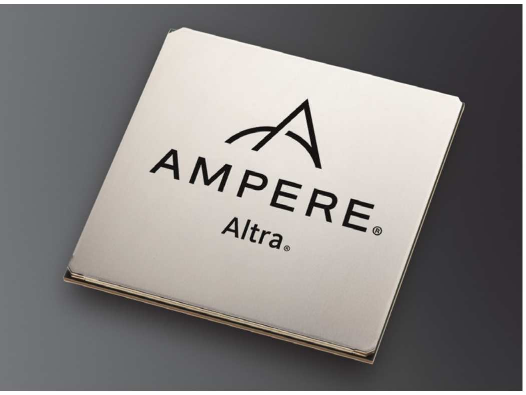


Object Detection with YOLOv8 — Running Machine Learning on GCP Tau T2A VMs

GCP Tau T2A, powered by Ampere® Altra CPU and high performance Ampere® AI inference engine, delivers best-in-class AI inference performance on standard frameworks, including PyTorch, TensorFlow, and ONNX-RT.

Ampere Altra Powered ML Inference on GCP

Ampere® Altra family of **Cloud-Native Processors** satisfies the performance requirements of widely used machine learning (ML) workloads while **providing the best price-performance**. This demo consists of multiple streams of video sources detecting still and moving objects such as pedestrians, laptop, chair, cup, and so on, using the popular YOLOv8 model.

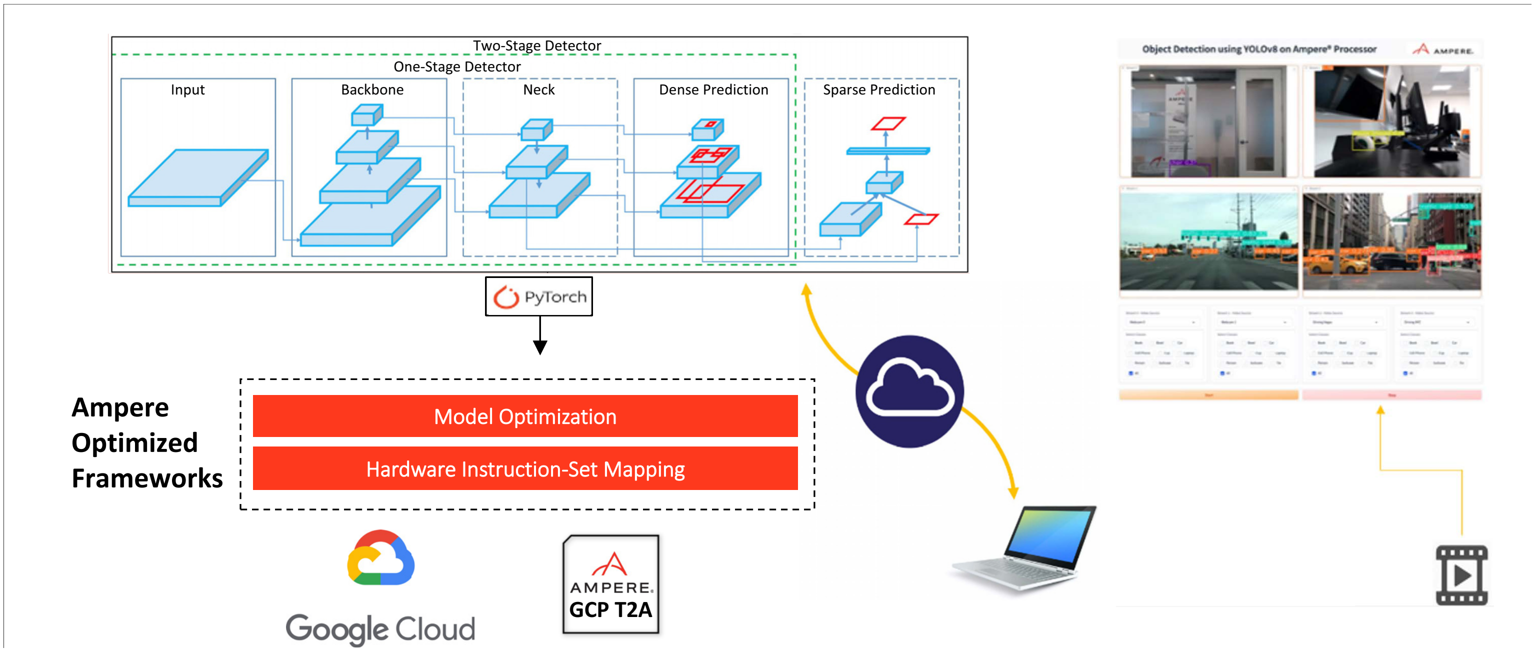
Setup

Deployment of the open-source **computer vision** object detection AI model YOLOv8 with **Ampere® Optimized PyTorch** running on Ampere Altra Max. The chosen model, YOLOv8, is a widely used algorithm for computer vision applications where both throughput and latency are critical. Implementation and performance details for the YOLOv8 model developed and released by Ultralytics can be found here.

