Voice to Text Chat with Lampi Fine-Tuned Llama 2 7B and Whisper Models

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 Generative AI Inference

Ampere® **Cloud Native Processors** satisfy the performance requirements of widely used generative AI models such as Whisper, originally developed by Open AI and used especially for ASR (automatic speech recognition), and Meta's Llama 2, especially for the smaller versions of the Llama model. Ampere CPUs provide the **best price-performance and optimized power draw** for all kinds of ML deployments.

About Lampi

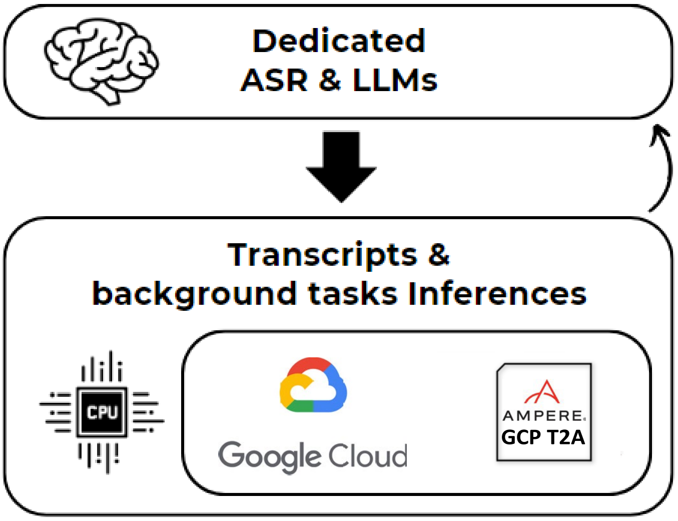
Lampi provides AI workflow copilot oriented to business use cases - agnostic and fully modular adapted on the fly to client's business needs. Lampi deploys multimodal AI on the fly allowing for a consistency in results to reach the desired quality and compliance levels.

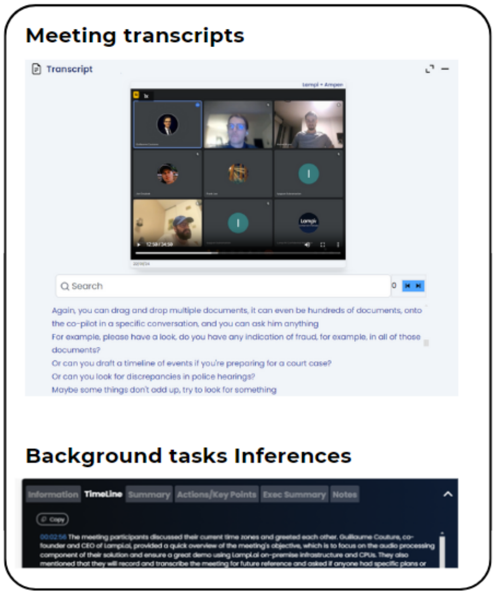
Setup

Deployment of the dynamic Lampi fine-tuned versions of the open-source generative AI **Llama 2 7B** and **Whisper** models running on Ampere-based **GCP T2A** cloud instance with **Ampere® Optimized AI Frameworks**. The demo showcases real-time chatbot interaction with the training set of Ampere AI and energy efficiency content chosen to inform the audience visiting the Ampere booth at Google Cloud Next.

Key Benefits Demonstrated

- Meets or exceeds the necessary **low latency** requirements for real-time generative AI chatbot interaction. (voice-to-text, voice-to-voice, text-to-voice, and text-to-text)
- Ampere delivers best throughput for Whisper model inference outcompeting some of the most commonly used GPUs.
- Delivers the best **price-performance** in GPU-Free AI inference in cloud deployment scenarios.
- Ampere Optimized AI Frameworks work right out-of-the-box and can be downloaded free of charge directly from Google Cloud Marketplace or, as docker images, from the Ampere AI solutions page.
- Ampere Cloud Native Processors enable **easy scaling** and can be **dynamically provisioned** based on the performance requirements of the user's applications.





