Llama 3.1 on OCI A1 VMs

OCI A1 VMs, powered by
Ampere® Altra Cloud Native Processor and
Ampere® Optimized AI Frameworks (AIO),
deliver best-in-class AI inference.

Ampere Altra Powered Generative AI I

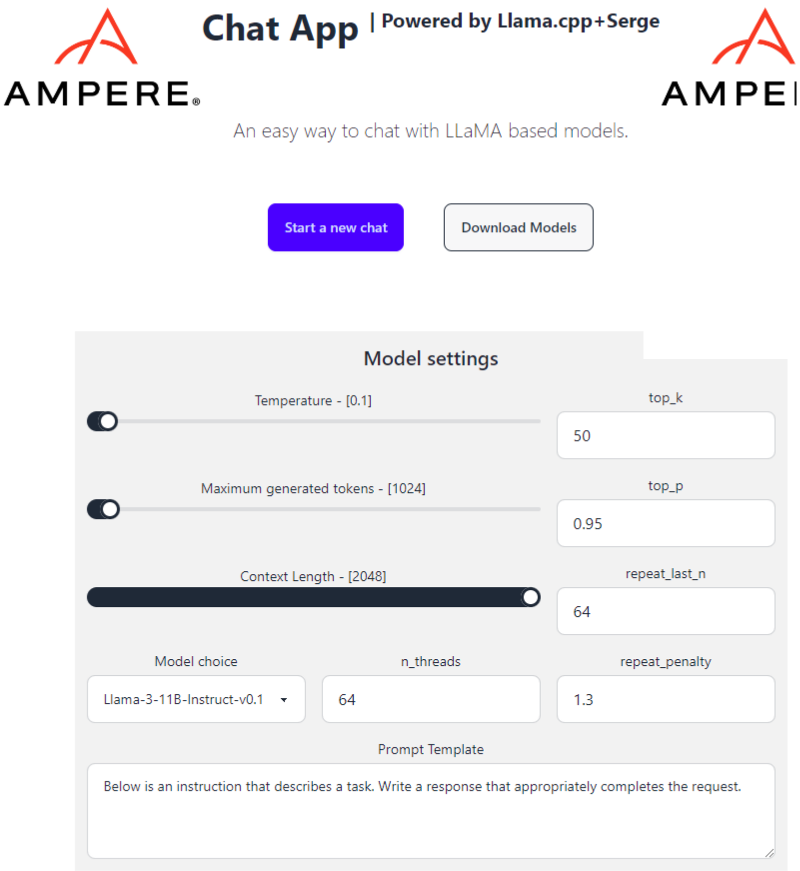
Ampere® **Cloud Native Processors** satisfy the performance requirements of widely used Large Language Models (LLMs), such as Meta's Llama 3.1, in particular for the smaller versions of these models. Ampere CPUs provide the **best price-performance and optimized power draw** for LLM deployments.

Setup

Deployment of the fine-tuned version of the open-source **Llama 3.1** model running on an Ampere-based **OCI A1** cloud instance with **Ampere® Optimized AI Frameworks (AIO) and Ampere Optimized llama.cpp**. The demo showcases real-time chatbot interaction on internally developed Ampere chatbot called Serge. The chatbot supports a variety of other models, e.g., Mistral, and allows for testing with different input or output token length and a chosen number of threads among other currently configurable parameters.

Key Benefits Demonstrated

- Meets or exceeds the necessary **low latency** requirements for real-time generative AI chatbot interaction.
- Delivers the **best price-performance** in LLM inference in cloud deployment scenarios.
- Provides the token generation rate needed for quality end-user experience.
- Ampere Cloud Native Processors enable **easy scaling** and can be **dynamically provisioned** based on the performance requirements of the user's applications.



Chat App | Powered by Llama.cpp+Serge

An easy way to chat with LLaMA based models.

