Intel Cloud Optimization Modules (ICOMs)

Date: August 22, 2023

Presenter: Benjamin Consolvo

AI Solutions Engineer Manager, Intel

Agenda

- Executive Summary
- Technical Stack
- Content Package
- ICOM for AWS: XGBoost on SageMaker
- ICOM for AWS: XGBoost on Kubernetes
- ICOM for Microsoft Azure: XGBoost Pipeline on Kubernetes
- ICOM for Microsoft Azure: XGBoost Kubeflow Pipeline
- ICOM for AWS: GPT2-Small Distributed Training
- Next Steps

Executive Summary

What are Intel Cloud Optimization Modules?

The Intel Cloud Optimization Modules (ICOMs) are open-source codebases with codified Intel AI software optimizations and instructions built specifically for each Cloud Service Provider (CSP). The ICOMs are built with production AI developers in mind, leveraging popular AI frameworks within the context of cloud services.

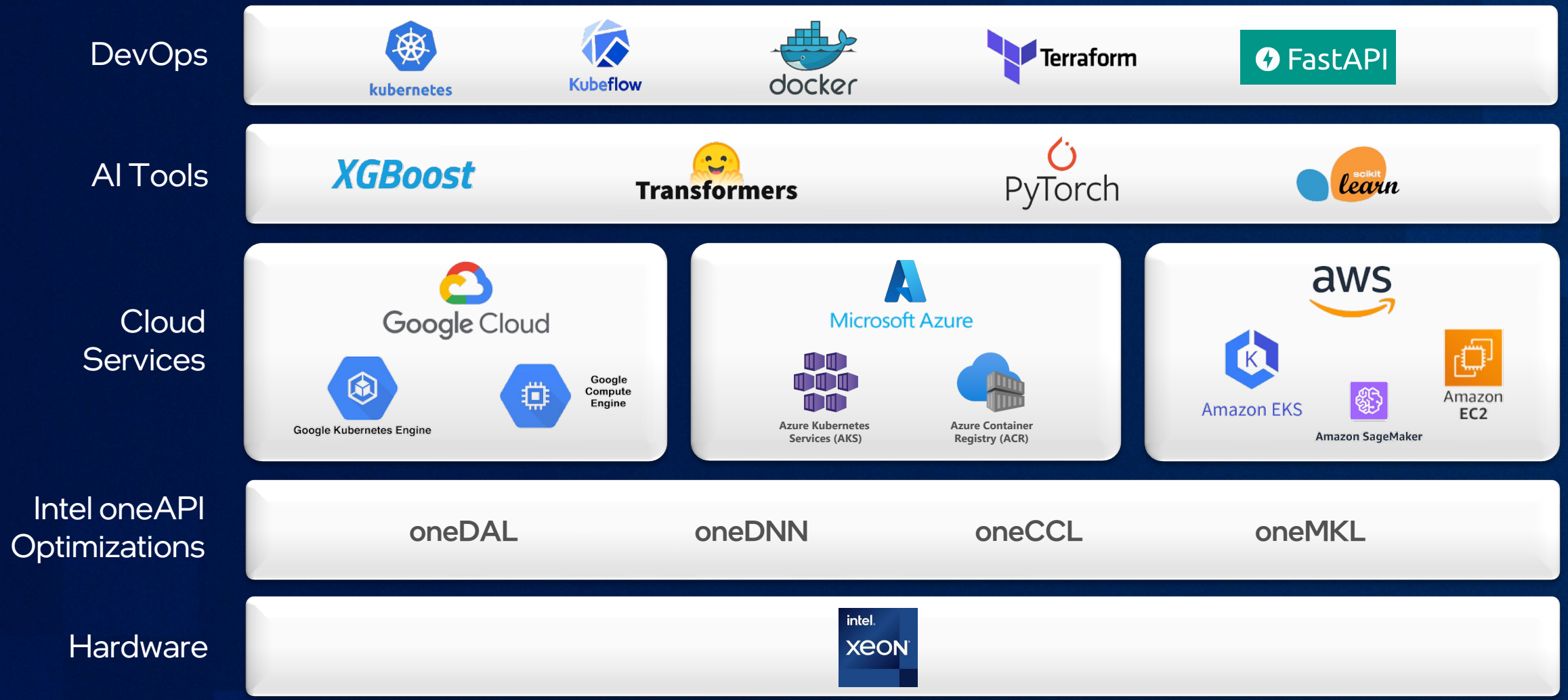
Target Audience:

