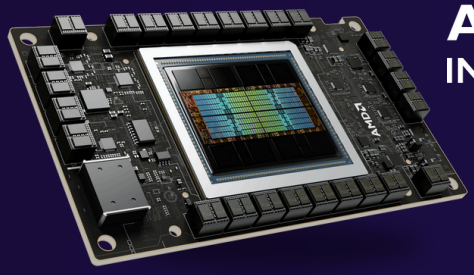
Enhancing Telecom Quality of Service with Generative AI Agentic RAG

This blog presents a telecom quality of service (QoS) Gen AI Agentic RAG solution powered by the AMD Instinct MI300X accelerators on Dell PowerEdge XE9680 servers.

| September 2024



AMD INSTINCT

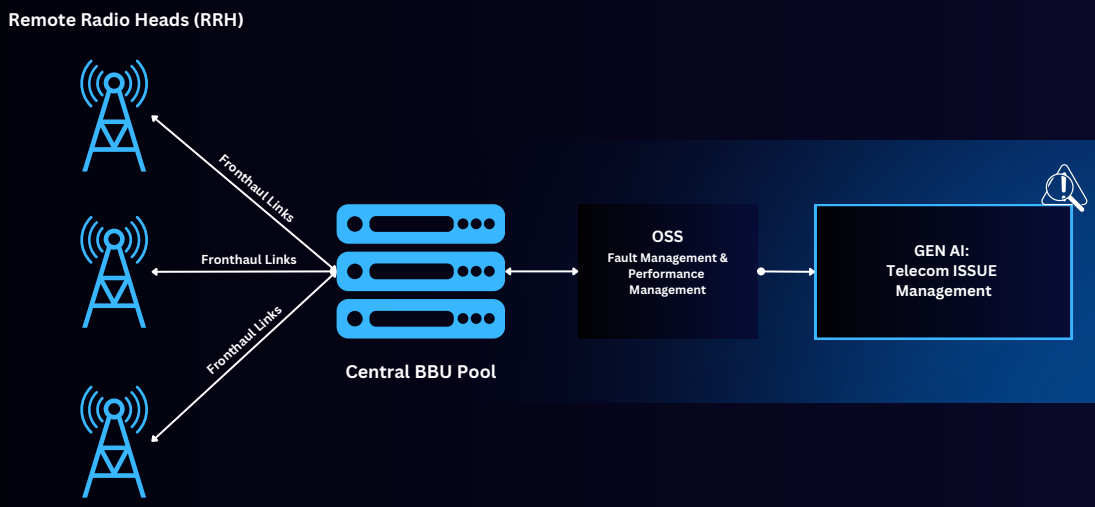
DELL Technologies AMD METRUM AI Hugging Face

The introduction of the AMD Instinct MI300X accelerator, now integrated with Dell's flagship PowerEdge XE9680 server, provides a robust platform for high-performance AI applications. Leveraging this powerful combination, we have developed a Telecom Issue Management solution to demonstrate the value of Generative AI in optimizing network operations for telecom companies and their enterprise clients. This solution is crucial for minimizing network downtime, ensuring consistent service quality, and enabling telecom operators to make informed decisions about network investments and improvements.

In this blog, we offer insights into the solution architecture built with industry-leading software and hardware components, showcasing the following:

- How to utilize LLM-based agentic RAG to build a telecom issue management solution
- How to deploy a cutting-edge language model, embeddings model, and vector database on a **Dell PowerEdge XE9680 server equipped with eight AMD Instinct MI300X accelerators**
- How to navigate an up-to-date and intelligent dashboard that tracks and recommends actions for telecom issues, with a focus on base station performance

Understanding the Telecom Landscape



The expected data volume over cellular networks is projected to exceed hundreds of exabytes per month by 2025, driven by human and machine data, representing tens of billions of devices. This explosive growth poses a challenge in ensuring that networks remain robust and resilient while handling increased demand for data throughput and low-latency services. This creates challenges in maintaining high-quality service and securing critical information. Telecom infrastructure providers must ensure the integrity and resilience of their networks to prevent costly disruptions. For example, network downtime caused by equipment failure at a Remote Radio Head (RRH) can lead to widespread service outages, particularly in densely populated areas. Even worse, cascading failures can occur when a single failure triggers a chain reaction, affecting multiple base stations across the network. With tens of thousands of RRHs deployed across geographically dispersed areas, addressing these issues manually—such as sending out repair teams—can be time-sensitive and extremely costly.

Other common challenges include signal degradation due to environmental factors, misconfigurations in the Baseband Unit (BBU), and congestion from unexpected traffic spikes. These problems not only affect the quality of service but also increase the risk of security vulnerabilities, making real-time monitoring and automated issue resolution critical for minimizing service disruptions and maintaining operational efficiency.

