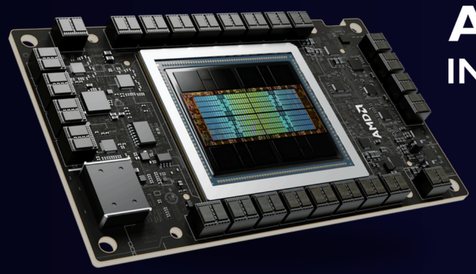
Industry’s First Multimodal RAG on Dell PowerEdge™ XE9680 Server with AMD Instinct™ MI300X Accelerators

| June 2024

In this blog, Metrum AI presents a multimodal RAG solution enabled by compute & memory capability of the AMD Instinct MI300X accelerators on Dell PowerEdge XE9680 servers.



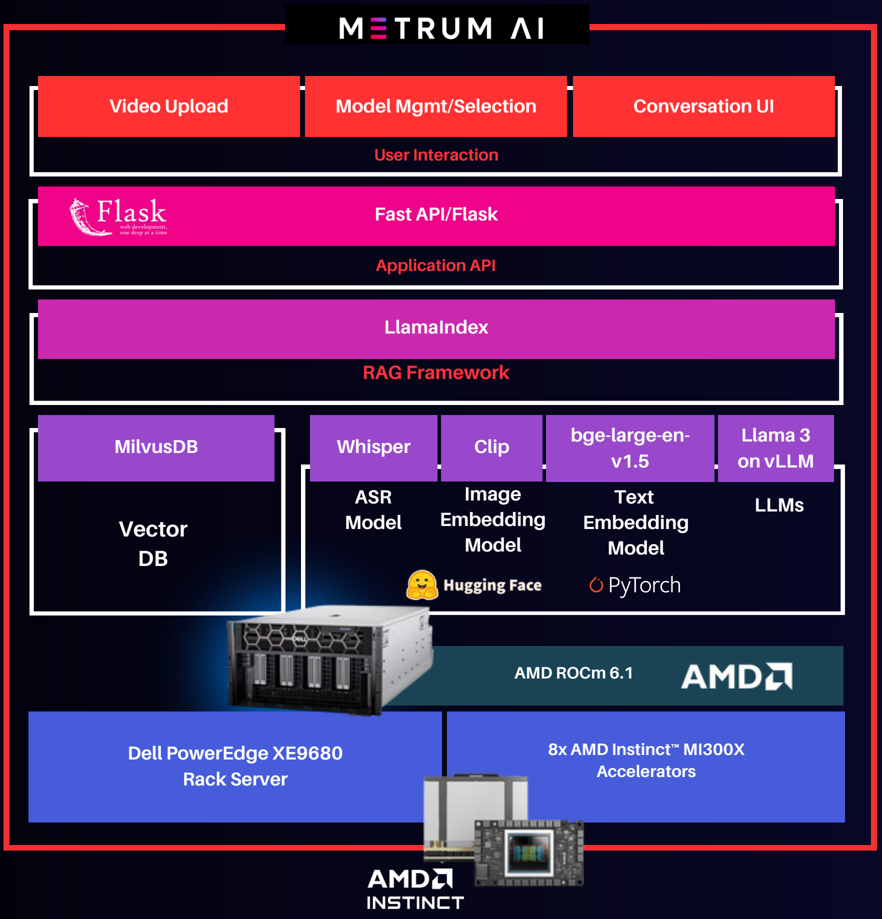
With the release of the AMD Instinct MI300X accelerator, we are now entering an era of choice for leading AI accelerators that power today’s retrieval-based generative AI solutions. Dell has paired the accelerators with its flagship PowerEdge XE9680 server for high performance AI applications. Leveraging this powerful combination, Metrum AI is excited to showcase a multimodal RAG (retrieval augmented generation) solution.

