DELL Technologies

AMD
together we advance_

2024 | White paper

Empowering Enterprises with AI: Entering the Era of Choice

Table of Contents

2024 | White paper

The Opportunity to Transform Industries with AI

Today, there is no greater opportunity to transform your business for the future of innovation thanks to AI. Data collected from Accenture Vision Technology 2023 has shown that 98% of global executives agree AI foundation models will play an important role in their organization's strategies in the next three to five years.[1]

AI has become incredibly useful for businesses in fields such as retail, healthcare, and financial services due to its ability to enhance efficiency of tasks, drive innovation, and improve decision-making processes. However, despite the advantages, there is still a perceived barrier to entry when it comes to integrating AI due to some common misconceptions.

You need a team of AI developers to get started:

While expertise in data science is still valuable for developing advanced AI solutions and understanding the underlying principles, it's no longer a prerequisite. There has been a proliferation of user-friendly AI tools, platforms such as Hugging Face, and task-specific models that abstract away much of the complexity involved in developing AI solutions.

You need to spend tens of millions in hardware to get results:

This misconception severely undermines the diversity of AI resources available today. While these commonly known resources are often powerful and well-supported, they may not always be the most suitable or cost-effective choice for every business.

You need to work tirelessly to acquire accelerators:

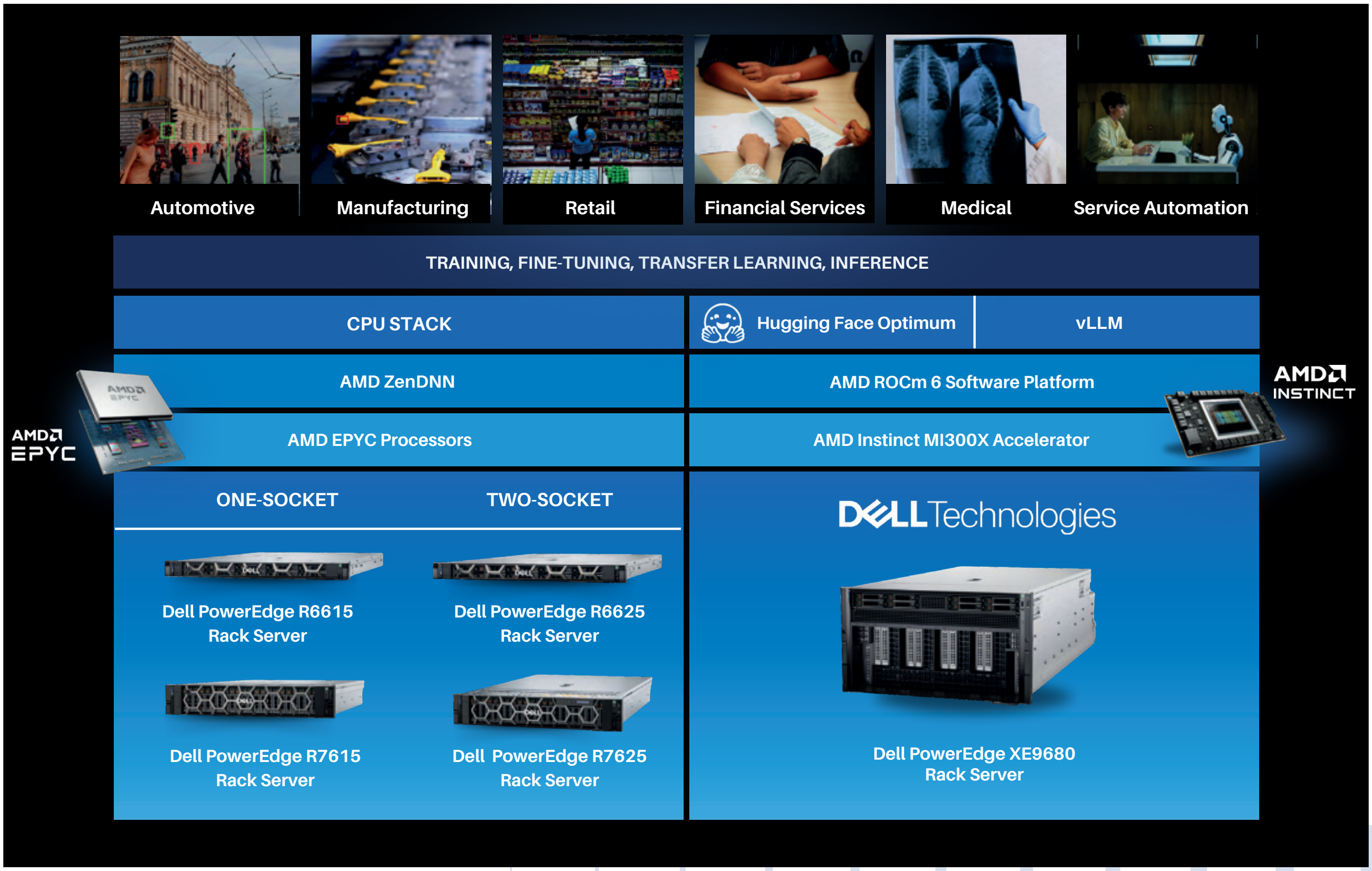
While accelerators do excel on heavy AI workloads, businesses may not need that much compute power for their AI applications. Waiting for an excessively long period of time to get access to market leading accelerators also simply isn't realistic. In many cases, AI-optimized CPUs can actually deliver the performance and efficiency necessary for producing AI assisted analyses and decisions in real time, and are a much more cost-effective and adaptable solution.

