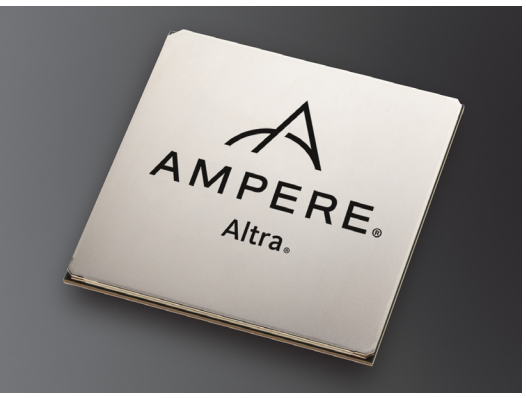


Object Detection with YOLOv5 — Running Machine Learning Workloads on Ampere® Altra®

Ampere Altra family of processors, with 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

Ampere Altra family of **cloud native processors** meets the needs of widely used ML workloads while **optimizing the total cost of ownership**. This demo consists of a video analytics use case for detecting still and moving objects, like vehicles, pedestrians, and traffic signs, using the popular YOLOv5 model.

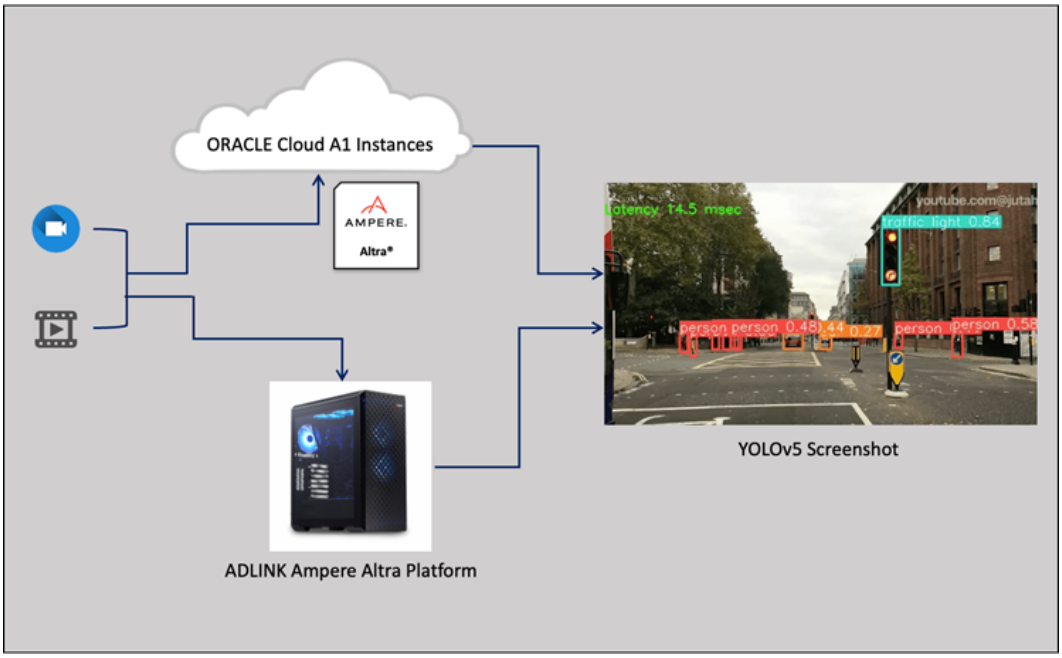
Setup

Deployment is done using an open source **computer vision** object detection AI model, YOLOv5 with **Ampere® Optimized PyTorch**, running on Ampere Altra. The chosen model, YOLOv5, is a widely used algorithm for real-time applications where both throughput and latency are critical. Implementation and performance details for the YOLOv5 model developed and released by Ultralytics can be found at https://github.com/ultralytics/yolov5.

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 inferencing in both cloud and edge deployment scenarios.
- The YOLOv5 model can be downloaded from Ampere AI Model Library (AML) and used as is without any modifications.
- Ampere Altra processor cores 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, and so on.

Figure 1: Ampere Altra YOLOv5 demo can be run on cloud instances or on a local Ampere Altra computer



Low Latency Demo – Real-time Object Detection and Classification

This demo performs object detection and classification inference with a pretrained YOLOv5 model. It processes images and video files from a web app or incoming real-time video streaming from a camera. The demo can be run either on a **local Ampere Altra computer** or on **OCI Ampere A1 instances with Ampere Altra** at real-time **performance levels (30 fps)**. The performance can be scaled up or down depending on application requirements by assigning processor cores and adding or removing CPU instances to meet a desired price-performance target.

