Automatic Speech Recognition (ASR) with Whisper Model

Running Machine Learning on Ampere® Cloud Native Processors

Ampere Cloud Native Processors with high performance Ampere® AI inference engine deliver best-in-class AI inference performance on standar frameworks, including PyTorch, TensorFlow, and ONNX-RT.

Ampere AI Powered ML Inference

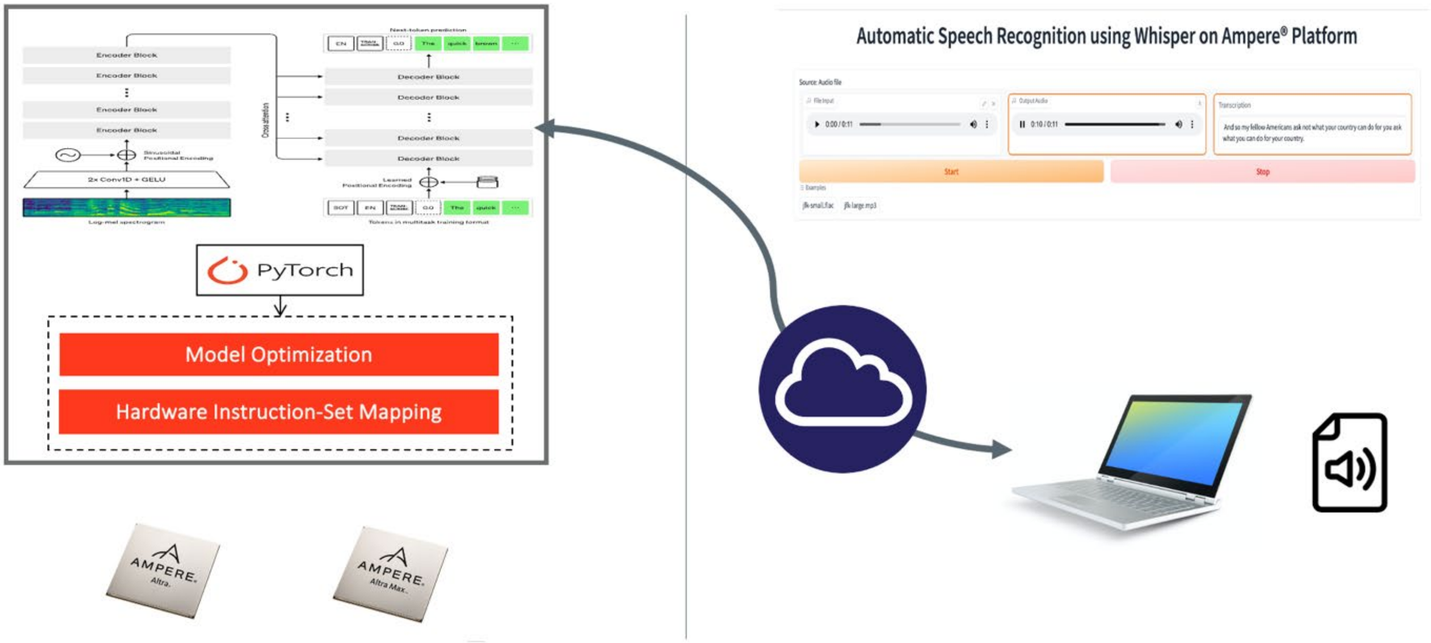
Ampere **Cloud-Native Processors** meet the needs of widely used machine learning (ML) workloads while **providing the best price-performance**. This demo performs live transcription of audio file into text, using the state-of-the-art Open AI Whisper model. Whisper offers the best-in-class accuracy and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

Setup

Deployment of the open-source ASR AI model Whisper with **Ampere® Optimized PyTorch** running on Ampere® Altra / Ampere® Altra Max / AmpereOne®. 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 here: 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® 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.



And so my fellow Americans ask not what your country can do for you ask what you can do for your country.

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

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 Ampere Altra / Altra Max / AmpereOne servers 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 **Ampere Cloud-Native Processors consistently outperform x86 platforms.**

Resources

The Whisper model 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 AI 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.

