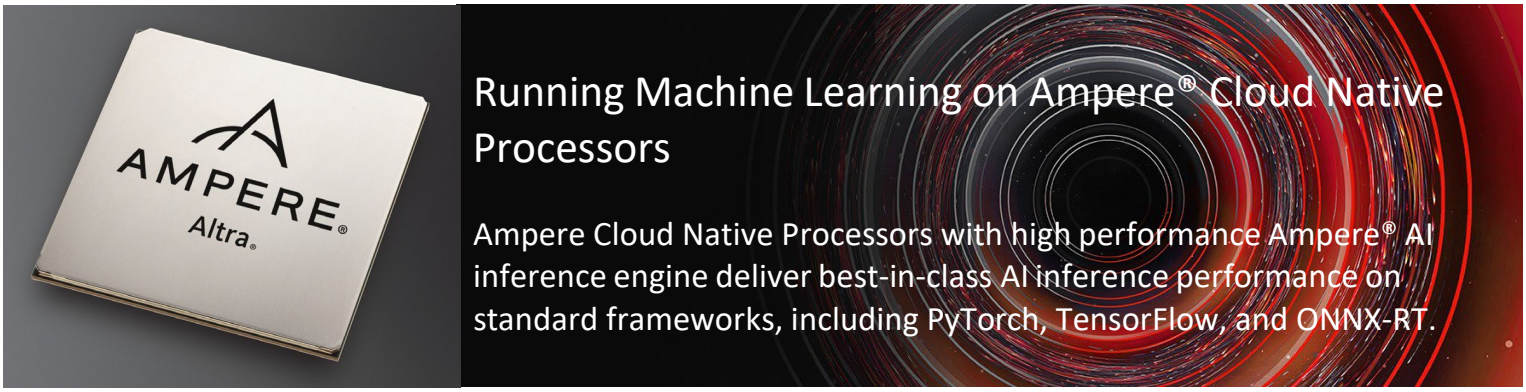
YOLOv8 Segmentation and Pose Estimation



Ampere AI Powered ML Inference

Ampere® **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. Semantic segmentation and pose estimation are performed on objects and humans with YOLOv8.

Setup

Deployment of the open-source **computer vision** YOLOv8 segmentation and pose estimation AI models with **Ampere® Optimized PyTorch** running on Ampere Altra Max / AmpereOne. 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: https://github.com/ultralytics/ultralytics/tree/main/ultralytics/yolo/v8.

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® 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.

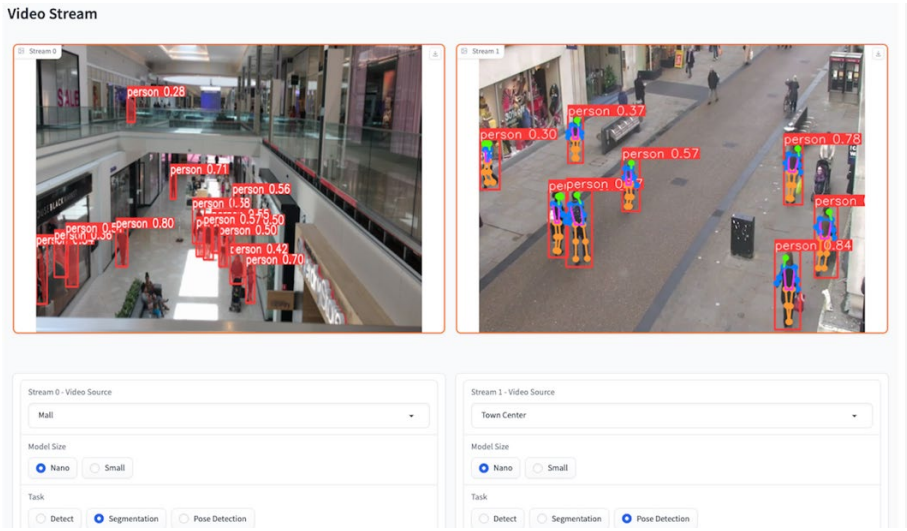


Figure 1: YOLOv8 demo runs on Ampere Altra Max / AmpereOne

Real-time Segmentation and Pose Estimation

This demo performs segmentation on common objects or human pose estimation with a pre-trained YOLOv8 model. It processes images and videos from an incoming real-time video streaming from video files. The demo runs on an **Ampere Altra Max/AmpereOne server** 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 **Ampere Cloud-Native Processors consistently outperform x86 platforms.**

Resources

The YOLOv8 models can be accessed from the Ampere 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.

