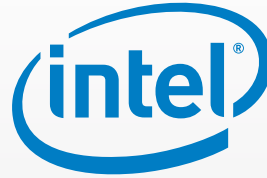


HIGH PERFORMANCE COMPUTING AND AI



BEST SERVED TOGETHER!

Amazon Web Services (AWS) and Intel share a passion for delivering constant innovation on the cloud. Together, they have developed a variety of resources and technologies for High Performance Computing, Big Data, Artificial Intelligence / Machine Learning and the Internet of Things.

Unlock the Potential of Your AWS Instance With Intel Parallel Studio XE

INSTALL INTEL PARALLEL STUDIO ON YOUR AWS INSTANCE TODAY!

POWER HIGH COMPUTE WITH EC2 INSTANCE

| C4 | C5 |
|---------------|--|
| 36 vCPUs | 72 vCPUs 2X vCPUs |
| "Haswell" | "Skylake" AVX 512 2X performance |
| 4 Gbps to EBS | 12 Gbps to EBS 3X throughput |
| 60 GiB memory | 144 GiB memory 2.4X memory |

Power of Amazon EC2 C5 instances

- Feature up to 72 vCPUs, a base clock frequency of 3.0 GHZ and 144 GiB of memory to run the most demanding HPC workloads
- 1.65 times higher energy efficiency and a system-level performance
- Ideal solution for financial operations, scientific modelling, and distributed analytics.
- Speed up the efficiency of deep learning training 2.2 times, and deliver 2.4 times higher deep learning inference performance
- 1.25% price/performance improvement over the C4

PARALLALIZE SOFTWARE TO MAXIMIZE PERFORMANCE



Deliver FASTER CODE FASTER

- Scale applications with improved parallel performance using Intel® Advanced Vector Extensions 512 (Intel® AVX-512) instructions
- Gain greater scalability and reduce latency with next generation Intel® MPI Library
- Vectorize and use threads (using OpenMP*) to take advantage of the latest SIMD-enabled hardware, including Intel AVX-512
- Speed up machine with Intel® Distribution for Python* and Intel® Data Analytics Acceleration Library (Intel® DAAL)

INTEL ACCELERATING AI

Solutions

Solution Architects

ARTIFICIAL INTELLIGENCE

AI Solutions Catalog (Public & Internal)

Platforms
Finance
Healthcare
Energy
Industrial
Transport
Retail
Home
More...

TOOLKITS

App Developers

DEEP LEARNING DEPLOYMENT

OpenVINO™

Open Visual Inference & Neural Network Optimization toolkit for inference deployment on CPU/GPU/FPGA for TF, Caffe* & MXNet*

Intel® Movidius™ SDK

Optimized inference deployment on Intel VPUs for TensorFlow* & Caffe*

Intel® Saffron™ AI

Cognitive solutions on CPU for anti-money laundering, predictive maintenance, more

Intel® Deep Learning Studio

Open-source tool to compress deep learning development cycle

libraries

Data Scientists

MACHINE LEARNING LIBRARIES

Python

- Scikit-learn
- Pandas
- NumPy

R

- Cart
- Random Forest
- e1071

Distributed

- MLlib (on Spark)
- Mahout

foundation

Library Developers

DEEP LEARNING FRAMEWORKS

Now optimized for CPU | Optimizations in progress

TensorFlow*

MXNet*

Caffe*

BigDL/Spark*

Caffe2*

PyTorch*

PaddlePaddle*

Hardware

IT System Architects

AI FOUNDATION

ATOM processor

CORE processor

XEON processor

Data Center Edge Device

DEEP LEARNING ACCELERATORS

NERVANA accelerator

STRATIX 10 accelerator

ARRIA 10 accelerator

MOVIDIUS accelerator

Training ← Inference →

AWS MACHINE LEARNING ACCELERATION WITH INTEL

AWS C5 Deep Learning AMI
Optimized for Intel CPU
(Training)

7.4X

faster than training on the
stock Tensorflow 1.6 binaries

AWS C5 Deep Learning AMI
Optimized for Intel CPU
(Inference)

12X

faster than default
configuration for NMT Inference
with MxNet

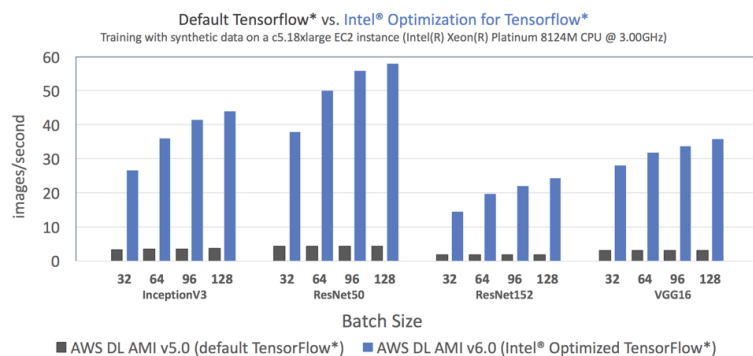
AWS SageMaker ML
Optimized for Intel CPU
(Training & Inference)

10

Machine Learning Algorithms
optimized for IA CPU

Optimized Deep Learning Frameworks on EC2 CPU Instances

AWS DL AMI v6.0 includes Intel® Optimizations for TensorFlow* - pre-built and ready to use with Intel® Math Kernel Library and Math Kernel Library for Deep Neural Networks (Intel® MKL and Intel® MKL-DNN) primitives



For the data scientist, consequences of using Intel® Optimization for TensorFlow manifest as significant speedups across the deep learning design space. Figure 1 compares default TensorFlow (previously available in AWS' DL AMI v5.0) vs. Intel Optimization for TensorFlow (currently available in AWS' DL AMI v6.0) training throughput on four benchmark topologies: InceptionV3, ResNet50, ResNet152 and VGG16.