Enterprise Cloud AI Developer

Landing Page:

https://www.intel.com/content/www/us/en/developer/topic-technology/cloud-optimization.html

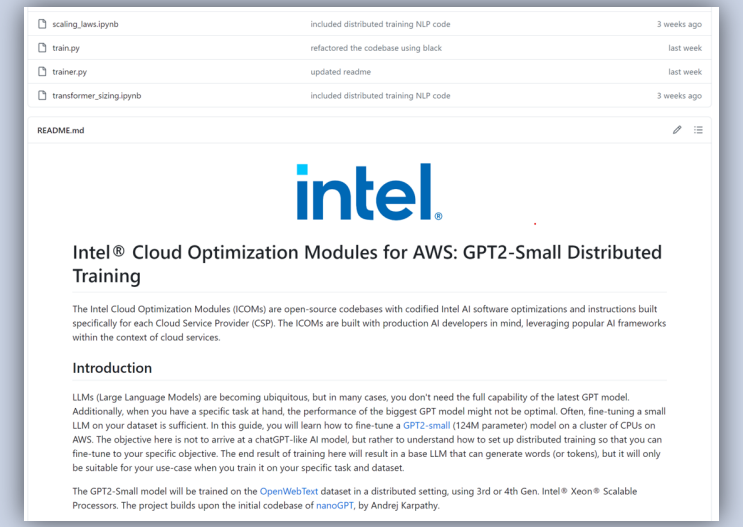
Technical Stack



Content Package

1. GitHub Repository

GitHub



2. Whitepaper

Authors: Ankur Singh, Benjamin Consolvo Date: August 9, 2023

Fine-tune GPT2-Small on the OpenWebText dataset in a distributed fashion on AWS on 3rd or 4th Generation Intel® Xeon® Scalable Processors.

3. Cheatsheet

AWS Cluster Setup

For distributed training, our example deploys a cluster consisting of three 3rd Gen. Xeon CPUs (Amazon Elastic Compute Cloud* (EC2) m6i.4xlarge instances) with an Ubuntu 22.04 AMI and 250 GB of storage. For maximum performance, we recommend using **bfloat16** precision on the 4th Gen. Xeon® CPUs (R7iz) on AWS with the deep learning acceleration engine called Intel® Advanced Matrix Extensions (AMX).

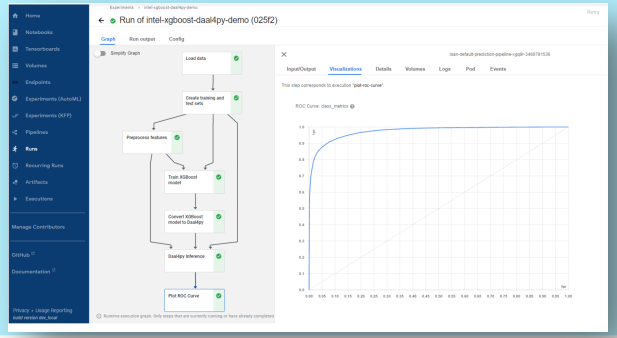
If you are using a 4th Gen. Xeon CPU, you can verify AMX is present by running:

```
lscpu | grep amx
```