[1] Accenture, March 30, 2023, "Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds", https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds

Fortunately, the AI landscape is evolving. Together, **Dell** and **AMD** are partnering to break these myths by making AI technologies and tools accessible to a broader range of users with end-to-end infrastructure designed to support the AI demands of today.

> You can get started with an already optimized model, a reliable software stack, and a versatile hardware system, all of which are openly available through Dell and AMD's partnership. Having access to increasingly scarce accelerators, a substantial group of skilled AI engineers, or resources for deploying massive cloud clusters is no longer a requirement for leveraging AI.

What **Dell** and **AMD**'s collaboration offers is a unified ecosystem of hardware and software, designed to allow developers to create end-to-end AI solutions that incorporate transfer learning, fine-tuning, and inferencing easily and efficiently. With support from **Hugging Face**, we now have a growing portfolio of models that run on Dell PowerEdge servers with AMD EPYC™ processors or AMD Instinct™ MI300X accelerators, so that developers can fine-tune, apply transfer learning, and deploy for inference. The investments in AMD ROCm™ and AMD ZenDNN™ as well as partnerships with PyTorch, Tensorflow, and ONNX Runtime frameworks, are the fundamental enablers of Applied AI developers experiencing the democratization of AI. The stack diagram below details the components that make up Dell and AMD unified AI ecosystem.



Automotive

Manufacturing

Retail

Financial Services

Medical

Service Automation

TRAINING, FINE-TUNING, TRANSFER LEARNING, INFERENCE

CPU STACK

Hugging Face Optimum

vLLM

AMD ZenDNN

AMD ROCm 6 Software Platform

AMD EPYC Processors

AMD Instinct MI300X Accelerator

AMD EPYC

AMD INSTINCT

ONE-SOCKET

TWO-SOCKET

Dell PowerEdge R6615 Rack Server

Dell PowerEdge R6625 Rack Server

Dell PowerEdge R7615 Rack Server

Dell PowerEdge R7625 Rack Server

DELL Technologies

Dell PowerEdge XE9680 Rack Server

AI In Industry

With the diversification of resources and emphasis on open-source innovation, AI is migrating into many different industries, including customer service, finance and banking, healthcare, and retail to name a few. Across these industries, however, AI collectively enables organizations to unlock the potential of their own proprietary data and reimagine their AI workflows by tackling the following key capabilities: data analysis, automation, personalization, and predictive analytics. AMD ROCm and ZenDNN libraries additionally accelerate these AI workflows to deliver results in near-real-time.

Take a closer look at how exactly AI influences various industries below.

Automotive

AI is used for object detection, lane tracking, and decision-making in autonomous vehicles. AI can also predict when a vehicle component is likely to fail, allowing for proactive maintenance and reducing downtime.

Manufacturing and Industry

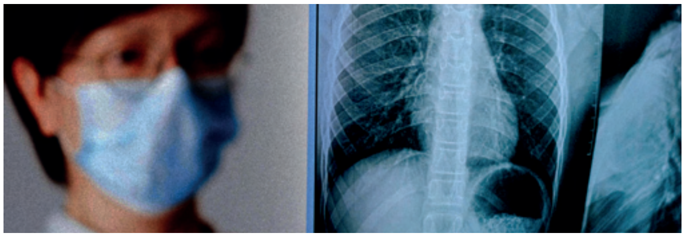
AI can be used in manufacturing and industry for predictive maintenance, quality control, process optimization, and supply chain management, leading to improved efficiency and reduced downtime.

Retail

AI can analyze customer behavior to provide personalized product recommendations, improving customer engagement and sales. It can also optimize inventory levels by predicting demand and minimizing overstock or stockouts.

Financial Services

AI can be used in finance and banking for fraud detection, risk assessment, customer service, and investment analysis, leading to improved security and more informed decision-making.



Medical

AI can be used in healthcare for a variety of applications, including medical image analysis, disease diagnosis, personalized treatment planning, and drug discovery, leading to improved patient outcomes and reduced costs.