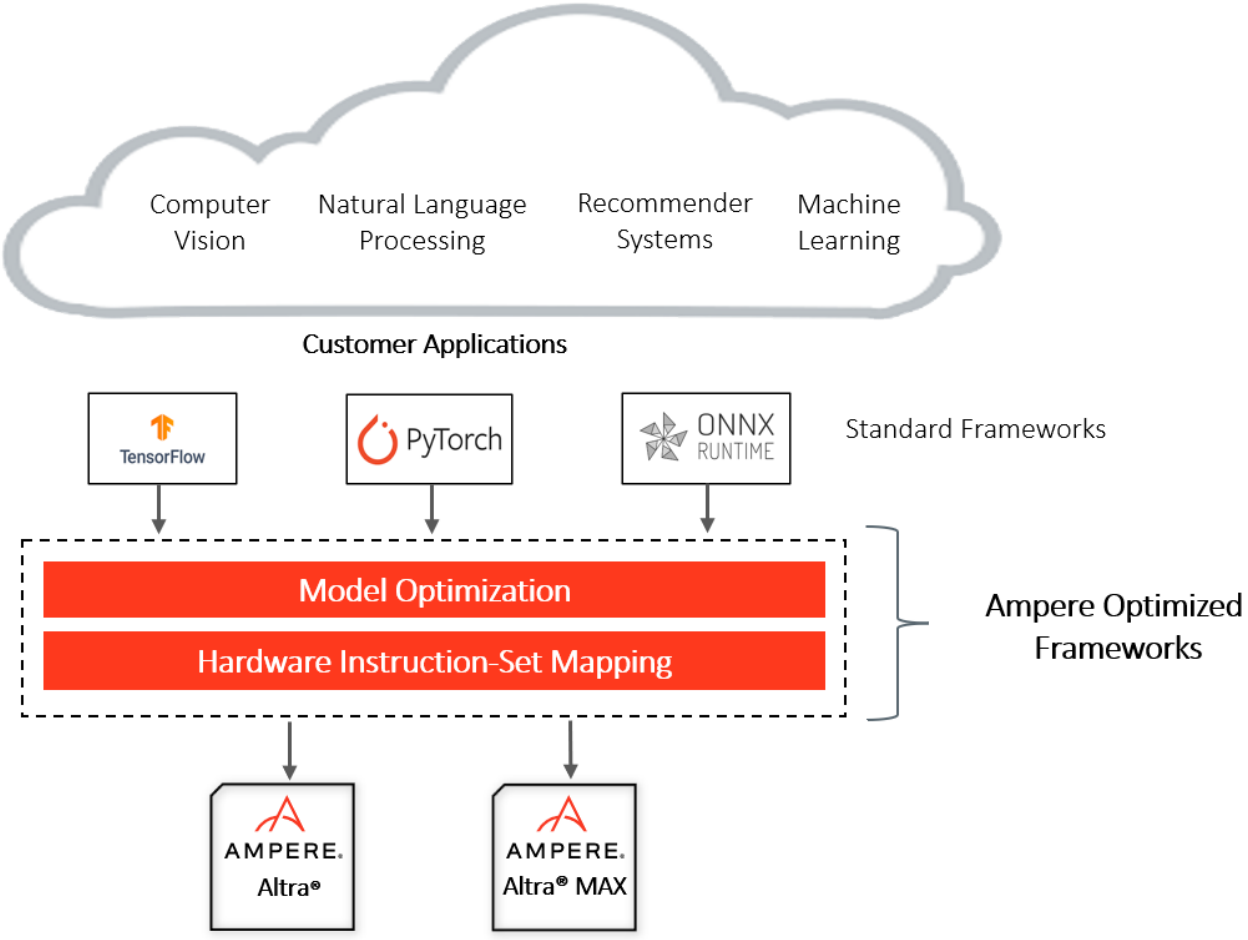


Figure 2: The integration of Ampere Optimized Frameworks with Ampere Altra Cloud Native Processors