ACCELERATE MACHINE LEARNING WITH ALLUXIO AND LATEST-GEN INTEL® XEON® SCALABLE PROCESSORS

Highlights

- Intel and Alluxio collaborate to recommend AWS instance types leveraging Intel® Deep Learning Boost's BFloat16 capability for PyTorch performance on Latest-Generation Intel® Xeon® Scalable Processors.
- Leveraging Alluxio to accelerate data access for end-to-end training pipelines, 3rd Gen Intel® Xeon® Scalable Processors deliver 20-25% better price/performance compared to the prior generation available on AWS.

Abstract

Intel Deep Learning (DL) Boost with BFloat16 (BF16) demonstrates benefits across deep learning training workloads with the same accuracy as 32-bit floating-point (single-precision) (FP32). Recently Amazon introduced EC2 M6i instances powered by the latest-generation Intel Xeon Scalable Processors. Intel and Alluxio collaborate to measure a 20-25% price/performance improvement over the prior generation for machine learning models with PyTorch on AWS. This collaboration demonstrates data preprocessing and training at lower cost on CPUs using Alluxio as the data access layer to cloud storage.

Introduction

Gartner expects that by the end of 2024, 75% of enterprises will shift from piloting to operationalizing AI, driving a 5X increase in data and analytics infrastructures[1]. Operationalized AI architectures are expected to approach maturity rapidly due to orchestration initiatives in the cloud. As the volume of data grows, with cloud storage as an increasingly popular choice for storing varied types of data for machine learning initiatives, new architectures & tools are being developed to extract and deliver value.

Amongst these architectures, the separation of compute and storage is becoming increasingly attractive as it enables companies to scale storage independently of compute, to reduce both capital expenditures and operating expenses. This disaggregated architecture introduces performance loss for certain types of workloads because of network latencies as data is not available locally for computation. Machine learning in the cloud experiences the same problem as compute instances are separated from cloud storage.

AWS S3 is a common storage choice for the vast amounts of data required by machine learning and deep learning initiatives. AWS announced EC2 M6i compute instances powered by the latest-generation Intel Xeon Scalable Processors (code-named Ice Lake) delivering price/performance benefits for a variety of workloads[2]. Alluxio bridges the gap between disaggregated storage and EC2 compute to accentuate the benefits from evolution in compute technology (see Figure 1).