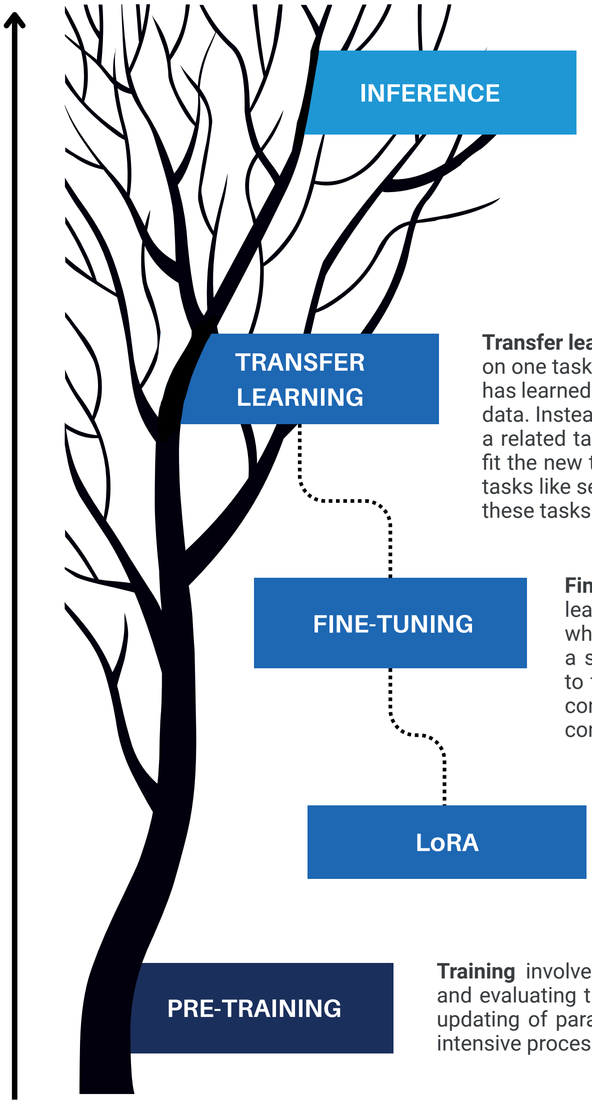
Service Automation

AI-powered chatbots can handle customer inquiries and provide support, reducing the need for human intervention. AI can also automate repetitive tasks such as data entry or document processing, improving efficiency and reducing errors.

What IT Decision Makers Must Consider

GETTING STARTED: BREAKING DOWN AI

Before navigating through these use cases, let's take a deeper look into the AI lifecycle. The AI (Artificial Intelligence) lifecycle refers to the stages involved in developing, deploying, and maintaining an AI system. While specific methodologies and terminology can vary, a typical AI lifecycle always includes model training and inferencing.



Inferencing is the process of using a trained AI model to make predictions, classifications, or decisions based on new, unseen data. In this stage of the AI lifecycle, the trained model is deployed and used in a production environment to perform the specific task it was designed for. During inference, the model takes input data, processes it using the learned patterns and relationships from the training phase, and produces an output, such as a prediction, a classification, or text in GenAI workloads. This output can then be used for various applications, such as recommending products to a user, detecting anomalies in a system, or recognizing objects in an image. In natural language processing, retrieval-augmented generation (RAG) improves accuracy when deploying models by allowing models to access and leverage external knowledge sources, enabling them to better understand and respond to natural language input in a wider range of contexts.

Transfer learning, also referred to as re-training, is a machine learning technique where a model trained on one task is re-purposed on a second related task. The general idea is to use the knowledge a model has learned from a task with a lot of available labeled training data in a new task that doesn't have much data. Instead of starting the learning process from scratch, we start with patterns learned from solving a related task, effectively updating the parameters of the model trained for the related task to better fit the new task. Transfer learning is mostly used in computer vision and natural language processing tasks like sentiment analysis to reduce the normally huge amount of computational power required for these tasks.

Fine-tuning in the AI life cycle refers to the process of further training a pre-trained machine learning model on a new dataset or task. This technique is commonly used in transfer learning, where a model trained on a large dataset for a specific task is adapted to a related task or a smaller dataset. Fine-tuning allows the transfer of knowledge from the pre-trained model to the new task or dataset, often resulting in faster convergence and improved performance compared to training a model from scratch. It is widely used in various AI applications, including computer vision, natural language processing, and speech recognition.

LoRA (Low Rank Adaptation) is a new technique for fine-tuning deep learning models that work by reducing the number of trainable parameters and enable efficient task switching, ultimately making the fine-tuning process less compute intensive and more accessible with widely available hardware.

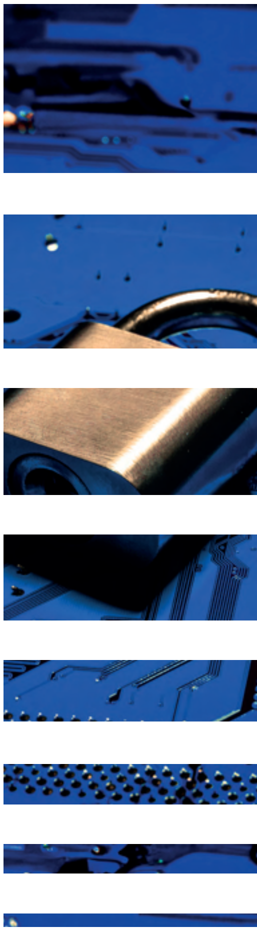
Training involves selecting the appropriate algorithm or model architecture, tuning model parameters, and evaluating the model's performance internally. Depending on the size of the model being trained, the updating of parameters in each iteration of the training process can be an extremely data and compute intensive process.

Figure 1: The AI Lifecycle

CRITICAL CHOICES

| Performance

In many of these real- world applications, real-time or near-real-time decision-making is critical for success. For example, fraudulent activities in financial transactions or insurance claims must be identified promptly to prevent financial losses and protect businesses assets. In a manufacturing scenario, defects in the assembly line or factory conditions must be monitored dynamically for quality assurance. Effectively, the processor handling your inference workload must be optimized for processing incoming data streams quickly and efficiently. Dell PowerEdge servers paired with AMD EPYC processors are a versatile combination, well-suited for handling edge inference workloads, as well as tasks involving high-performance computing, cloud computing, and big data analytics.



| Data Security

Data security is crucial for the success of AI systems, especially those leveraging generative AI, and is an important concern for technology leaders aiming to incorporate AI into their operations. AI systems typically rely on massive amounts of data, which may include sensitive and confidential information such as personal details, financial data, or proprietary information. Safeguarding this data is critical to prevent unauthorized access or data theft, as well as to ensure the precision, dependability, and consistency of AI models and predictions.