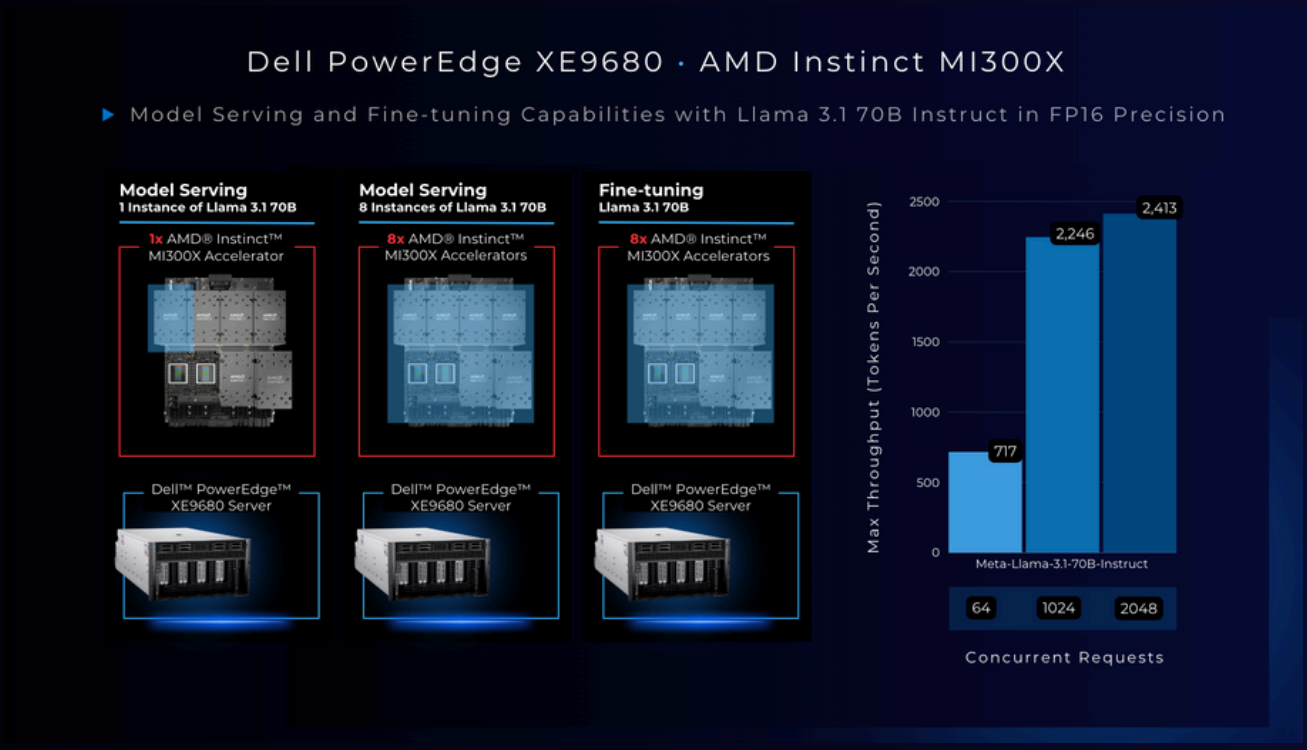
To better understand the impact of this solution, it is important to gain some foundational knowledge on the telecom landscape (as represented in the simplified diagram above):

- **RRH (Remote Radio Head):**
  - RRHs are deployed across multiple sites to perform basic signal transmission and reception functions, and handle radio frequency (RF) processing at cell sites.
- **BBU (Baseband Unit):**
  - BBUs are aggregated within a centralized BBU pool to provide robust computing capabilities for baseband signal processing. They handle the digital processing of signals.
- **OSS (Operations Support Systems):**
  - OSS comprises software tools that analyze and manage the telecom network, including network monitoring, fault management, and performance optimization.

In a typical telecom network, RRHs are connected to the BBU pool through fronthaul links. BBUs can be dynamically assigned to serve clusters of RRHs in a many-to-one configuration. The OSS provides a comprehensive network view by integrating data from RRHs and BBUs, ensuring that the telecom network meets Quality of Service (QoS) requirements.

In this implementation, we simulate a telecom network with several RRHs, a centralized BBU pool, and an OSS using synthetic data streams, specifically system logs from various telecom devices and OSS event logs. The following section provides a detailed walkthrough of the solution architecture and user interface. In doing so, we demonstrate how telecommunications service providers can use **agentic RAG solutions** to analyze and address quality of service and security challenges in near real-time, increasing the uptime of their networks and improving the quality of service.

