



Computer Vision for Smart Cities and Retail

Ampere® Cloud Native Processors with Ampere® Optimized AI Frameworks, deliver best GPU-Free AI inference performance for applications developed in PyTorch, TensorFlow, and ONNX-RT.



Ampere Altra Powered ML Inference

Ampere® Cloud-Native Processors satisfies the performance requirements of widely used machine learning (ML) workloads while providing the **best price-performance and optimized power draw**. This demo consists of multiple streams of video sources. Semantic segmentation and pose estimation are performed on objects and humans with YOLOv8.

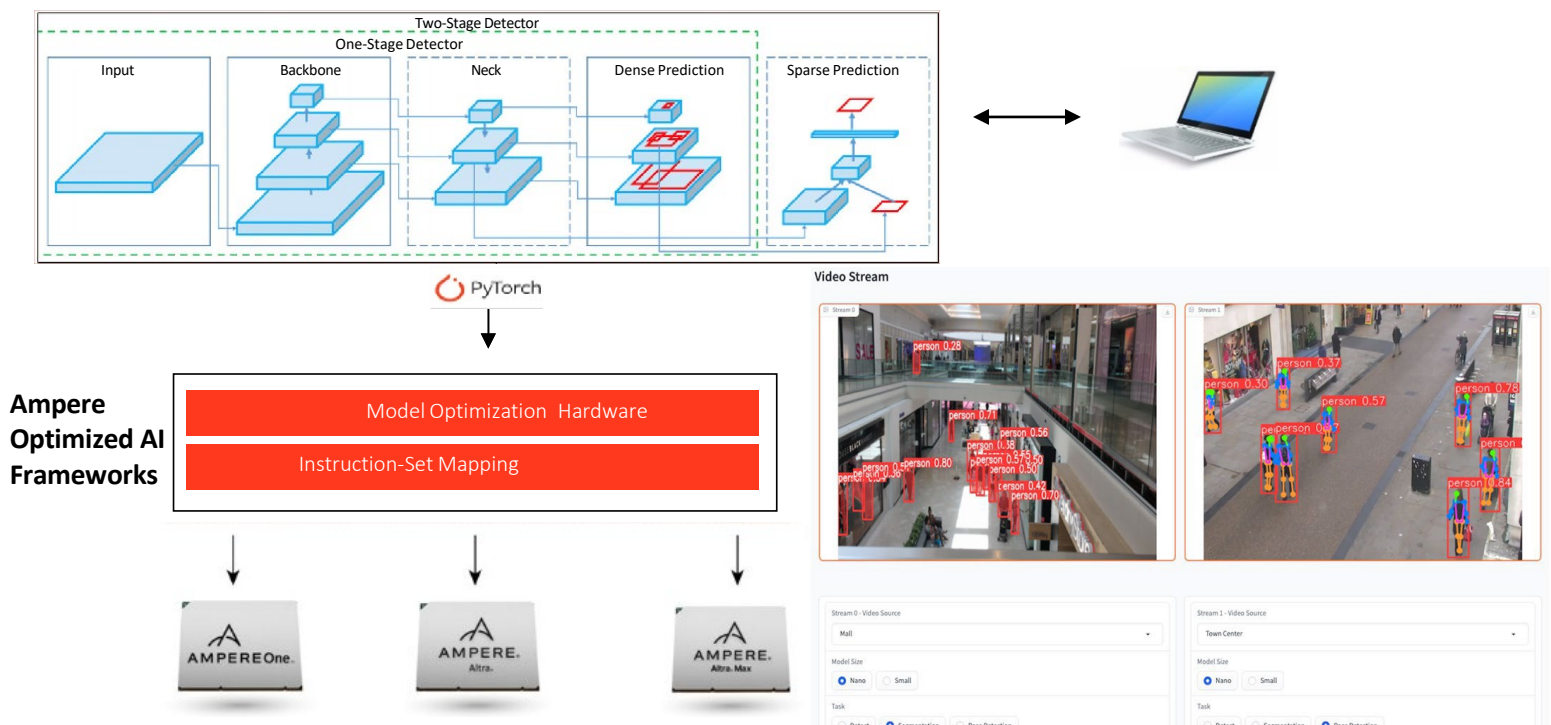
Setup

Deployment of the open-source **computer vision** YOLOv8 AI model 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: [ultralytics/yolo/v8](https://ultralytics.com/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 GPU-Free 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.