Key Benefits Demonstrated

- Meets or exceeds the necessary **low latency** requirements for real-time ML object detection applications.
- Delivers the best **price-performance** in CPU-only AI inference in both cloud and edge deployment scenarios.
- The YOLOv8 model can be downloaded from Ampere® AI Model Library (AML) and used as is without any modifications.
- Ampere Altra processor can **easily be scaled** and **dynamically provisioned** based on the performance requirements of the user's application such as target frame rate, number of video channels, etc.

Figure 1: YOLOv8 Demo Runs on GCP Tau T2A Instance with Ampere Altra



Real-time Object Detection and Classification

This demo performs object detection and classification with a pre-trained YOLOv8 model. It processes images and videos from an incoming real-time video streaming from video files. It runs on a **GCP T2A VM** at real-time **performance level**. The performance can be scaled depending on application requirements by allocating the number of vCPUs to meet the desired price-performance target.

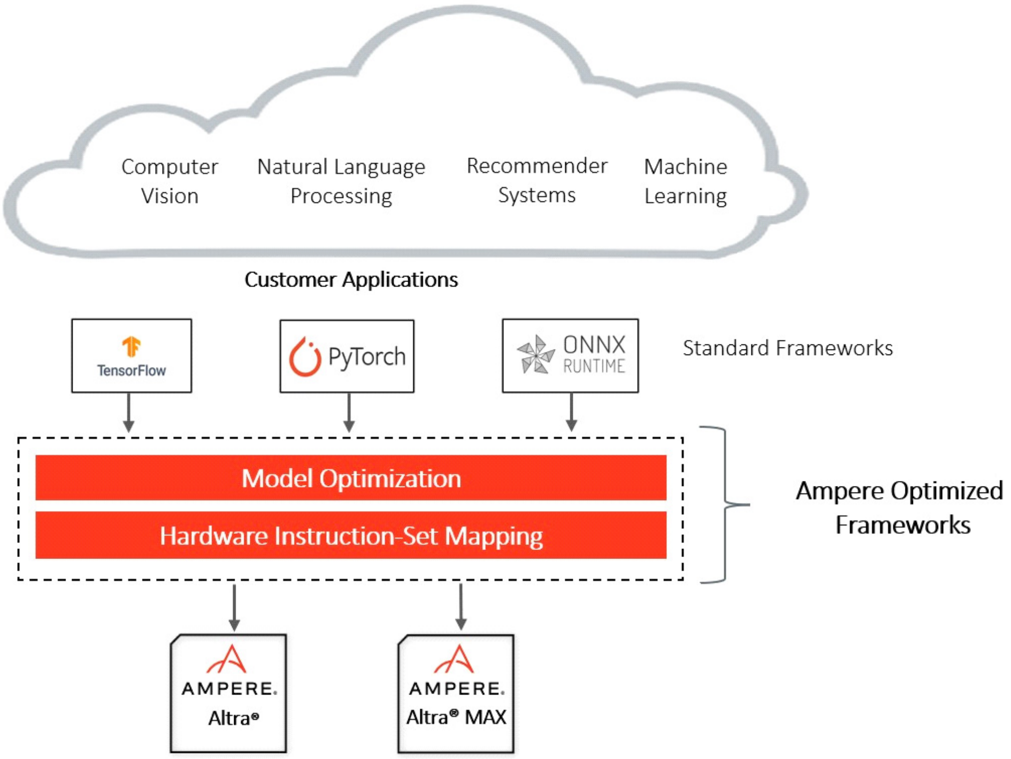
The same workload also runs on x86 for comparison purposes. We demonstrate that the **Ampere Altra family of cloud-native processors consistently outperforms x86 platforms**.

Resources

The YOLOv8 model can be accessed from the Ampere AI Model Library. The docker image of Ampere Optimized PyTorch is available in the downloads section of Ampere AI Solutions web page. Other Ampere® Optimized Frameworks can also be accessed from the same location.

Ampere Optimized TensorFlow, PyTorch, ONNX-RT can also be downloaded and installed free of charge on any edge workstation or server through Ampere AI Solutions web page.

Figure 2. Integration of Ampere Optimized Frameworks with Ampere Altra Cloud Native Processors



Ampere Computing reserves the right to make changes to its products, its datasheets, or related documentation, without notice and warrants its products solely pursuant to its terms and conditions of sale, only to substantially comply with the latest available datasheet.

Ampere Computing® / 4655 Great America Parkway, Suite 601 / Santa Clara, CA 95054 / www.amperecomputing.com