Solution Architecture



We selected the Dell PowerEdge XE9680 equipped with AMD Instinct MI300X accelerators for our solution due to its exceptional performance and memory capacity, which is crucial for handling the latest high parameter count large language models. With 192GB of HBM3 memory per accelerator, we can comfortably **run the entire Llama 3.1 70B model on a single accelerator**. Memory and compute-intensive workloads, such as serving multiple model instances and fine-tuning, are also possible using only one hardware system with eight accelerators. As shown in the chart above, max token throughput with vLLM model serving of Llama 3.1 70B scales by a factor of 3 with an increase in concurrent requests, achievable due to the unparalleled memory capacity of the AMD Instinct MI300X accelerator combined with the Dell PowerEdge XE9680 server.

To deliver an industry specific solution, we paired a large language model with critical software components as shown in the architecture below, such as a cutting-edge text embeddings model and vector database. The memory and performance capabilities of Dell PowerEdge XE9680 with AMD Instinct MI300X accelerators make it possible to support this extensive software stack without compromising accuracy or efficiency.

This solution leverages the following technologies:

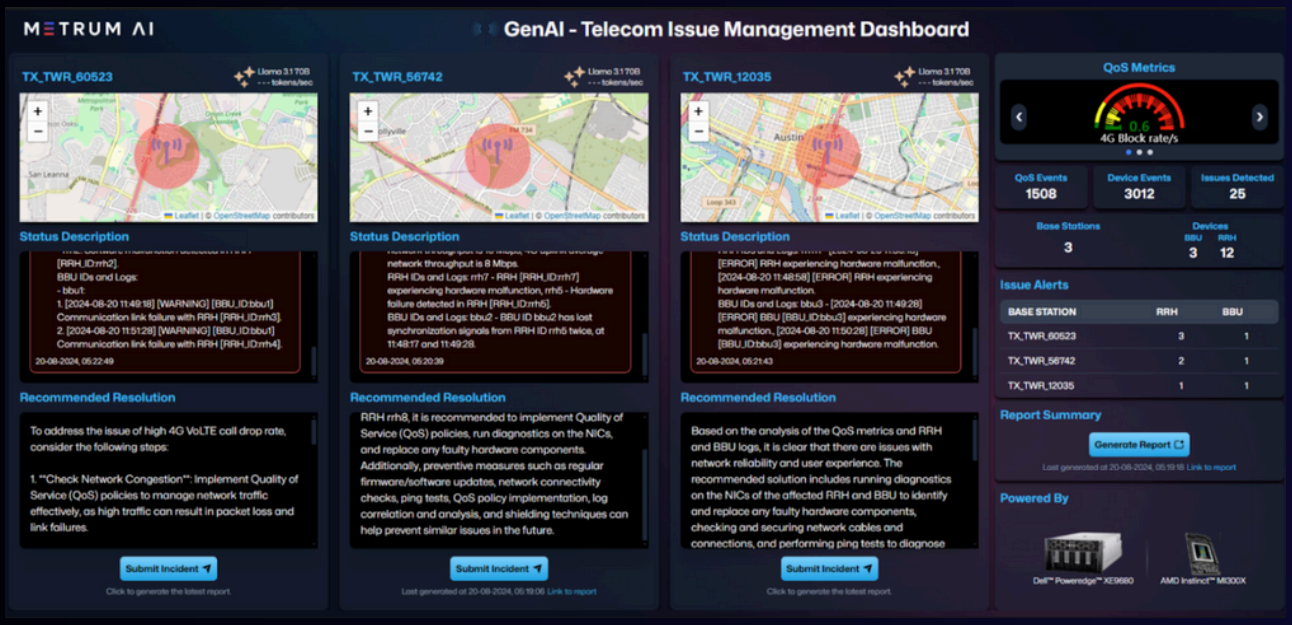
- **Utilization of LLM Agents**: LLM agents are AI systems that leverage large language models to understand and respond to natural language inputs. These agents can perform tasks such as answering queries, generating text, or autonomously controlling software applications with minimal human intervention. These agents can adapt to different tasks, making them valuable in customer service, automation, and content creation.
- **Plug-and-Play Architecture:** This architecture enables the seamless integration of essential components, including vector databases, embedding models, and LLMs, into existing systems.
- **Extensibility**: The system is designed to efficiently process thousands of messages daily across hundreds of RRHs on a single server, with the architectural support to scale out as needed.

To enable these features, the software stack includes the following key components:

- **vLLM (v0.5.3.post1),** an industry-standard library for optimized open-source large language model (LLM) serving, with support for AMD ROCm 6.1.
- **llama-agents,** an async-first framework for building, iterating, and productionizing multi-agent systems.
- **Llama 3.1 70B Model,** an industry-leading open-weight language model with 70 billion parameters, served using vLLM with AMD ROCm optimizations.
- **LlamaIndex,** a popular open-source retrieval augmented generation framework.
- **bge-large-en embeddings model**, one of the top ranked text embeddings models running with Hugging Face APIs.
- **MilvusDB,** an open-source vector database with high performance embedding and similarity search.
- **kafka broker**, a key component of Apache Kafka's distributed architecture. It manages data storage and transfer between producers and consumers while ensuring scalability, fault tolerance, and high performance.