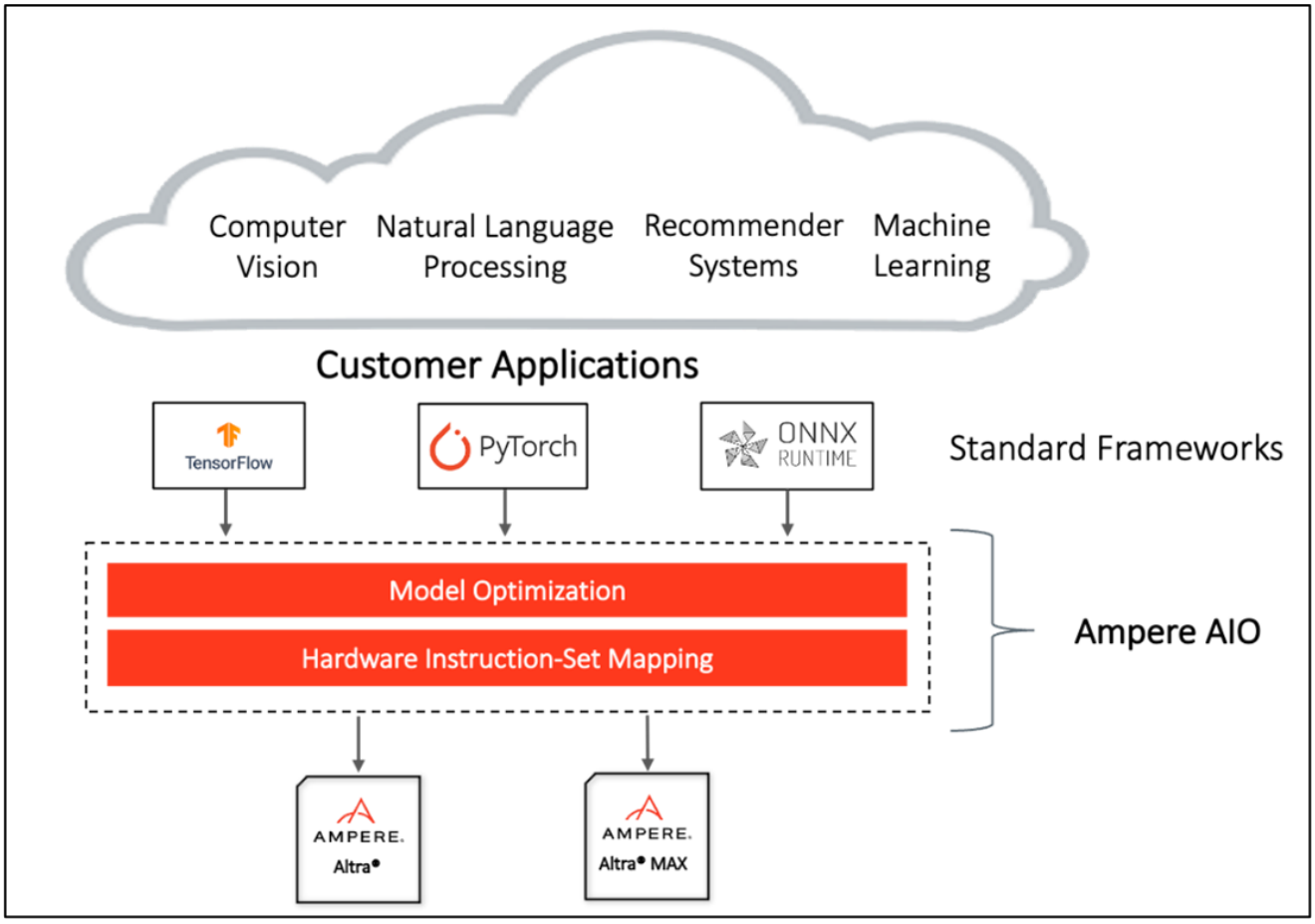
The same workload also runs on x86 for comparison purposes. We demonstrate that the **Ampere Altra family of cloud native processors consistently outperforms x86 platforms**.

Resources

YOLOv5 demo github.com/AmpereComputingAI/yolov5-demo supports AI inference on Ampere Altra processors and Ampere Altra with NVIDIA GPU. The YOLOv5 model can be accessed from Ampere AI Model Library. The docker image of Ampere Optimized PyTorch is available in the downloads section on the Ampere AI Solutions website. Other Ampere optimized frameworks can be accessed from the same location. Ampere A1 instances can be accessed and evaluated via Oracle's free tier. Additional information on Ampere A1 Compute and Ampere Optimized PyTorch is available on OCI marketplace.

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

Figure 2: Ampere Altra instances are available from a variety of cloud providers, including Oracle Cloud Infrastructure, Google Cloud, Microsoft Azure, Tencent, Equinix, and others



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