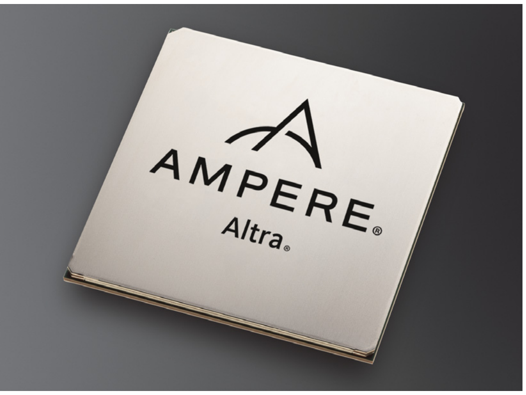


Automatic Speech Recognition (ASR) with Whisper Model — Running Machine Learning on GCP Tau T2A VMs

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** meets the needs of widely used machine learning (ML) workloads while **providing the best price-performance**. This demo performs live transcription of audio files into text, using the state-of-the-art Open AI Whisper model. Whisper offers the best-in-class accuracy and capabilities for Automatic Speech Recognition (ASR) use cases.

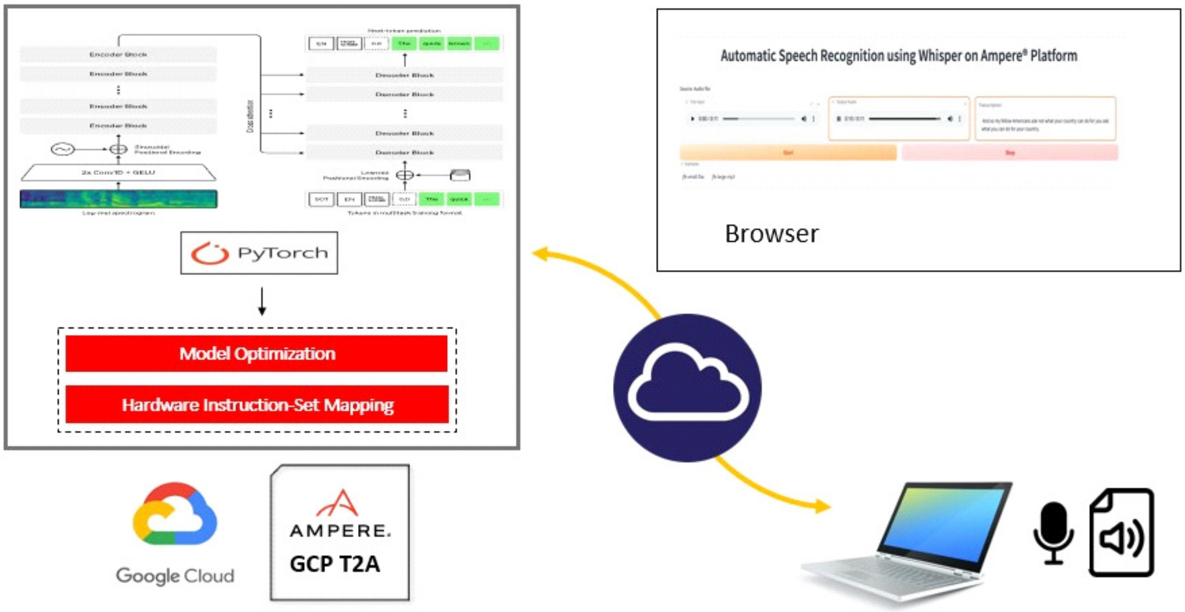
Setup

The setup includes the deployment of the open-source ASR AI model Whisper, with **Ampere® Optimized PyTorch** running on GCP T2A Ampere Altra. The chosen model, Whisper Medium, is a widely used algorithm for ASR applications where both throughput and latency are critical. Implementation and performance details for the Whisper model by Open AI can be found at https://github.com/openai/whisper.

Key Benefits Demonstrated

- Meets or exceeds the necessary **low latency** requirements for real-time ML Automatic Speech Recognition (ASR) applications.
- Delivers the best **price-performance** in CPU-only AI inference in both cloud and edge deployment scenarios.
- The Whisper 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.

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



Real-time Automatic Speech Recognition (ASR)

This demo performs ASR inference with a pre-trained Whisper model. It processes audio streams read from audio files. The demo runs on a **GCP T2A VM** at real-time **performance level** (the rate of speech-to-text processing is faster than the rate of the audio stream). The performance can be scaled depending on application requirements by allocating the number of vCPUs to meet the desired price-performance target.

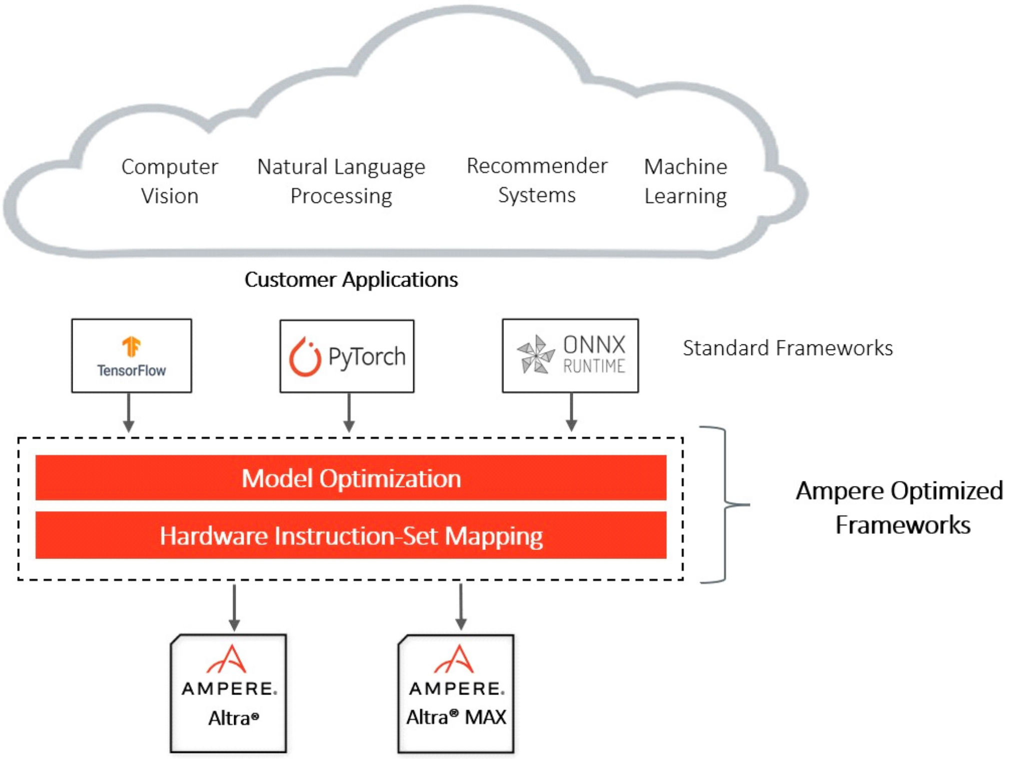
The same workload also runs on x86 for comparison purposes. We demonstrate that the **Ampere Altra family of cloud-native processors consistently outperforms x86 platforms**.

Resources

The Whisper 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, and 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.