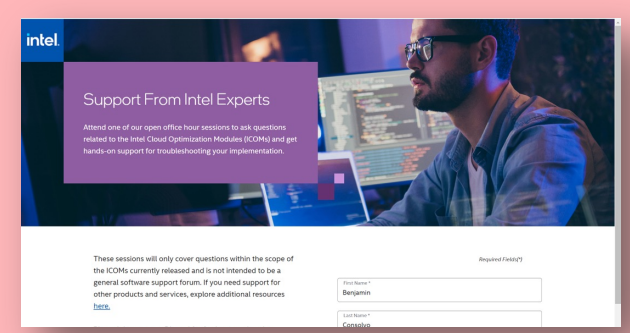
You should see the following flags:

```
amx_bf16 amx_tile amx_int8
```

4. Video Series (Coming Soon)



5. Office Hours – Registration

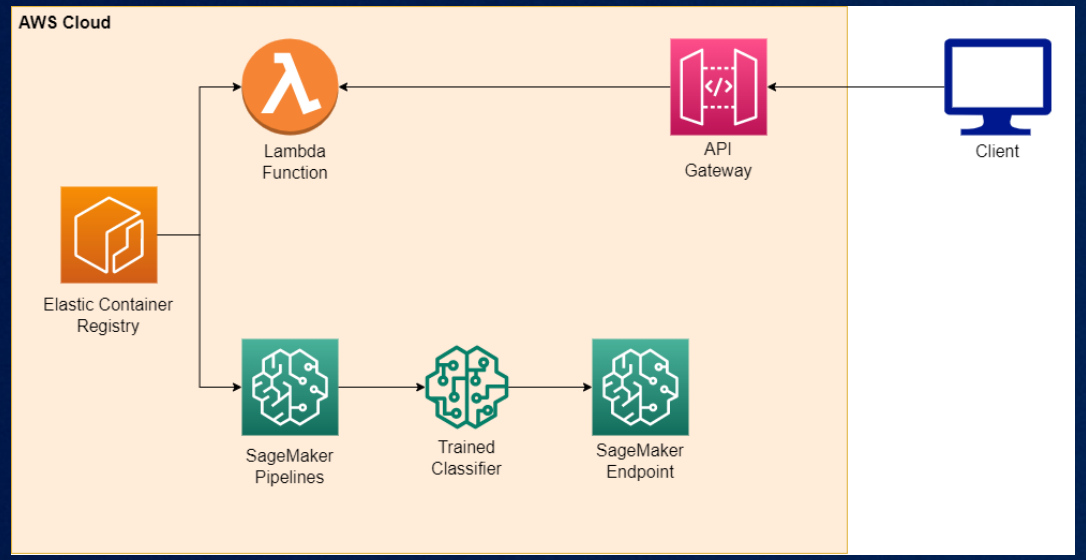


**Dependency on enterprise account managers*

ICOM for AWS: XGBoost on SageMaker

Overview: SageMaker is a fully managed machine learning service on the AWS cloud. The motivation behind this platform is to make it easy to build robust machine learning pipelines on top of managed AWS cloud services. You can learn how to inject your custom training and inference code into a prebuilt SageMaker pipeline. This module enables you to use Intel® AI Analytics Toolkit accelerated software in SageMaker pipelines.