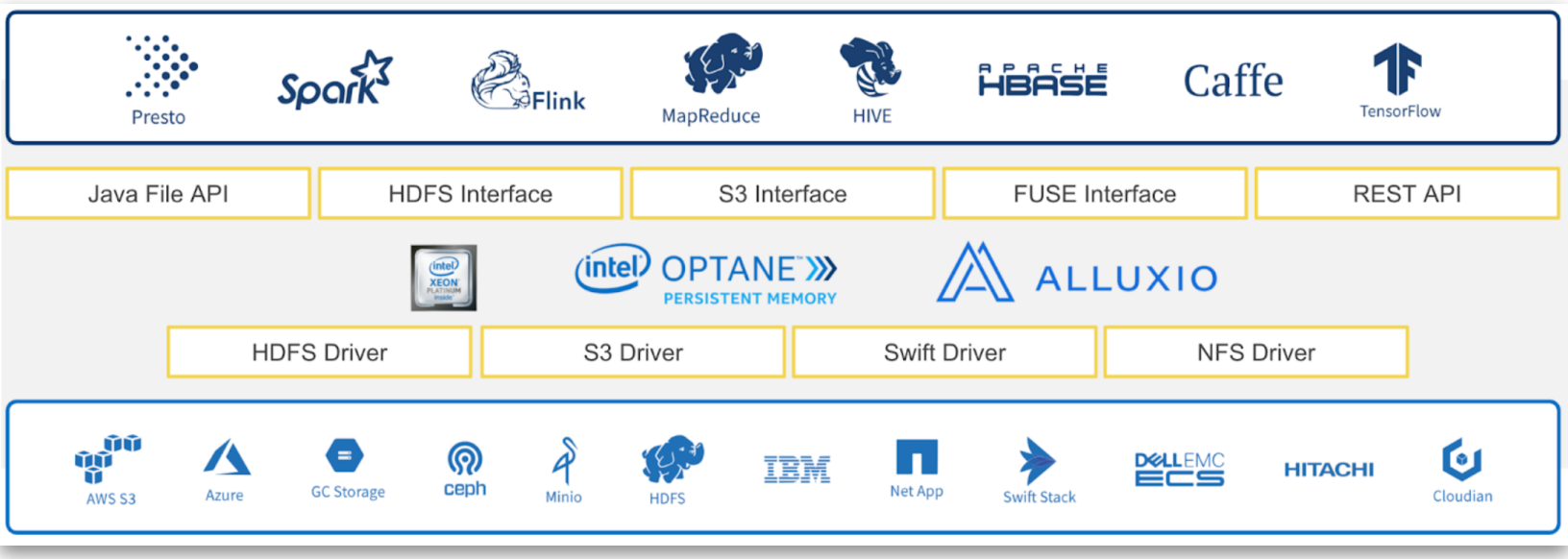

Figure 1: Intel + Alluxio for Analytics & AI

Alluxio is an open source data orchestration platform for analytics and machine learning applications. It is a software solution which sits between disaggregated compute and storage to connect data analytics applications to various heterogeneous data sources and bring data close to compute[3].

In this article, we quantify the benefits of the latest-generation Intel Xeon processors over the prior generation in AWS using two workloads to measure the price/performance of CPU instances for deep learning. The first workload is Resnet-50[4] training optimized using Intel Extension for PyTorch (IPEX)[5], demonstrating 25% better cost/performance with AWS M6i EC2 instances over M5n EC2 instances. The second is a data loading[6] workload, exercising the data loading and pre-processing step of a training pipeline on CPUs, demonstrating 20% better cost/performance using the same instance types as the prior workload.

Architecture

The drive for better accuracy, faster speed, and lower cost for machine learning/deep learning training leads many companies to adopt distributed training in the cloud. Growing datasets for training accuracy often do not fit onto a single machine. Also, computation tasks need to be distributed to multiple machines to speed up training.

Cloud storage, such as AWS S3, is becoming more attractive with high scalability, reduced cost and ease of use. However, providing a high throughput to sustain high compute utilization can be challenging. Data access becomes the main bottleneck when serving distributed training in the cloud due to the following challenges:

a. Training tasks distributed across a cluster all need access to the training data
b. Data throughput needs to be high to guarantee a high CPU utilization rate

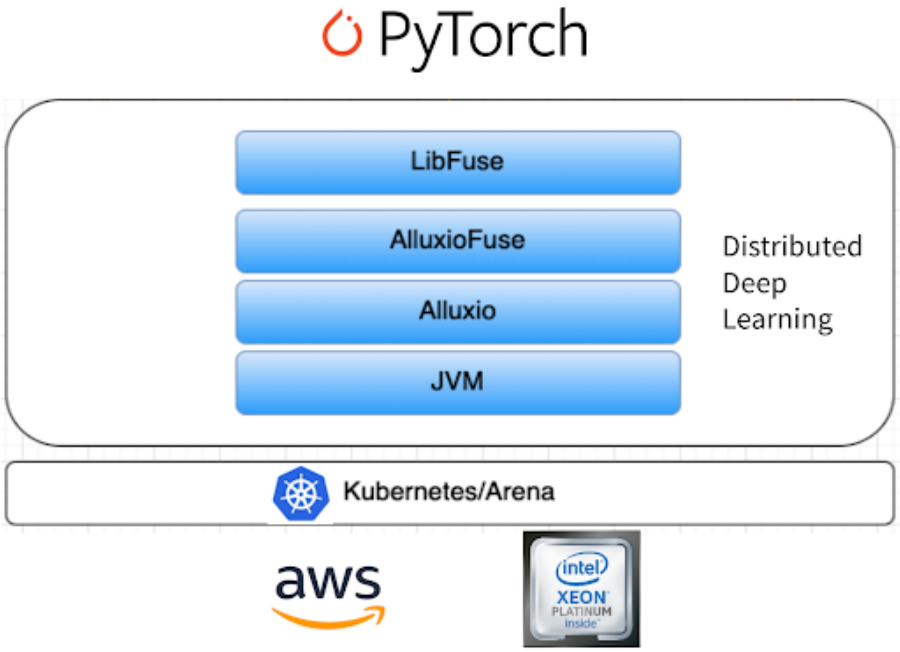

Figure 2: Architecture evaluated as part of this article.

