



# Automatic Speech Recognition (ASR) with Whisper Model on GCP T2A VMs

GCP T2A VMs, powered by Ampere® Altra CPU and high performance Ampere® AI inference engine, deliver best-in-class GPU-Free AI inference performance on standard AI frameworks, including PyTorch, TensorFlow, and ONNX-RT.

## Ampere Altra Powered ML Inference On GCP

Ampere® Altra family of **Cloud Native Processors** meet the needs of widely used machine learning (ML) workloads while **providing the best price-performance and optimized power draw**. 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 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 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 GPU-Free AI inference in cloud 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.

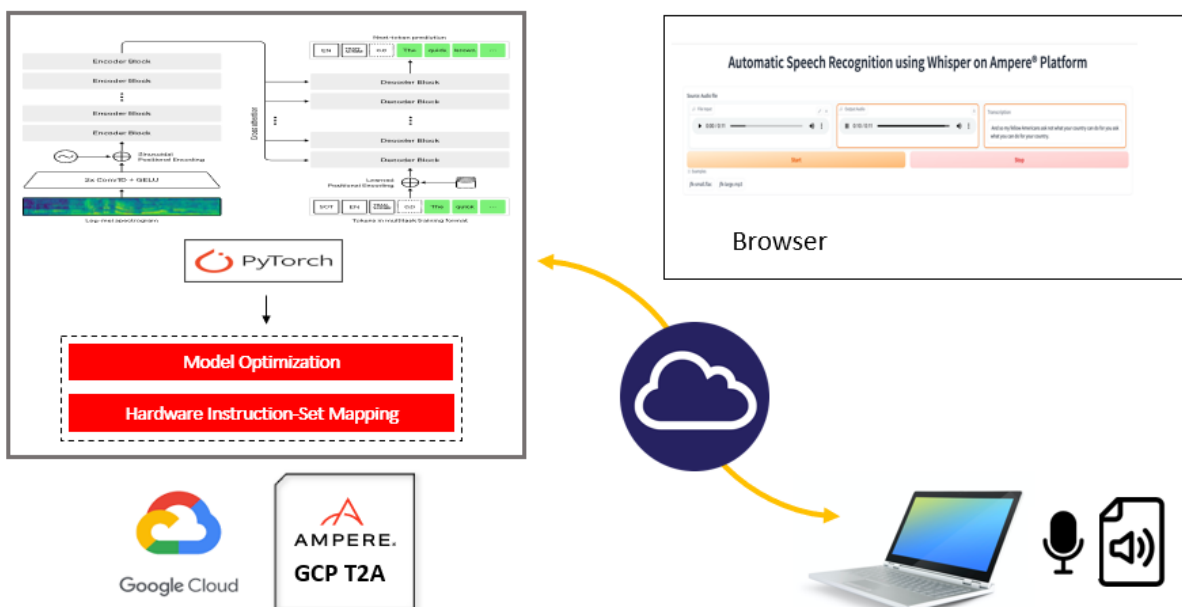


Figure 1: Whisper demo runs on GCP T2A Instance with Ampere Altra

## Real-time Automatic Speech Recognition (ASR)

This demo performs AI 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.

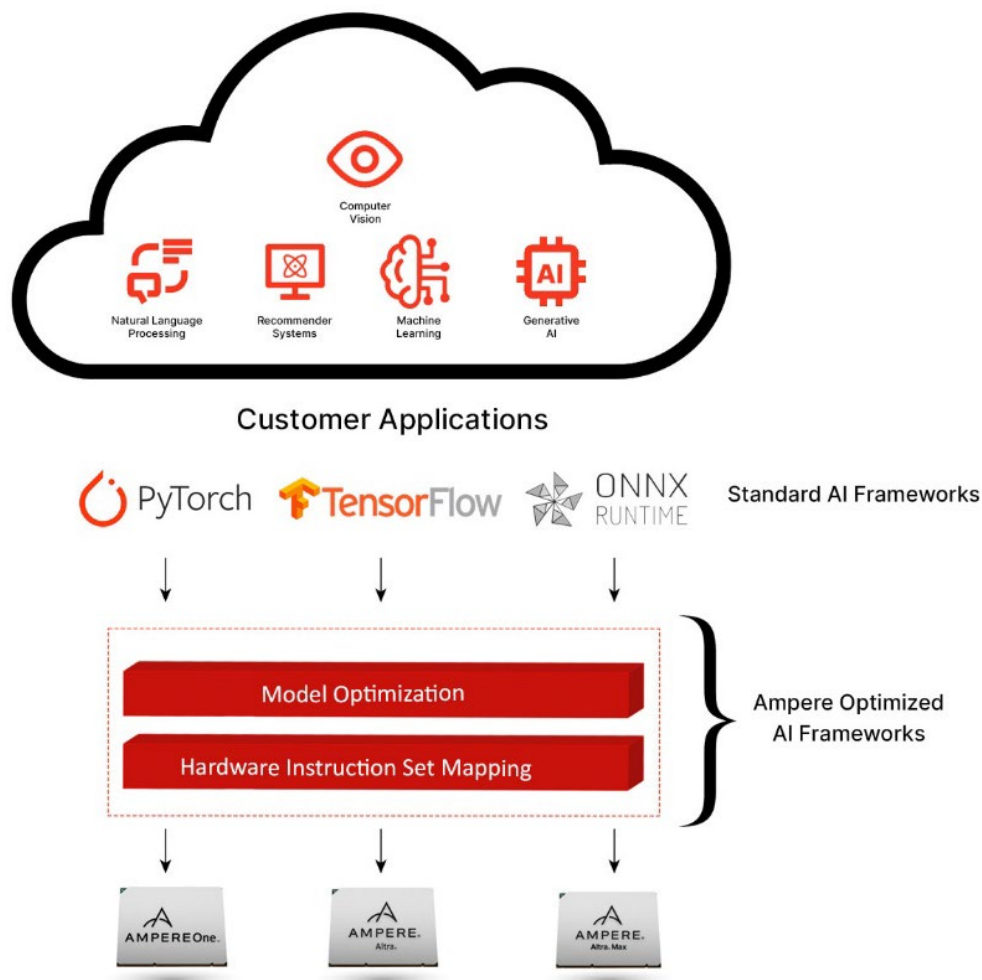
The same workload also runs on x86 for comparison purposes. We demonstrate that the **Ampere Altra family of Cloud Native Processors** consistently outperforms x86 platforms.

## Resources

The Whisper model can be accessed from the [Ampere AI Model Library](#). Ampere Optimized Pytorch, and other Ampere Optimized AI Frameworks, can be accessed directly from [Google Cloud Marketplace](#).

The docker images are also available in the downloads section of [Ampere AI Solutions web page](#). All software is available free of charge and runs straight out-of-the-box with no additional coding required.

**Figure 2:** The integration of Ampere Optimized AI 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.

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](http://www.amperecomputing.com)**