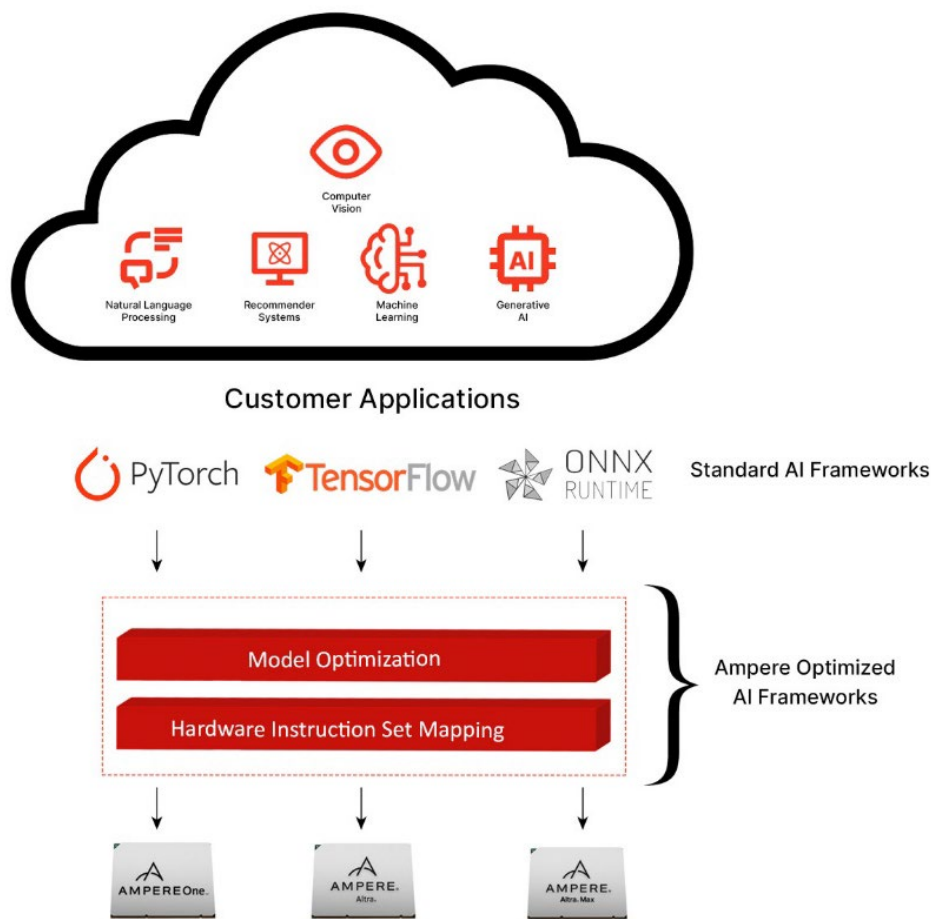
The same workload also runs on x86 for comparison purposes. We demonstrate that **Ampere Cloud-Native Processors consistently outperform x86 platforms**.

Resources

The YOLOv8 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 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, Ampere Computing, the Ampere Computing and 'A' logos, and Altra are registered trademarks of Ampere Computing.

Arm is a registered trademark of Arm Limited (or its subsidiaries) in the US and/or elsewhere. All other trademarks are the property of their respective holders.

Copyright © 2024 Ampere Computing. All Rights Reserved.

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