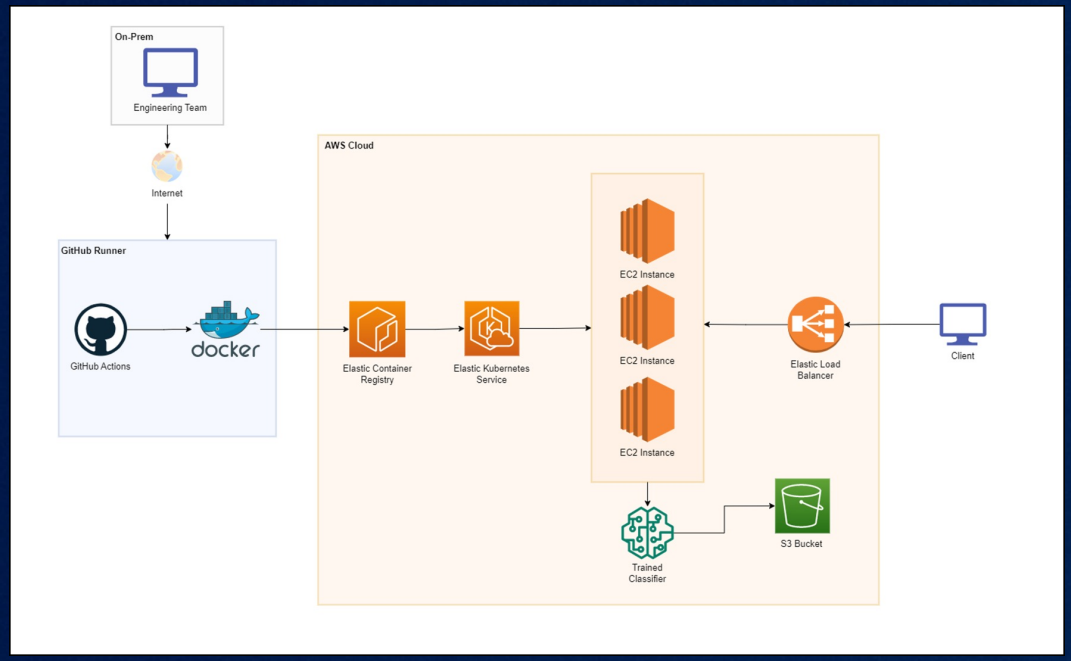
Open-Source Implementation: https://github.com/intel/intel-cloud-optimizations-aws



ICOM for AWS: XGBoost on Kubernetes

Overview: Build and deploy ML applications with XGBoost on AWS with Kubernetes with built-in Intel AI optimizations. We introduce the AWS services that we will use in the process, including Amazon Elastic Kubernetes Service (EKS), Amazon Elastic Container Registry (ECR), Amazon Elastic Compute Cloud (EC2), and Elastic Load Balancer (ELB).

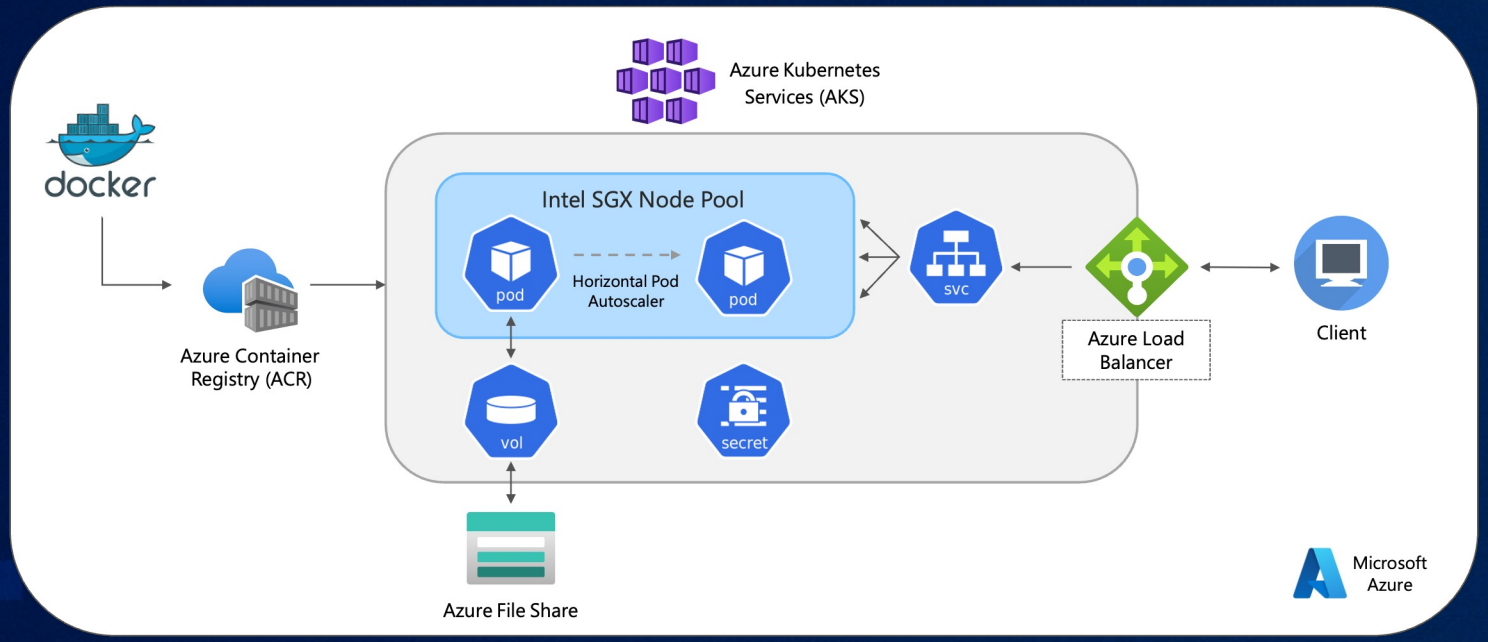
Open-Source Implementation: https://github.com/intel/intel-cloud-optimizations-aws