Being able to access data is a basic requirement and making sure the data access solution provides a high CPU utilization rate is critical. By using Alluxio's preloading and on-demand caching abilities, loading data from the source with data caching can be done in parallel with the actual training task[7]. Training benefits from high data throughput when accessing data cached in Alluxio without the need to wait to fully cache the data before training.

Caching data eliminates the need for repeated data access and reduces the network bandwidth required between AWS EC2 and S3. EC2 M6i instances offer the latest-generation Intel Xeon Scalable processors with lower network bandwidth over the prior-generation M5n instance types. By using Alluxio, we demonstrate better cost/performance using the newer instance types even though the network bandwidth for data loading and pre-processing is limited.

With traditional deep learning training alternatives, data is pre-processed through transformation with CPU instances and written back to cloud storage. The pre-processed data is then read back for the training step, incurring high network transfer costs. With Alluxio, the Intel Xeon scalable processors are kept busy without waiting for network I/O as data is cached locally on compute instances across different steps of the training pipeline.

Benchmark Results

To evaluate the difference between generations of Intel Xeon Scalable processors on AWS, we ran two workloads with the same system configuration and testing methodology listed below.

System Configuration

We ran a cluster with 1 edge and 4 worker instances on AWS EC2. The configuration for each worker instance is captured in the table below. Hourly cost reflects AWS EC2 on-demand instance pricing as of this writing[8].

Intel Xeon Scalable Processor Generation	Instance Type	Hourly Cost	Worker Count	vCPU	Memory (GiB)
3rd gen (code name Ice Lake)	m6i.8xlarge	$1.54	4	32	128
2nd gen (code name Cascade Lake)	m5n.8xlarge	$1.90	4	32	128

Table 1: AWS Instance Configuration

Alluxio has master and worker processes. A single instance of the Alluxio master was used. An instance of the Alluxio worker process was run on each of the 4 worker instances. Local SSDs were used for caching data.

CPU	Memory	SSD total (GB)
8	32	640

Table 2: Alluxio Resource Configuration

PyTorch is given a majority of the resources on each worker instance.

CPU	Memory (G)
24	64
24	64

Table 3: PyTorch Resource Configuration

Testing Methodology

The benchmark used the Imagenet dataset[9]. The required datasets are named ILSVRC2012_img_train.tar and ILSVRC2012_img_val.tar. Training data was preloaded from AWS S3 into Alluxio managed SSDs local to EC2 instances before the first iteration.

For the data loading benchmark, only data loading and preprocessing is performed with the training step omitted. This benchmark stresses the CPU loading data from Alluxio.

For Resnet-50 training, the Intel extension for PyTorch is used to measure the impact of Intel Xeon Scalable processors with Intel Deep Learning Boost and BFloat16 for model training.

Results

	Average Speed (Images / sec)		Total Time (sec)	
Intel Xeon Scalable Processor Generation	Data loading	Training	Data Loading	Training
3rd gen (code name Ice Lake)	1580	379	805	3347
2nd gen (code name Cascade Lake)	1564	350	813	3645

Table 4: Benchmarking results compare performance for M6i vs. M5n AWS EC2 instance type.

References

[1] Gartner Top 10 Trends in Data and Analytics for 2020
https://www.gartner.com/smarterwithgartner/gartner-top-10-trends-in-data-and-analytics-for-2020

[2] New – Amazon EC2 M6i Instances Powered by the Latest-Generation Intel Xeon Scalable Processors
https://aws.amazon.com/blogs/aws/new-amazon-ec2-m6i-instances-powered-by-the-latest-generation-intel-xeon-scalable-processors/

[3] Alluxio - Data Orchestration Platform for Analytics and AI in the cloud
https://www.alluxio.io/

[4] Intel Optimized Models
https://github.com/intel/optimized-models/tree/master/pytorch/ResNet50

[5] Intel Extension for PyTorch (IPEX)
https://github.com/intel/intel-extension-for-pytorch

[6] Data Loading Library (DALI)
https://developer.nvidia.com/dali

[7] Alluxio pre-loading and caching capabilities
https://docs.alluxio.io/ee/user/stable/en/core-services/Caching.html#loading-data-into-alluxio-storage

[8] Amazon EC2 On-Demand Pricing
https://aws.amazon.com/ec2/pricing/on-demand/

[9] ImageNet
https://www.image-net.org/