SOLUTION BRIEF | AI & MACHINE LEARNING

POWER LLMS WITH NUTANIX CLOUD INFRASTRUCTURE (NCI)

AI IN A BOX

Powered by 4th Gen AMD EPYC™ Processors

August 2024

Large Language Models (LLMs) such as GPT-4 and Llama2 can impact many sectors of society, including education, art, medicine, healthcare, public service, business, communication, and entertainment. Academia, businesses, and governments are increasingly leveraging LLMs.

Traditional LLM deployments rely on complex, expensive Graphics Processing Unit (GPUs). This solution brief explains how you can quickly and easily deploy LLMs using the Nutanix Cloud Infrastructure (NCI) powered by 4th Gen AMD EPYC™ processors with no need for GPUs.

ABOUT THE SOLUTION

This solution uses the following hardware and software:

- **Nutanix NX-8155A-G9 Appliance:**
  - **4th Gen AMD EPYC processors:** AMD EPYC 9004 Series Processors are built on the proven x86 architecture and "Zen 4" cores that deliver efficient, optimized performance by combining high frequencies, the largest-available L3 cache, 128 lanes of PCIe® 5 I/O (1P), and synchronized fabric and memory clock speeds, plus support for up to 6 TB of DDR5-4800 memory. Built-in security features, such as Secure Memory Encryption (SME) and Secure Encrypted Virtualization (SEV-SNP), collectively known as AMD Infinity Guard, help protect data while in use.[1]
  - **Nutanix NCI:** This platform delivers easy management, scalability, performance, security, and reliability by supporting rapid deployment on hybrid and multi-cloud environments.
  - **Nutanix NKE:** This Kubernetes engine simplifies LLM deployment, management, and scalability while providing both high availability and security features.
- **Llama2:** This open-source LLM boasts advanced natural language capabilities, customizability, scalability, versatility, cost-effectiveness, and strong community support that make it well suited for a wide range of use cases.
- **Deployment and serving infrastructure:**
  - **Kubeflow:** This scalable, containerized open-source ML platform streamlines deploying and managing ML pipelines.
  - **TorchServe:** This open-source framework simplifies deploying Llama2 as a web service, allowing applications to easily interact with its capabilities.
  - **Llama.cpp:** This inference engine is efficiently written for the Meta LLaMA model. It supports various quantization methods to optimize model size and computation and is implemented in C++ for fast execution.

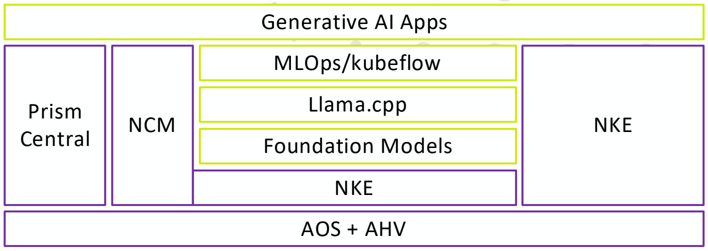

Figure 1: The Nutanix LLM software stack

KEY BENEFITS AND USE CASES

Key benefits of deploying Llama2 on Nutanix NX-8155A-G9 appliances include:

- No need for expensive GPUs enhances access to AI technology.
- 4th Gen AMD EPYC 9004 processors optimize Llama2 performance for fast text generation, natural language processing, and content creation.
- Nutanix NCI provides reliability, scalability, and easy management of your Llama2 deployment.
- Measure performance using detailed performance benchmarks.

Some use cases for this solution include:

- Automating tasks to streamline workflows, boost efficiency, and empower your workforce by automating repetitive tasks, such as data analysis and report generation.
- Enhancing customer service by delivering personalized experiences through intelligent chatbots, sentiment analysis, and AI-powered recommendations.
- Driving research efforts with LLM-powered automated data analysis, document summarization, and hypothesis generation.
- Enhancing competitiveness by leveraging AI-driven opportunities for innovation across text and language-driven domains.

TAKE THE FIRST STEP

Please see *4th Gen AMD EPYC™ Processors Power Nutanix® LLM in a Box* at the AMD Documentation Hub. Next, explore the provided resources and guidance to plan, deploy, and integrate Llama2 into your business operations. Leverage performance benchmark findings to unlock the full potential of your LLM deployment.

Don't let expensive hardware hinder your AI aspirations. Embrace the future of accessible and powerful AI with AMD, Nutanix, and Llama2.

[1] GD-183A: AMD Infinity Guard features vary by EPYC™ Processor generations and/or series. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/products/processors/server/epyc/infinity-guard.html.

DISCLAIMERS

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale. GD-18u

COPYRIGHT NOTICE

READY TO CONNECT? VISIT www.amd.com/epyc