ICOM for Microsoft Azure: XGBoost Pipeline on Kubernetes

Overview: Build and deploy highly available and scalable AI applications on Microsoft Azure with Kubernetes. The machine learning component of the module focuses on predicting the probability of a loan default using Intel® Optimization for XGBoost* and Intel® oneAPI Data Analytics Library (oneDAL) to accelerate model training and inference. We also demonstrate how to use incremental training of the XGBoost model as new data becomes available.

Open-Source Implementation: https://github.com/intel/intel-cloud-optimizations-azure



ICOM for Microsoft Azure: XGBoost Kubeflow Pipeline

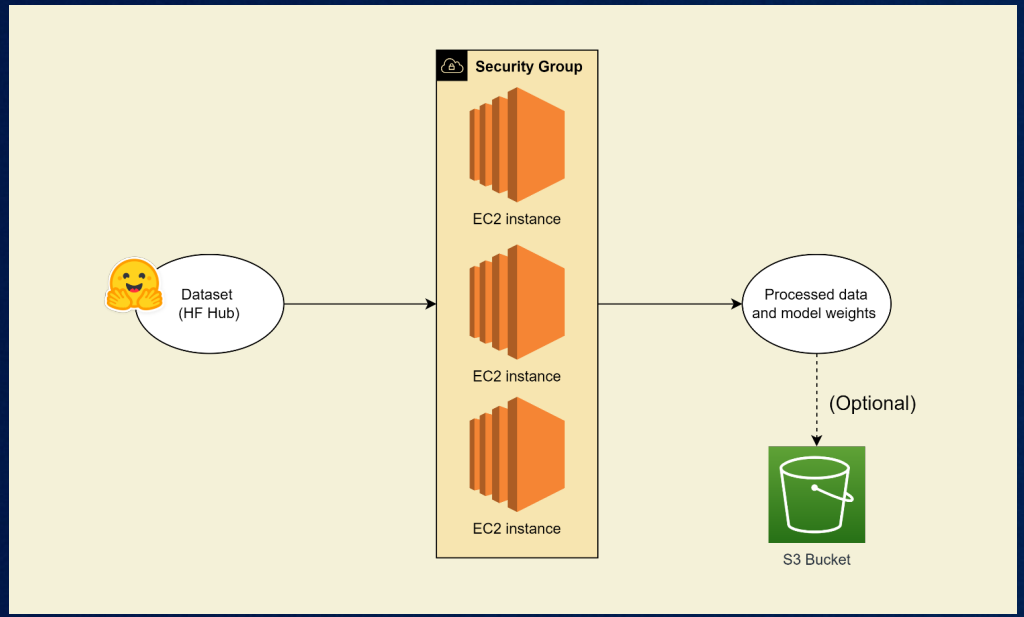
Overview: Build and deploy accelerated AI applications on Kubeflow. The module is designed to maximize the performance and productivity of XGBoost with a loan default prediction problem. This set of reference architectures for Microsoft Azure also takes advantage of secure and confidential computing virtual machines leveraging Intel® Software Guard Extensions (Intel® SGX) on the Azure cloud.