Confidential computing is a technology that facilitates data processing in a secure enclave, protecting it from unauthorized access or manipulation by unauthorized parties, including the cloud provider and other users.[2] Encryption and other security measures are used to isolate the data during processing. The AMD Infinity Guard, a collection of sophisticated security features integrated into AMD EPYC processors, supports confidential computing by employing Secure Encrypted Virtualization (SEV), which encrypts virtual machines (VMs) using a key known only to the processor. These services aim to provide hardware-based trusted execution environments using AMD SEV-Secure Nested Paging (SEV-SNP), which enhances guest protections to help defend against external threats.

Federated learning is another method for maintaining data security. It trains a central model across decentralized devices or servers.[3] Rather than transferring all data to a central location, each device trains the model locally, and only the model updates are shared. This approach preserves privacy and enables collaborative learning without sharing raw data. The Federated AI platform from Dell Technologies enables computational processes, AI, and ML algorithms to be run on datasets at the network edge as they are collected, sharing only mathematical models, metadata, and query results over the network to other edge devices, data centers, or the cloud. This exchange enhances results by enabling the near real-time extraction of actionable insights from large, distributed datasets without revealing the data and any intellectual property.

[2] Advanced Micro Devices, Inc. 2023, August 30, "AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings", https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html
Advanced Micro Devices, Inc., 2021, "Data Center Solutions, Confidential Computing" Solution Brief, https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

[3] Analytics Vidhya, 2023, Dec, "Federated Learning: A Beginner's Guide", https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data
Dell Technologies, 2021, "A federated learning platform for real-time artificial intelligence" Solution Brief, https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

SCALING YOUR SOLUTION

| Balancing Cost And Innovation

Striking the right balance between cost and innovation, ensures that AI solutions are not only financially feasible but also impactful, driving real value for both businesses and users. A key component to finding this balance, lies in identifying a hardware that both solves your use cases and integrates easily into existing infrastructure. In the modern AI hardware market, the increased demand for accelerators from various industries, on top of production capacity constraints, logistics challenges, and semiconductor shortages, are all contributing to accelerator shortages.

CPUs, however, are already a standard component in most data centers, making integration simpler and more cost-effective compared to adding entirely new accelerator hardware. AI-optimized CPUs can leverage existing software and tooling, reducing the need for extensive retooling or retraining. CPUs also offer greater flexibility and efficiency for a wide range of tasks beyond AI, allowing for a more versatile use of resources within the data center. Refreshing your data center with Dell PowerEdge servers running AMD EPYC processors supports your existing workloads to be fulfilled, while remaining ready for advancements towards more innovation and efficiency driven by AI.

| Simplicity And Flexibility

Simplicity and flexibility of your AI system are essential for building AI solutions that are effective, adaptable, and scalable in the long run. Having access to a suite of software frameworks and optimizations that complement your hardware, enhances performance without spending extra time and effort for cross-platform integration. These qualities are especially important for tackling mixed AI workloads, which involve a combination of different types of AI tasks such as training, inference, and data processing.

AMD and Dell Technologies tackle mixed AI workloads through a combination of hardware and software solutions. AMD EPYC processors provide high-performance computing power, with features like simultaneous multithreading (SMT) and a high core count, enabling efficient parallel processing for AI workloads. These processors are optimized for AI tasks, offering strong performance for both training and inference workloads. Dell PowerEdge servers, equipped with AMD EPYC processors, provide a scalable and flexible platform for deploying AI workloads. Additionally, Dell OpenManage software suite offers management tools to optimize resource allocation and performance monitoring for mixed AI workloads.

AMD also offers the Unified Inference Frontend (UIF), which taps into the performance-enhanced versions of each of today's software stacks and draws upon the AMD ZenDNN library for AMD EPYC processors, the open-source AMD ROCm stack for AMD Instinct Accelerators, as well as a software stack for AMD adaptive SoCs. AMD ROCm is also designed to work with a wide range of AMD CPUs and accelerators, including both professional and consumer-grade products.

| Ensuring Explainability

Explainable AI plays a pivotal role in ensuring transparency, trustworthiness, and effectiveness in artificial intelligence applications. Explainable AI provides insights into how AI models make decisions, shedding light on the underlying factors and reasoning processes. This transparency is crucial for gaining stakeholders' trust, especially in sensitive domains like healthcare, finance, and criminal justice, where decisions directly impact individuals' lives.

Human-in-the-loop AI systems leverage human intelligence to enhance AI performance and mitigate algorithmic biases. By integrating human oversight, these systems can handle complex and ambiguous situations more effectively, ensuring that AI solutions align with ethical and social norms. Moreover, human involvement enables continuous refinement and adaptation of AI models based on real-world feedback, fostering iterative improvement and long-term reliability. These approaches are essential for building responsible, accountable, and inclusive AI systems that serve the best interests of society.

Real-World Scenarios

Scalers AI collaborated with Dell and AMD to showcase the capabilities of Dell PowerEdge servers equipped with AMD processors. Check out how these technologies are harnessed for training, transfer learning, and inference in retail and healthcare scenarios.

RETAIL

Scalers AI built the Retail Inventory Management Reference Solution, a system designed to monitor and manage stock levels on retail shelves through the implementation of an object detection AI model. This reference solution leverages the SSD_MobileNet_V2 model for identifying and recognizing products on store shelves, ultimately enabling automatic inventory counts and precise monitoring of stock levels. The model underwent transfer learning using the SKU110K image dataset, comprising 23,000 images from Roboflow. By leveraging computer vision and machine learning algorithms, the system can detect when items are running low or out of stock, providing alerts to store personnel for timely restocking or replenishment.