Solution Overview

This solution is centered on ***real-time*** monitoring and management of telecom services, highlighting the potential of Generative AI in creating detailed incident reports and recommending appropriate responses. The image below depicts the user interface of the telecom issue management solution, through which users can track and summarize incidents at various base stations, submit incident alerts, and generate comprehensive summaries of all ongoing incidents along with suggested resolution plans.



- **Base Station Simulated Logs**
  - The logs are generated at each base station from their respective RRH. The corresponding colors indicate the following:
    - Green - List of logs, where all are [INFO] logs and all RRHs are operating normally.
    - Red - Unique list of RRHs at each base station, where logs have a [WARNING] or [ERROR] status that needs to be resolved.
- **Incident Report**
  - As soon as logs of [WARNING] or [ERROR] status (indicated by the color red) are clicked, an incident report is raised.
- **QoS Metrics** (top right)
  - Continuous live QoS Metrics averaged across all base stations, are visible using Gauge.
- **Events and Issues** (middle right)
  - The events and issues panel provides a count of all events coming from RRH and BBU logs, as well as QoS metrics. The "Issues Detected" quantity represents the count of all logs of [WARNING] or [ERROR] status.
- **Issue Alerts**
  - This panel provides information about issue mapping between RRHs, BBUs, and base stations.
- **Report Summary**
  - This feature allows users to generate a concise summary report of recent incidents across all base stations.

The image above illustrates each segment of the workflow, and details how the RAG-based agentic workload and vector database interact with the simulated telecom network data.

As demonstrated in our implementation, telecom operators can now leverage Generative AI to enhance the monitoring and management of telecom services, while ensuring the privacy of their proprietary data and workflows. Dell's flagship PowerEdge XE9680 server, equipped with eight AMD Instinct MI300X accelerators, provides the necessary memory footprint to support these rich multimodal data and model-intensive use cases.

In this blog, we demonstrated how enterprises deploying applied AI can leverage their proprietary data to harness multimodal RAG capabilities in the context of a telecom issue management tool. We explored the capabilities of the Dell PowerEdge XE9680 server equipped with AMD Instinct MI300X accelerators, achieving the following milestones:

- Developed a telecom issue management solution using LLM-based agentic RAG.
- Deployed cutting-edge language model, embeddings model, and vector database on **Dell PowerEdge XE9680 server with eight AMD Instinct MI300X accelerators.**
- Integrated an intelligent, real-time dashboard that monitors and recommends actions for telecom issues, with a focus on base station performance.

To learn more, please request access to our reference code by contacting us at contact@metrum.ai.

Additional Criteria for IT Decision Makers

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

Retrieval-Augmented Generation (RAG), 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 deliver more accurate, informative, and contextually relevant outputs.

The key advantage of RAG is its ability to dynamically leverage a large amount of external knowledge, allowing the model to generate responses that are informed not only based on its training data but also by up-to-date and detailed information from the retrieval phase. This makes RAG particularly valuable in applications where factual accuracy and comprehensive details are essential, such as in customer support, academic research, and other fields that require precise information.

Ultimately, RAG provides enterprises with a powerful 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, the 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).

Effectively implementing Retrieval-Augmented Generation (RAG) solutions requires a robust hardware infrastructure that can handle both the retrieval and generation components. Key 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. Optimized for generative AI, the AMD Instinct MI300X accelerator 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.

What are AI agents, and what is an agentic workflow?

AI agents are autonomous software tools designed to perceive their environment, make decisions, and take actions to achieve specific goals. They utilize artificial intelligence techniques, such as machine learning and natural language processing, to interact with their surroundings, process information, and perform tasks with varying degrees of independence and complexity.

An agentic workflow in AI refers to a sophisticated, iterative approach to task completion using multiple AI agents and advanced prompt engineering techniques. Unlike traditional single-prompt interactions, agentic workflows break complex tasks into smaller, manageable steps, allowing for continuous refinement and collaboration between specialized AI agents. These workflows leverage planning, self-reflection, and adaptive decision-making to achieve higher accuracy and efficiency in task execution. By employing multiple AI agents with distinct roles and capabilities, agentic workflows can handle complex problems more effectively, often producing results that are significantly more accurate than conventional methods. This approach represents a shift towards more autonomous, goal-oriented AI systems capable of tackling intricate challenges across various domains.

Resources

AMD images: AMD Library, https://library.amd.com/account/dashboard/

Dell images: Dell.com

***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.