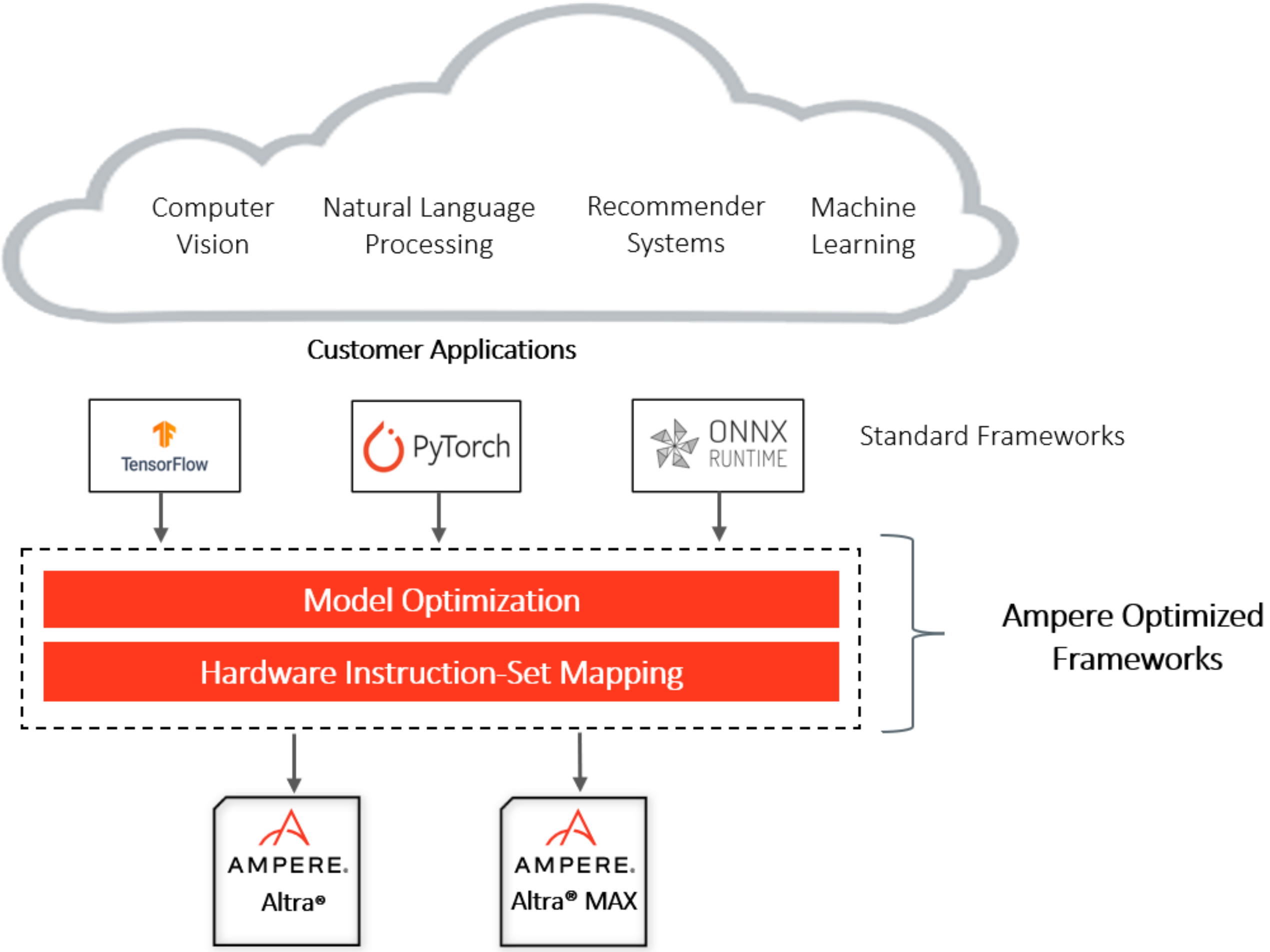


Figure 2: The integration of Ampere Optimized Frameworks with Ampere Altra Cloud Native Processors