This offering can analyze audio, video, and text content, which is critical for enterprises as they operate on multimodal inputs whether the use cases are customer support calls, product quality images, employee training videos and more. In this blog, Metrum AI walks through an Earnings webcast scenario, provides insights into the solution architecture and user interface, and uncovers the following critical value drivers:

- Built multimodal RAG on AMD Instinct MI300X accelerator on Dell PowerEdge XE9680 server.
- Deployed four different models (language, image embeddings, text embeddings and voice) on a single **Dell PowerEdge XE9680 server with AMD Instinct MI300X accelerators.**
- Showcased live at Dell Tech World ‘24 in an Earning Webcast application demonstrating the use of language, voice, and vision for rapid analyst insights.

Solution Architecture

This solution leverages Dell PowerEdge XE9680 server equipped with eight AMD Instinct MI300X accelerators, along with AMD ROCm™ supporting an array of optimizations for generative AI workloads.



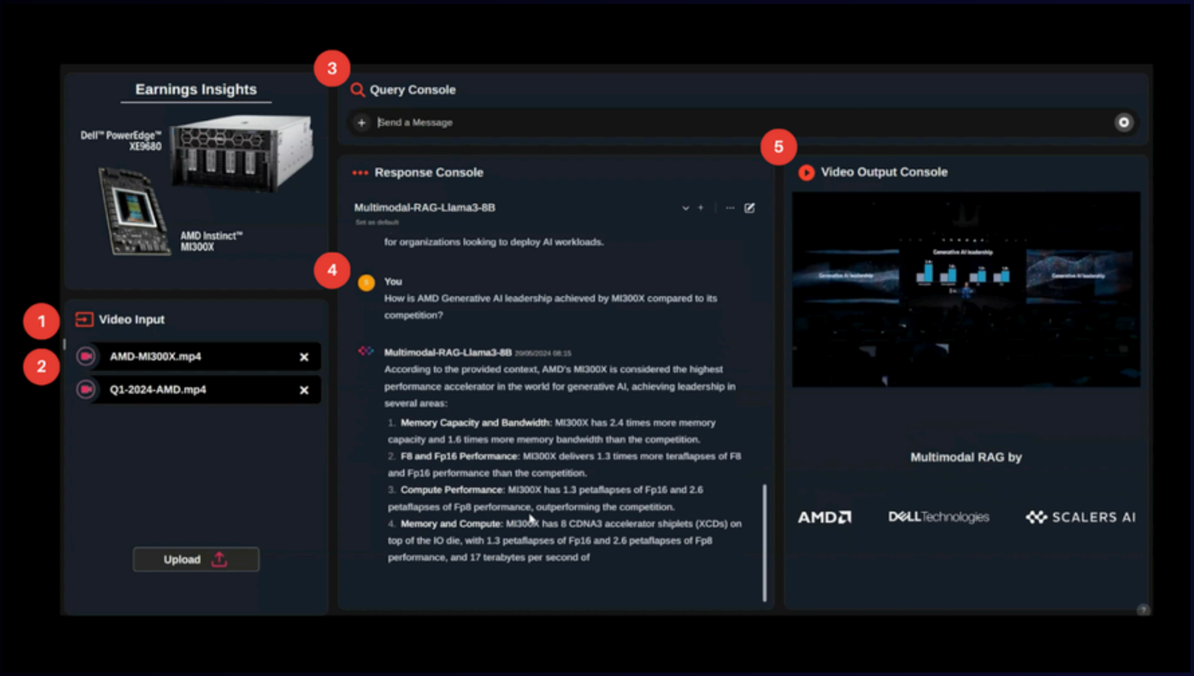
The software stack includes the following key components:

- MilvusDB, an open-source vector database
- OpenAI Whisper, an automatic speech recognition system (ASR)
- CLIP, an image embedding model
- bge-large-en-v1.5, a text embedding model, which captures syntactic and semantic information from text data and encapsulates the information in numerical vectors
- Llama 3 with vLLM, with LlamaIndex as a RAG Framework
- Fast API, to interface with user interactions including video file uploads, model selection, and using the conversation console.

Additional context on software components are available in the github repository.