Model settings

Temperature - [0.1]

top_k: 50

Maximum generated tokens - [1024]

top_p: 0.95

Context Length - [2048]

repeat_last_n: 64

Model choice: Llama-3-11B-Instruct-v0.1

n_threads: 64

repeat_penalty: 1.3

Prompt Template

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Figure 1: Llama 3 11B (11 billion parameters) Serge chatbot deployment on an Ampere Altra-based OCI A1 instance

Real-time Chatbot

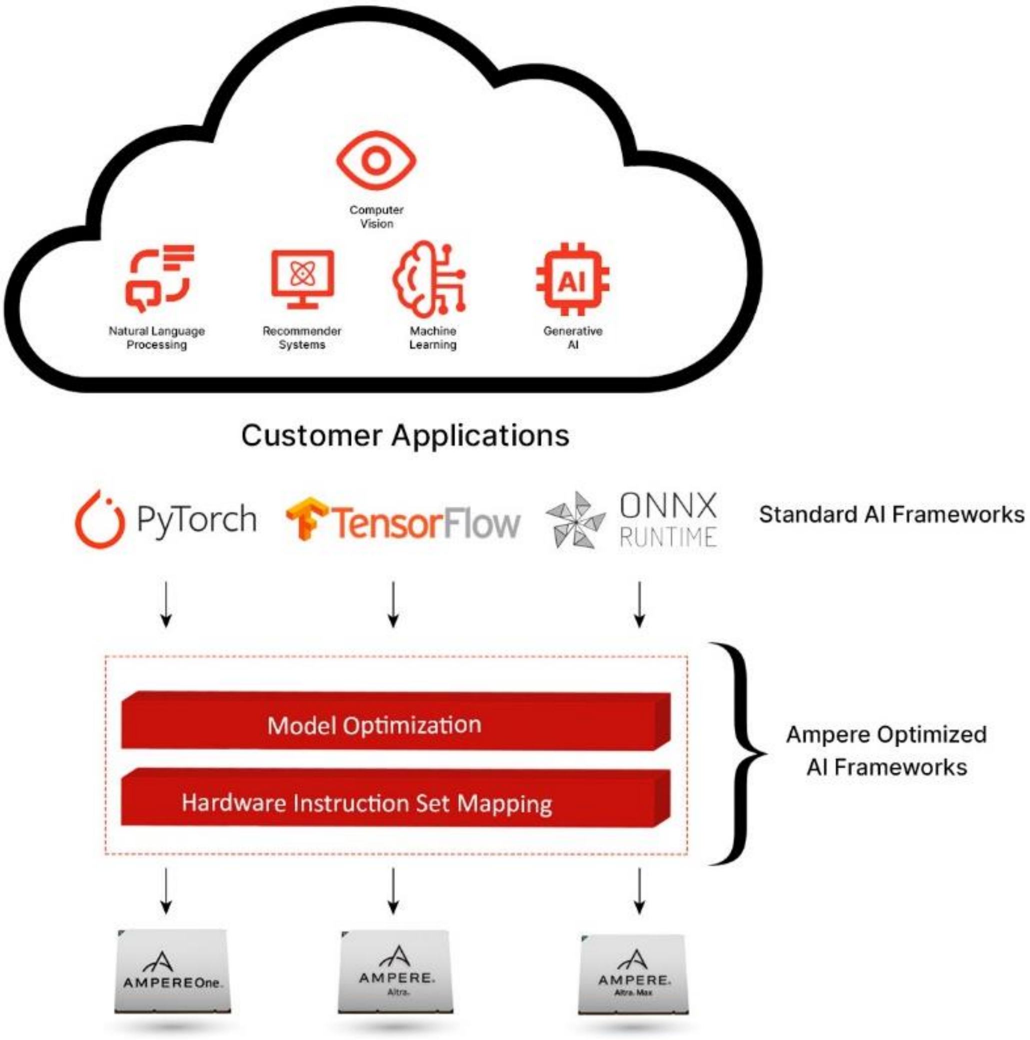
This demo shows a **generative AI chatbot deployment on OCI A1**. It processes incoming real-time input from in the form of a user prompt and generates the output per user's instructions. The demo runs at a **real-time performance level** with sufficient **low latency** and per second token generation rate to satisfy the user needs.

Resources

Ampere Optimized Pytorch, and other Ampere Optimized AI Frameworks (AIO), can be accessed directly from the Oracle Cloud Marketplace.

The docker images are also available in the downloads section of Ampere AI Developers 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 (AIO) 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.