This solution utilizes the Dell PowerEdge R7615 server with the AMD EPYC 9354P 32-Core processor.

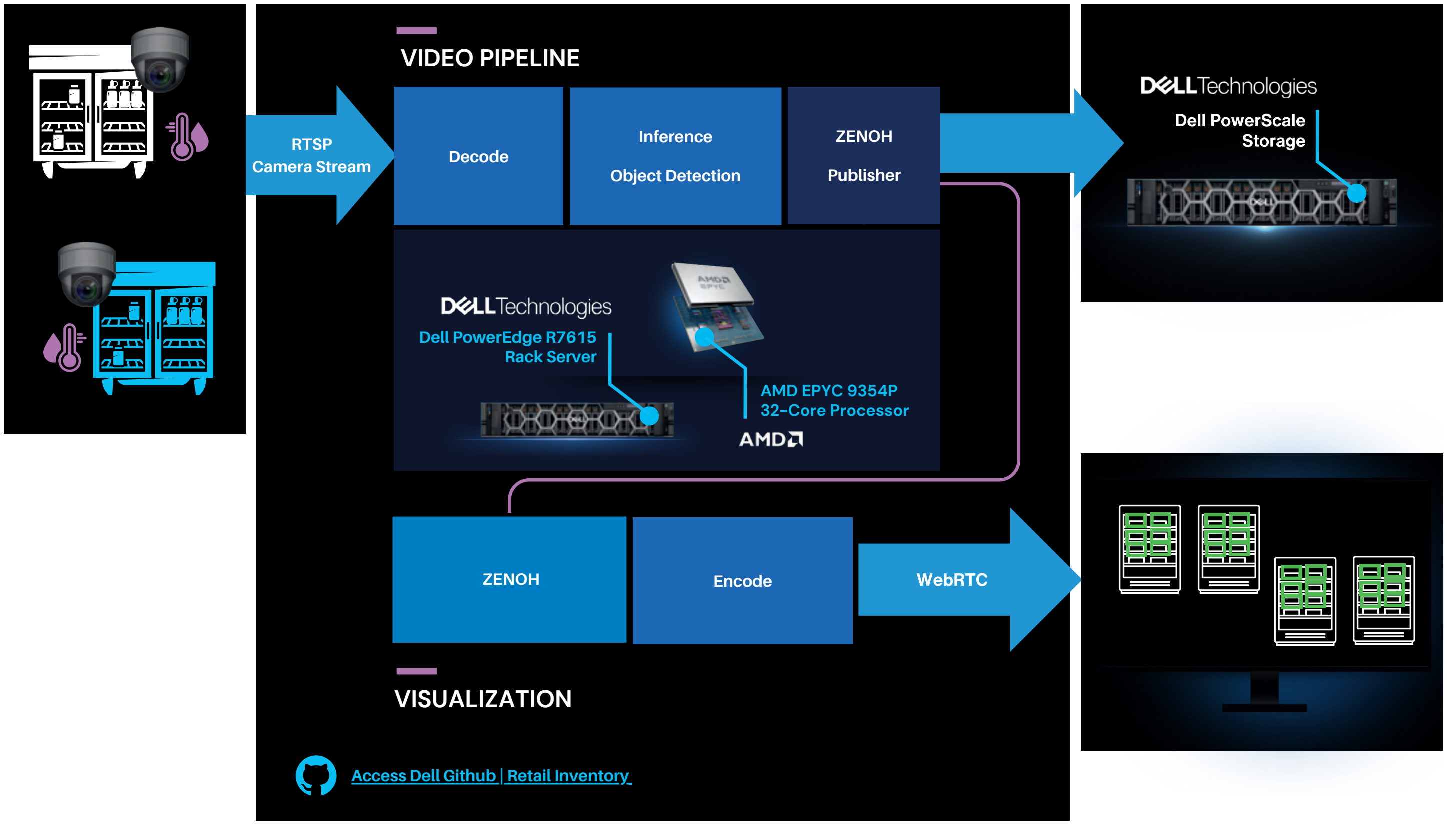


Figure 2: Architecture Diagram of the Retail Inventory Management Reference Solution

HEALTHCARE

AI-powered medical imaging is immensely valuable for its ability to enhance healthcare by improving diagnostic accuracy and efficiency and providing healthcare professionals with precise insights into conditions that may be difficult to detect with the naked eye. By automating the analysis of medical images, AI reduces the time required for diagnosis, enabling faster treatment decisions and ultimately improving patient outcomes.

Scalers AI harnessed the capabilities of the Dell PowerEdge R7625 server equipped with AMD EPYC 9554 64-Core processors to create an AI-powered medical imaging solution for pneumonia detection. Using advanced algorithms and machine learning techniques to analyze medical images, such as X-rays or CT scans, the solution helps increase the speed and accuracy of diagnosis of pneumonia in patients. Ultimately, this introduces an additional layer of computer assisted review, creating potential to assist healthcare professionals in handling large volumes of imaging data more efficiently.

This reference solution utilizes the ResNet50 model to analyze chest X-ray images obtained from the NIH Clinical Center dataset. Its primary objective is to detect the presence or absence of pneumonia, essentially performing a binary classification. The model was trained using the Xray DICOM dataset from the NIH Clinical Center dataset, involving transfer learning with the ResNet50 architecture.