Step-by-Step Demo Walkthrough

Earnings webcasts are a critical tool for publicly traded companies to communicate with their investors, analysts, and the broader market, and typically involve large amounts of data collected from the company’s knowledge base. RAG solutions enhance the accuracy, efficiency, and effectiveness of the earnings webcast process by extracting insights from company proprietary data, providing significant benefits to enterprises in terms of communication, decision-making, and stakeholder engagement. The user interface provided below illustrates the earnings webcast RAG solution user interface, through which users can upload relevant video files and query the application, which will then compile relevant information and provide references to the uploaded files from which the answers were extracted.



Multimodal-RAG-Llama3-8B

Set as default

for organizations looking to deploy AI workloads.

You

How is AMD Generative AI leadership achieved by MI300X compared to its competition?

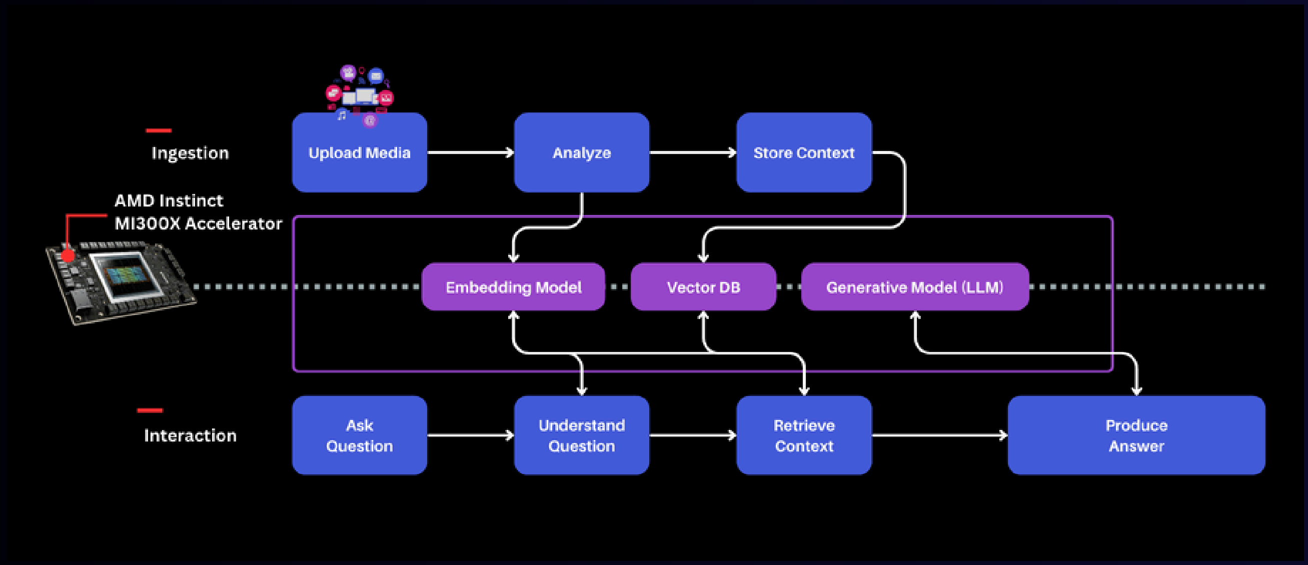
Multimodal-RAG-Llama3-8B 20/05/2024 08:15

According to the provided context, AMD's MI300X is considered the highest performance accelerator in the world for generative AI, achieving leadership in several areas:

1. **Memory Capacity and Bandwidth**: MI300X has 2.4 times more memory capacity and 1.6 times more memory bandwidth than the competition.
2. **F8 and Fp16 Performance**: MI300X delivers 1.3 times more teraflapses of F8 and Fp16 performance than the competition.
3. **Compute Performance**: MI300X has 1.3 petaflapses of Fp16 and 2.6 petaflapses of Fp8 performance, outperforming the competition.
4. **Memory and Compute**: MI300X has 8 CDNA3 accelerator shiplets (XCDs) on top of the IO die, with 1.3 petaflapses of Fp16 and 2.6 petaflapses of Fp8 performance, and 17 terabytes per second of

The steps below detail the flow of the solution, each step is marked in the interface illustration above:

1. User uploads input video files.
2. Solution creates image and text embedding from input video frames.
3. User converses with the application through the conversation console about information presented in the input video.
4. Solution generates text responses in the conversation console.
5. Solution retrieves a relevant video clip from the list of input video files. This provides the user with reference content from which the text response was generated.