Meeting transcripts

Again, you can drag and drop multiple documents, it can even be hundreds of documents, onto the co-pilot in a specific conversation, and you can ask him anything
For example, please have a look, do you have any indication of fraud, for example, in all of those documents?
Or can you draft a timeline of events if you're preparing for a court case?
Or can you look for discrepancies in police hearings?
Maybe some things don't add up, try to look for something

Background tasks Inferences

00:02:56 The meeting participants discussed their current time zones and greeted each other. Guillaume Couture, co-founder and CEO of Lampi.ai, provided a quick overview of the meeting's objective, which is to focus on the audio processing component of their solution and ensure a great demo using Lampi.ai on-premise infrastructure and CPUs. They also mentioned that they will record and transcribe the meeting for future reference and asked if anyone had specific plans or

Real-time Voice-to-Text Chat

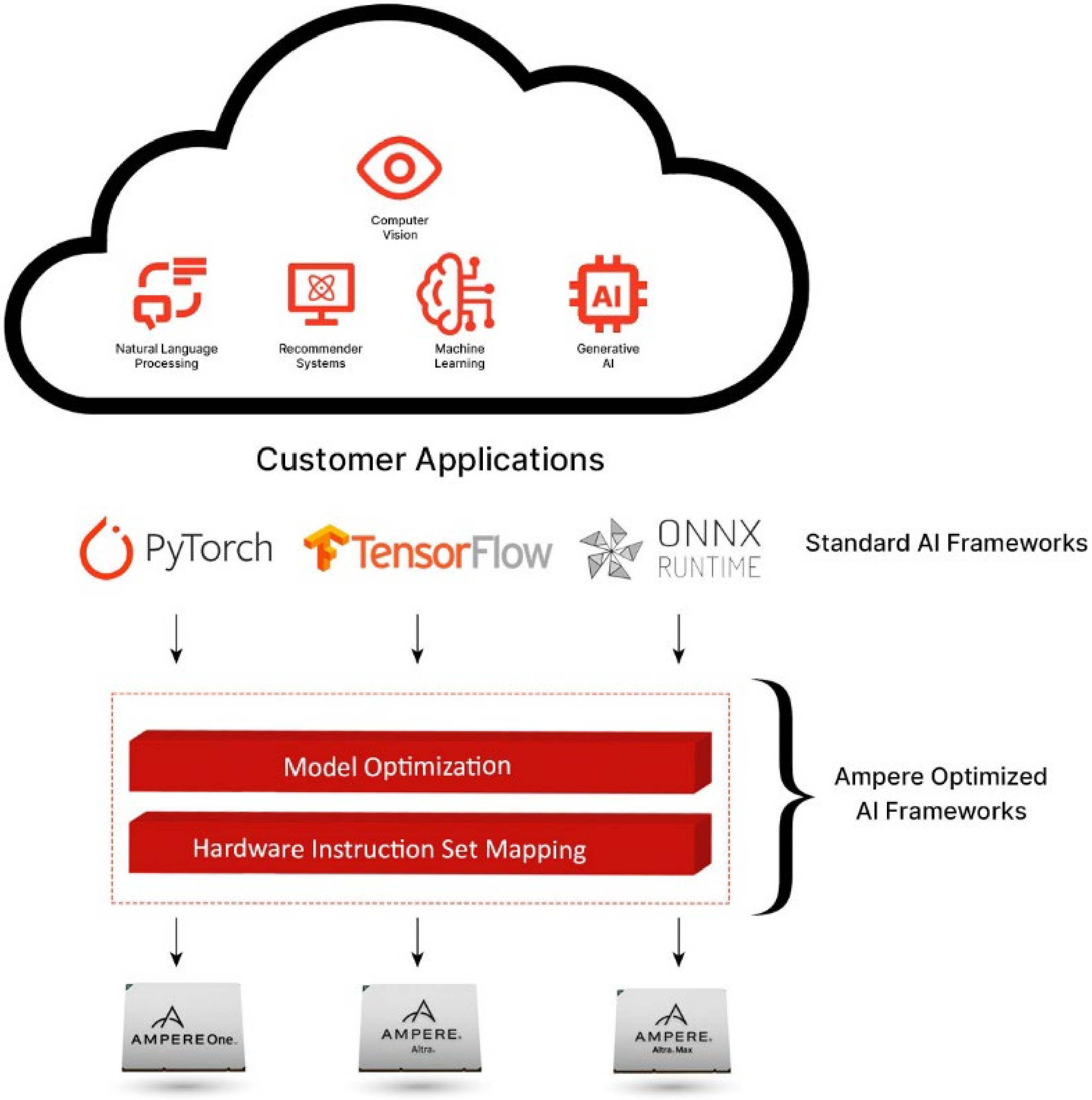
This demo shows a partner solution of a **generative AI chatbot deployment**. It processes incoming real-time input from a dedicated external microphone available to the visitors of the Ampere booth at Google Cloud Next. The demo runs at a **real-time performance level** with sufficient **low latency** and per second token generation rate to satisfy the user needs.

Resources

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.