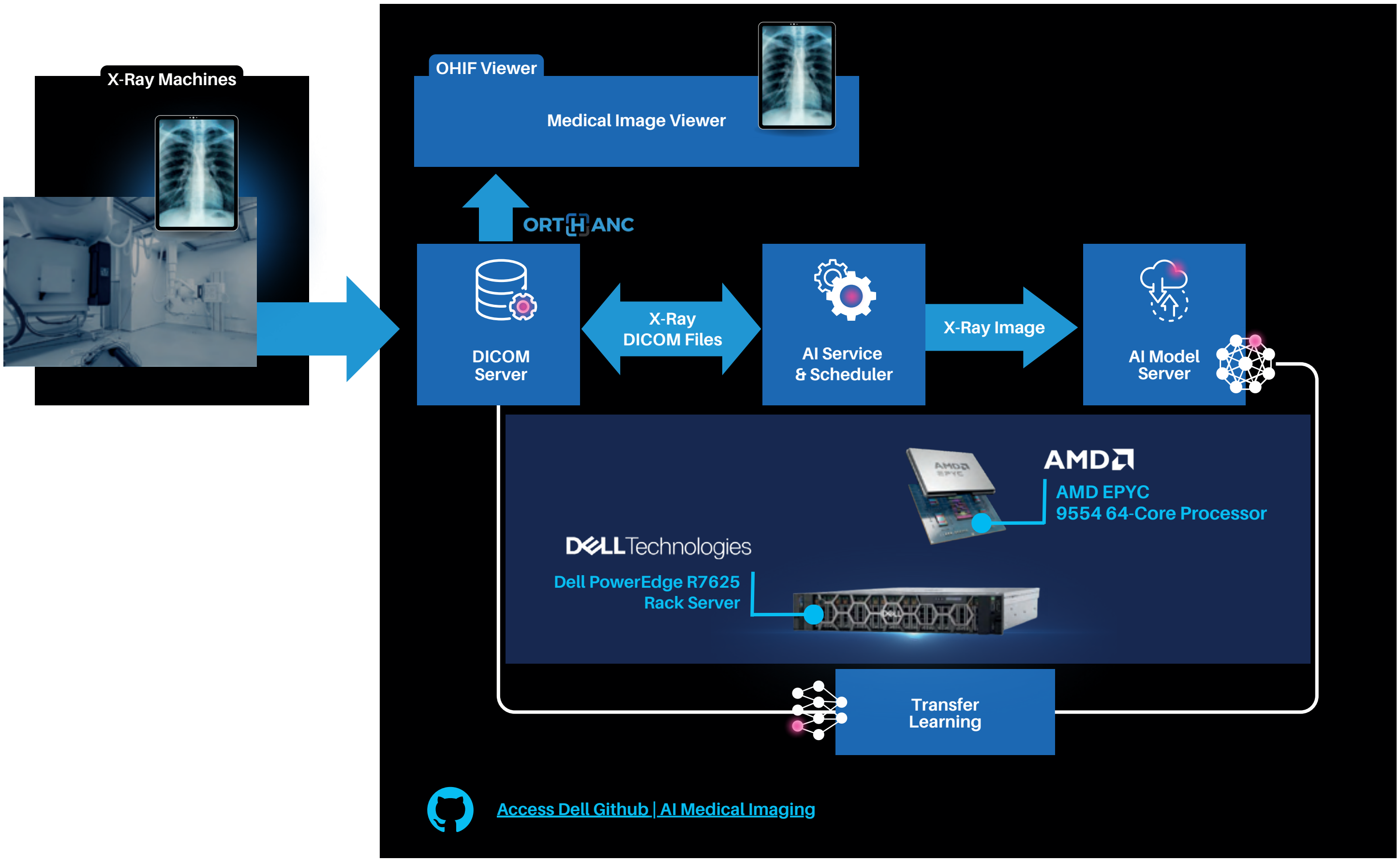
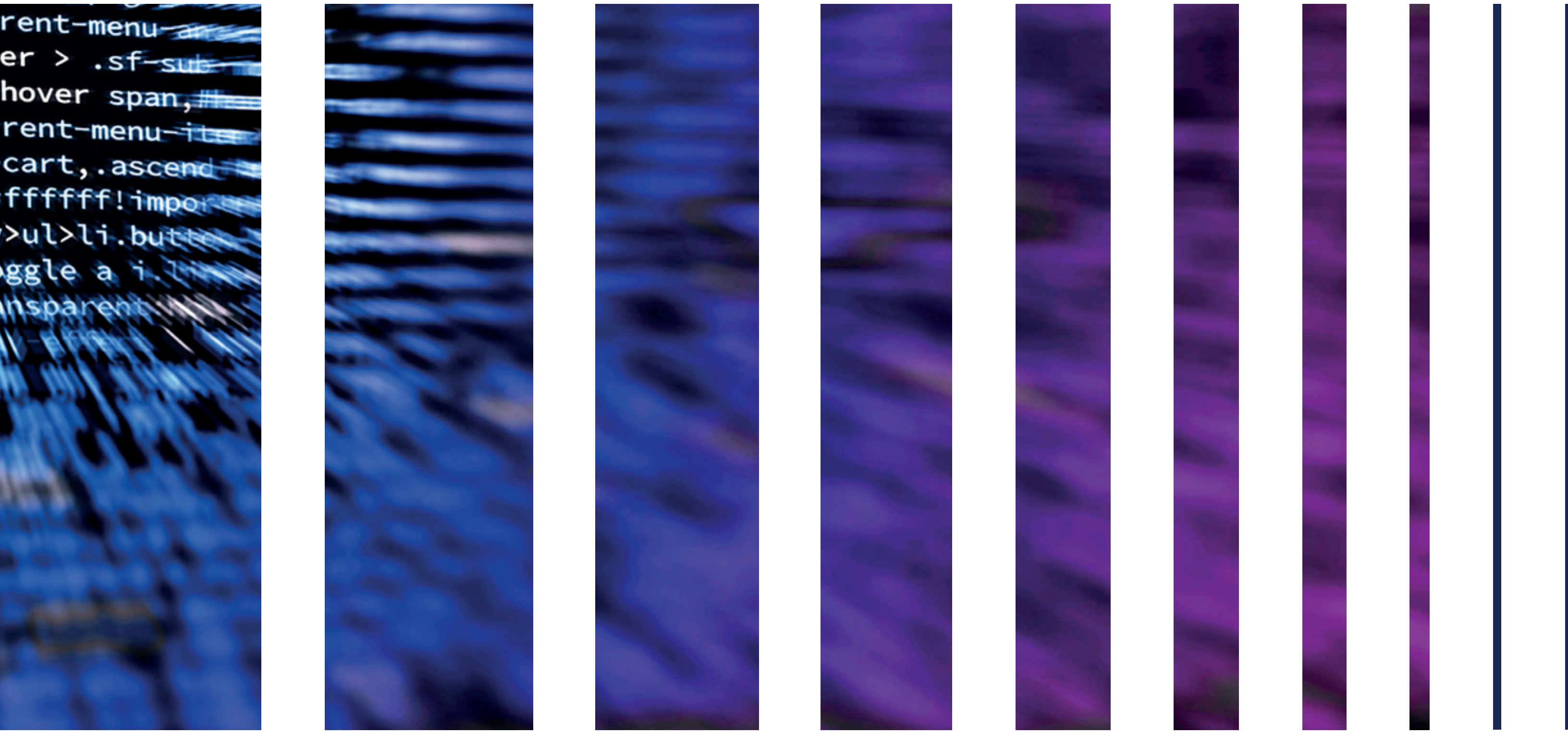


