidly improving.
NEURAL MACHINE TRANSLATION: 1 CPU NODE

• Translate one sentence from German to English in milliseconds

4X higher Inference Performance on Intel® Xeon® Scalable Processors COMPARED TO NVIDIA V100 on Neural machine translation

HIGH-RISK INSPECTION BY DRONES: 1 CPU NODE

Time to train in **six hours**

FRAMEWORK

HARDWARE

SILICON PACKAGE DETECTION: 8 CPU NODES

Training within one hour

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Configuration Details: see the backup slides. Source: Intel measured as of February 2018.

Optimization notice: Other names and brands may be claimed as the property of others.

V. Codreanu et al. "Achieving Deep Learning Training in less than 40 Minutes"  
NATURAL LANGUAGE PROCESSING - 1.1M VCPUS

Topic Modeling on a large compute cluster in the cloud with 1.1 million vCPUs for over half million documents

Used cost-optimized Amazon EC2 Spot CPU instances

Clemson university

Use optimized frameworks at ai.intel.com/framework-optimizations/
Visit www.intel.com/ai for more information