The image above illustrates each segment of the workflow, and details how the embeddings model, vector databases, and generative AI models interact with the data and user queries within the application.

As you can see from this demo, showcased live at Dell Tech World '24, enterprises can now take advantage of their various data types, whether they involve voice, video, images, or text. This enables them to scale and enhance multiple use cases such as employee onboarding, customer support assistants, critical document generation, all while keeping their proprietary data and workflows private. Dell's flagship PowerEdge XE9680 server with eight AMD Instinct MI300X accelerators supports the memory footprint needed for these rich multimodal data and model intensive use cases.

In this blog, we showcased how enterprises deploying applied AI can use their own proprietary data to take advantage of multimodal RAG capabilities in the context of an Earnings Webcast Insights and explored the capabilities of Dell PowerEdge XE9680 server equipped with AMD Instinct MI300X accelerators with the following milestones:

- Built multimodal RAG on AMD Instinct MI300X accelerator on Dell PowerEdge XE9680 server.
- Deployed four different models (language, image embeddings, text embeddings and voice) on a single **Dell PowerEdge XE9680 server with AMD Instinct MI300X accelerators.**
- Showcased live at Dell Tech World '24 in an Earning Webcast application demonstrating the use of language, voice, and vision for rapid analyst insights.

The reference code along with more information can be found here.

Additional Criteria for IT Decision Makers:

What is RAG, and why is it critical for enterprises?

RAG, which stands for Retrieval-Augmented Generation, is a method in natural language processing (NLP) that enhances the generation of responses or information by incorporating external knowledge retrieved from a large corpus or database. This approach combines the strengths of retrieval-based models and generative models to provide more accurate, informative, and contextually relevant outputs.

The key advantage of RAG is that it leverages a large amount of external knowledge dynamically, enabling the model to generate responses that are not just based on its training data but also on up-to-date and detailed information from the retrieval phase. This makes RAG particularly useful in applications where factual accuracy and detail are crucial, such as in customer support, academic research, and other domains requiring precise information. Ultimately, RAG provides enterprises with a robust tool for improving the accuracy, relevance, and efficiency of their information systems, leading to better customer service, cost savings, and competitive advantages.

Why is the Dell PowerEdge XE9680 Server with AMD Instinct MI300X Accelerators Well-suited for RAG Solutions?

Designed especially for AI tasks, Dell PowerEdge XE9680 server is a powerful data-processing server equipped with eight AMD Instinct MI300X accelerators, making it well-suited for AI-workloads, especially for those involving training, fine-tuning, and conducting inference with Large Language Models (LLMs). AMD Instinct MI300X accelerator is a high-performance AI accelerator intended to operate in groups of eight within AMD's generative AI platform.

Implementing Retrieval-Augmented Generation (RAG) solutions effectively requires a robust hardware infrastructure that can handle both the retrieval and generation components efficiently. Critical hardware features for RAG solutions include high performance accelerator units and large RAM and storage capacity. With 192 GB of GPU memory, a single AMD Instinct MI300X accelerator can host an entire Llama 3 70B parameter model for inference. It is optimized for generative AI and can deliver up to 10.4 Petaflops of performance (BF16/FP16), and provides 1.5TB of total HBM3 memory in a group of eight accelerators.

***DISCLAIMER - Performance varies by hardware and software configurations, including testing conditions, system settings, application complexity, the quantity of data, batch sizes, software versions, libraries used, and other factors. The results of performance testing provided are intended for informational purposes only and should not be considered as a guarantee of actual performance.