Figure 3: Architecture Diagram of the Medical AI Imaging Solution

Our Solutions

AI IS FOR EVERYONE: DELL & AMD DEMOCRATIZING AI

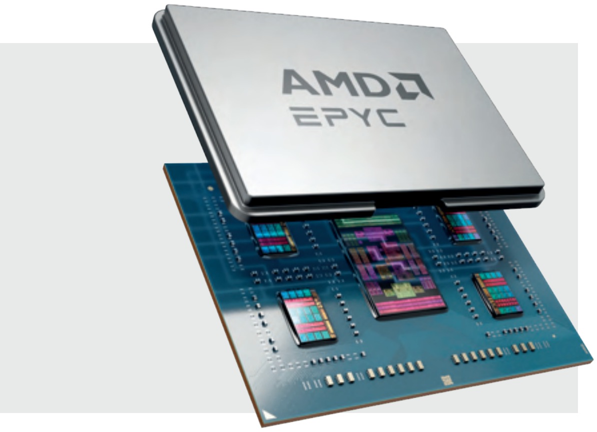
This collaboration lays the foundation for the democratization of AI, which is essential for fostering innovation and promoting inclusivity in the AI ecosystem. Dell and AMD are accomplishing this result by empowering individuals and organizations to leverage AI and solve unique challenges in their respective fields with an accessible suite of powerful servers equipped with state-of-the-art AMD CPU and accelerator technologies. Dell PowerEdge servers with the AMD Instinct MI300X accelerators are capable of handling large AI workloads such as training and fine-tuning large language models (LLMs), while Dell PowerEdge servers equipped with AMD EPYC processors excel in handling edge inference workloads. On top of the underlying hardware platform, AMD also offers the ZenDNN software library for the optimization of deep learning inference on AMD CPUs, as well as the AMD ROCm software library to improve training, fine-tuning, and inference capabilities on AMD Instinct Accelerators. All of these options are seamlessly tied together in AMD's Unified Inferencing Model (UIF), through which users can construct end-to-end AI solutions, with flexibility in choice of software frameworks, software optimizations, and hardware platform choices.



HUGGING FACE COLLABORATION

Businesses eager to adopt AI can start by leveraging pre-existing models or AI workflows tailored for their specific needs directly from Hugging Face, an open-source platform dedicated to data science and machine learning. AMD has entered a collaboration with Hugging Face, with the shared aim of delivering top-notch transformer performance by adding AMD-specific software optimizations to software libraries and frameworks that already integrate seamlessly with AMD platforms. Hugging Face is actively collaborating with the engineering team at AMD to optimize key models for peak performance, incorporating AMD ROCm into the Transformers library, and improving Optimum-AMD, a library specifically designed for AMD platforms, to assist Hugging Face users in utilizing them with minimal code changes.

Dell Technologies has also recently joined forces with Hugging Face to simplify the process for enterprises to develop, fine-tune, and apply their own open-source generative AI (Gen AI) models using the Hugging Face community, all on industry-leading Dell infrastructure products and services. A new Dell portal is being developed on the Hugging Face platform, which will include custom, dedicated containers and scripts to aid users in securely and effortlessly deploying open-source models available on Hugging Face using Dell's servers and data storage systems. Businesses can now take full advantage of Hugging Face resources to directly deploy models on Dell PowerEdge servers with AMD processors and construct end-to-end AI solutions using their own proprietary data.