Open-Source Implementation: https://github.com/intel/intel-cloud-optimizations-azure

ICOM for AWS: GPT2-Small Distributed Training

Overview: Learn how to fine-tune a GPT2-small (124M parameter) model on a cluster of CPUs on AWS. The objective here is not to arrive at a chatGPT-like AI model, but rather to understand how to set up distributed training so that you can fine-tune to your specific objective. The result will be a base LLM that can generate words (or tokens), but it will only be suitable for your use-case when you train it on your specific task and dataset. The GPT2-Small model is trained on the OpenWebText dataset in a distributed setting, using 3rd or 4th Gen. Intel® Xeon® Scalable Processors. The project builds upon the initial codebase of nanoGPT, by Andrej Karpathy.

Open-Source Implementation: https://github.com/intel/intel-cloud-optimizations-aws



And more!

Intel® Cloud Optimization Modules

This set of cloud-native open source reference architectures helps developers build and deploy Intel®-optimized AI solutions on leading cloud providers, including Amazon Web Services (AWS)*, Microsoft Azure*, and Google Cloud Platform* service.

Each module or reference architecture includes a complete instruction set, all source code published on GitHub*, and a video walk-through.

Third-Party Cloud Optimization Modules

Amazon Web Services (AWS*) | Microsoft Azure* | Google Cloud Platform* Service

Amazon SageMaker*

Learn how to build and register images to Amazon Elastic Container Register (Amazon ECR) for the xgboost-daal4py app and the Lambda inference handler.

Amazon Elastic Kubernetes* Services

Learn how to build accelerated machine learning pipelines using Amazon Elastic Kubernetes Services* (Amazon EKS) with Intel® Optimization for XGBoost*.

Amazon Elastic Container Service (ECS)

This module creates an ECS cluster using the latest Intel® architecture available for the AWS ECS Service.

Amazon EKS Module with Terraform

This repository provides an example to create an Amazon EKS cluster optimized on 3rd generation Intel® Xeon® Scalable processors (formerly code named Ice Lake). The example will be creating an Amazon EKS cluster with an Amazon EKS managed node group.

Amazon ElastiCache* for Redis

This module creates an Amazon ElastiCache* for Redis cluster based on Intel architecture and creates a new or existing VPC. This module uses the cache.r5.large by default, which is the latest Intel Xeon processor available at the time of this module publication.

Amazon Elastic Compute Cloud (EC2)

Configuration in this directory creates an AWS VM (instance). The instance is created on a 3rd generation Intel® Xeon® Scalable processor by default.

Next Steps

- Learn more about all of our Intel Cloud Optimization Modules here:
  - https://www.intel.com/content/www/us/en/developer/topic-technology/cloud-optimization.html
- Register for Office Hours here for help on your ICOM implementation:
  - https://software.seek.intel.com/SupportFromIntelExperts-Reg
- Come chat with us on our Intel DevHub Discord server to keep interacting with fellow developers:
  - https://discord.gg/rv2Gp55UJQ

- Stay connected with me on social media:
  - Benjamin Consolvo | AI Solutions Engineer Manager, Intel
  - LinkedIn: https://linkedin.com/in/bconsolvo
  - Twitter: https://twitter.com/bpconsolvo


intel