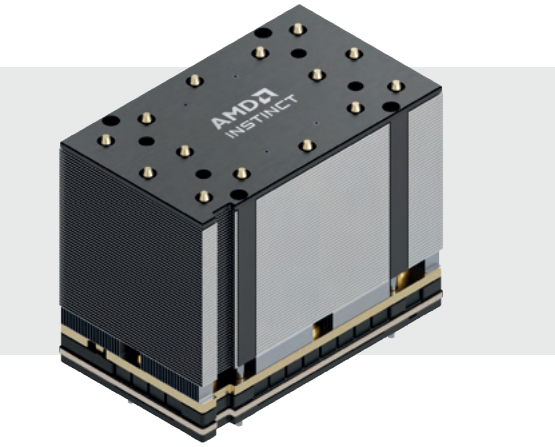


AMD EPYC PROCESSORS

AMD provides the technological advancements necessary for modern cloud-based datacenters through their AMD EPYC processors. These processors are a system on chip (SoC) designed from scratch to efficiently address the demands of current and future data centers. The AMD EPYC 9000 Series processors equip the datacenter with up to 128 cores, 256 threads, 12 memory channels that support up to 6 TB of memory per socket, and 128 PCIe Gen5 lanes. This is paired with the industry's pioneering hardware-embedded x86 server security solution. By integrating essential compute, memory, I/O, and security resources into the SoC, AMD EPYC processors yield top-tier performance and facilitate a lower Total Cost of Ownership (TCO).

AMD INSTINCT MI300X ACCELERATORS

The AMD Instinct MI300X accelerator, built on the cutting-edge AMD CDNA 3 architecture, offers industry-leading efficiency and performance for the most intensive AI and HPC applications. It is equipped with 304 high-performance compute units and features AI-specific functions such as support for new data types and photo and video decoding, as well as an unparalleled 192 GB of HBM3 memory on a single accelerator.

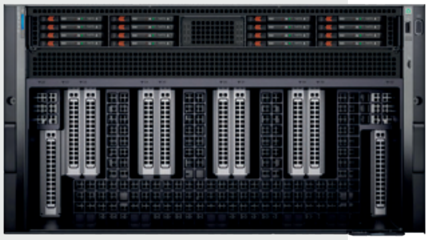


AMD ROCM 6 OPEN-SOURCE SOFTWARE PLATFORM

The AMD ROCm 6 open-source software platform is optimized to maximize high-performance computing (HPC) and AI workload performance from AMD Instinct MI300X accelerators. It also extends support for AMD Instinct MI300X accelerators, ensuring compatibility with industry software frameworks. The AMD ROCm platform encapsulates a variety of drivers, development tools, and APIs that facilitate accelerator programming from the kernel level to end-user applications, and can be tailored to align with your specific requirements. AMD ROCm is especially adept for applications in high-performance computing (HPC), artificial intelligence (AI), and scientific computing. Additionally, the AMD ROCm platform offers support for multi-accelerator computing, including remote direct memory access (RDMA) for server-node communication.

AMD ROCm

DELL POWEREDGE SERVERS PORTFOLIO



Dell's investment in AMD creates a critical choice in the market to democratize AI, as evidenced by their four server platforms with EPYC and their flagship Dell PowerEdge XE9680 rack server with AMD Instinct MI300X accelerators. The latest generation of Dell PowerEdge servers powered by AMD EPYC processors enhance both business agility and time to market, with the ability to support transformative workloads such as databases and analytics, virtualization, software-defined storage, virtual desktop infrastructure (VDI), containerization, high performance computing (HPC), AI, and Machine Learning (ML). Their one-socket (single CPU) rack servers provide a cost-efficient balance of performance and storage capacity, designed to grow seamlessly with your business, while their two-socket (dual CPU) rack servers accommodate more demanding workloads with a wide array of features.

The Dell PowerEdge XE9680 rack server is a robust data-processing machine tailored specifically for AI tasks. It supports eight accelerators, making it perfect for machine learning (ML) / deep learning (DL) training and inference workloads, particularly for training Large Language Models (LLMs). Equipped with eight MI300X accelerators, each with 192GB of 5.3 TB/s High Bandwidth Memory (HBM3), leading to a total HBM3 capacity of 1.5 TB per server and over 21 petaflops of FP16 performance, the Dell PowerEdge XE9680 rack server with AMD Instinct MI300X accelerators is ready to further extend Gen AI accessibility to enterprises. This enables them to train larger models, minimize data center footprints, reduce TCO, and gain a competitive advantage.

Summary

The rapid pace of innovation fueled by AI is revolutionizing data center workloads faster than any other technological transformation. To support these technological advancements, Dell and AMD are working towards a more inclusive, innovative, and ethically developed AI ecosystem that encourages developers from all industries to collaborate on open source resources and drive the Gen AI innovation of today. Whether your AI solution achieves your performance requirements on AMD EPYC processors or on servers powered with AMD Instinct Accelerators, we provide the flexibility to run your AI workload across our hardware platforms, allowing you to leverage the best of what Dell and AMD have to offer.

REFERENCES

AMD images: AMD.com, AMD Partner Resource Library, **https://www.amd.com/en/partner/resources/resource-library.html**

Dell images: **Dell.com**

Learn more about Dell servers solutions

Contact a Dell Technologies Expert

View more resources

Join the conversation with #PowerEdge