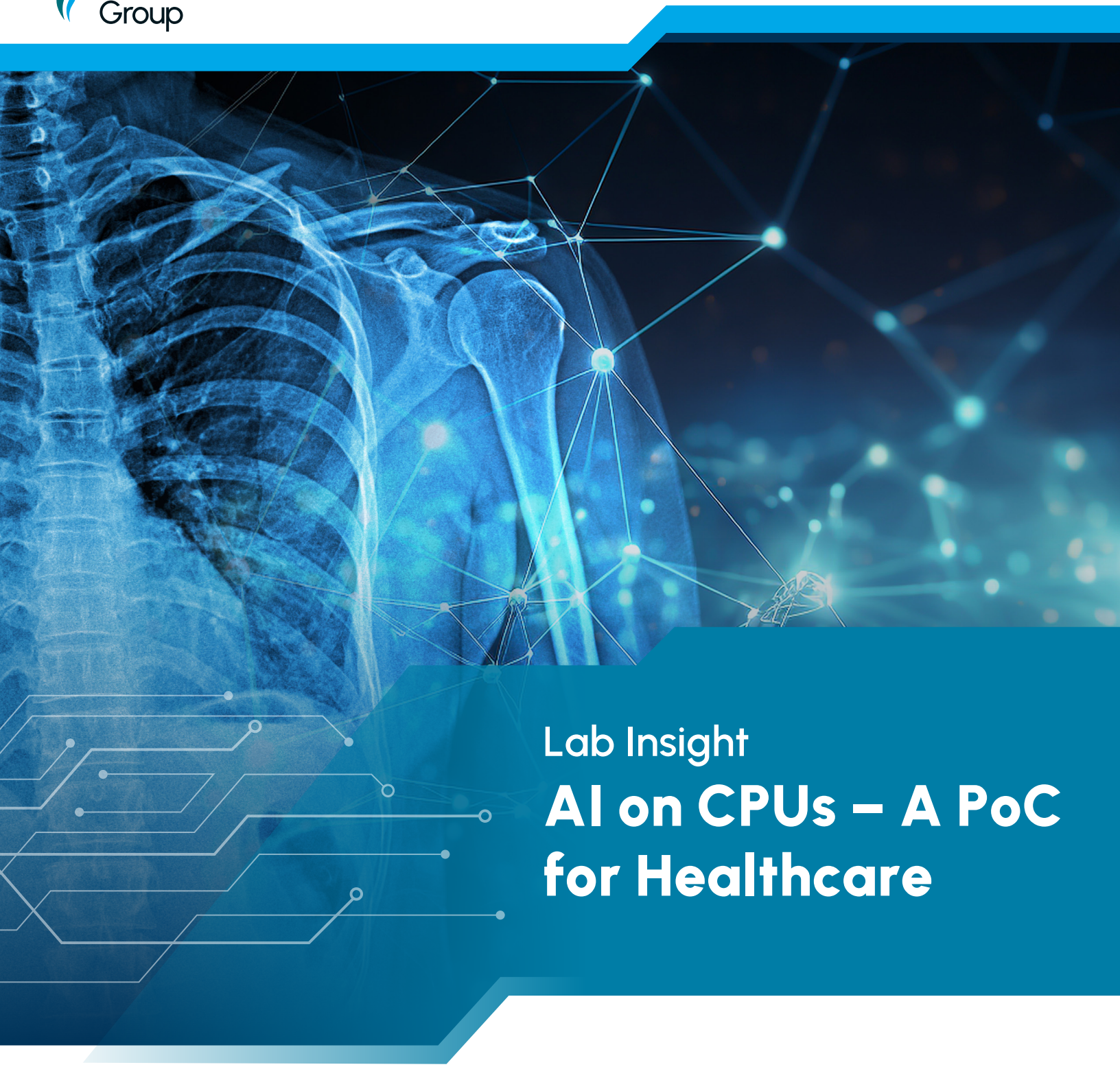
The Futurum Group



Lab Insight

AI on CPUs – A PoC for Healthcare

AUTHOR **Mitch Lewis**
Research Analyst | The Futurum Group

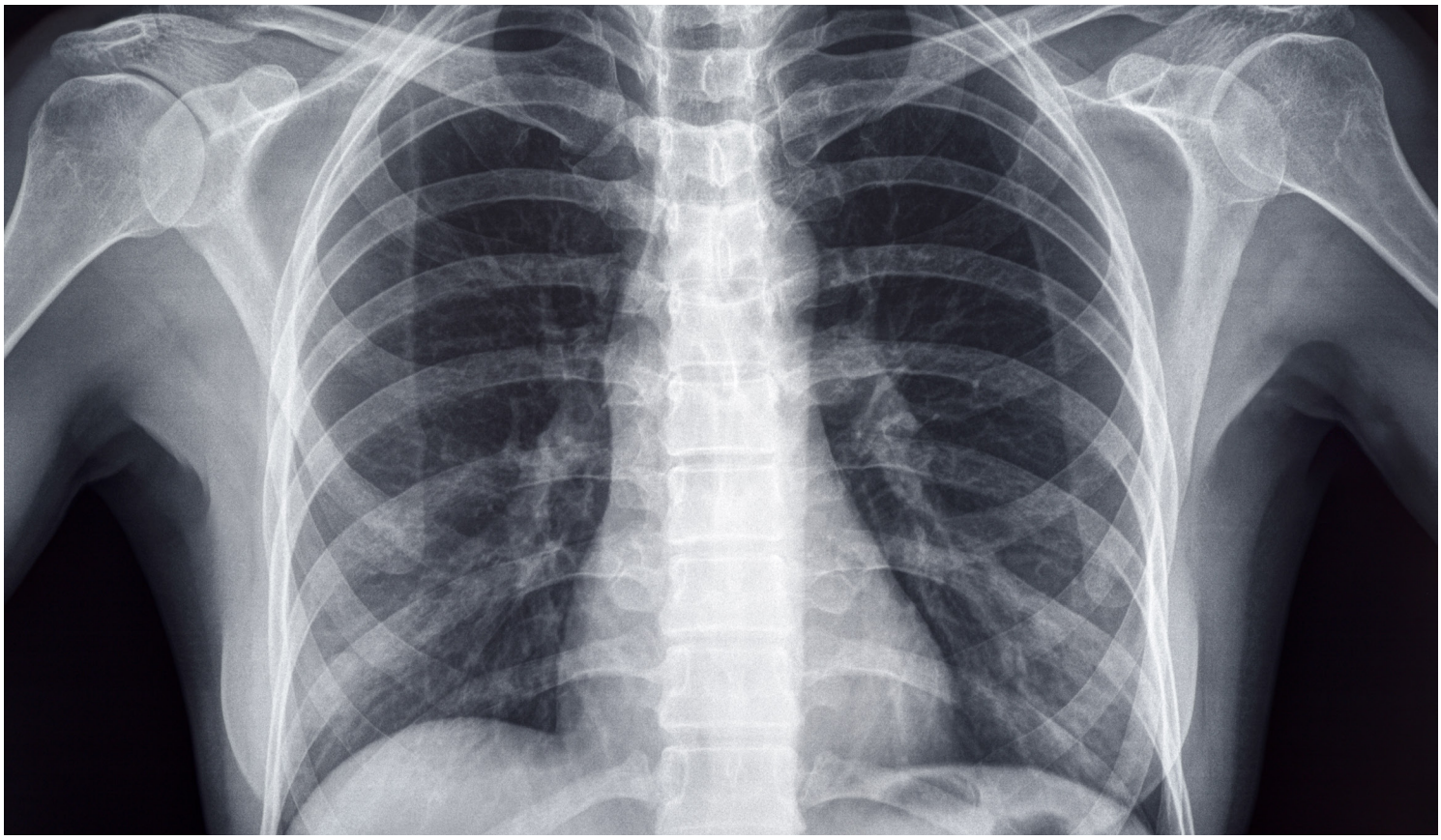
APRIL 2024

Introduction

Recent advances in AI have significantly accelerated interest in the technology and how it can be applied to revolutionize processes across many different industries. One such industry that is well positioned to benefit from leveraging AI is healthcare. AI is increasingly being used to assist in medical diagnostics, specifically to improve the accuracy and speed of diagnoses made by radiologists and physicians. This paper presents a proof-of-concept (PoC) solution that utilizes an AI image classification model to quickly and precisely detect pneumonia from patient X-ray images.

The PoC shows the practicality of bringing AI into image analysis for healthcare, and sets the stage for healthcare organizations to quickly adopt and deploy AI solutions. Building on the success of the pneumonia detection PoC, the approach can be further extended to modalities such as CT-Scans, MRIs, and others. The PoC overcomes common challenges found both generally in new AI deployments and more specifically in healthcare environments by leveraging and optimizing common CPU-based hardware, customizing a model for a healthcare-specific use case, and deploying a secure, on-premises solution to address healthcare data regulations and privacy concerns.

The PoC was deployed on standard Dell™ PowerEdge hardware with 64 core 4th Gen AMD™ EPYC CPUs. The PoC demonstrated impressive performance for both model training and inferencing, without requiring GPUs. Model training was completed in 9 hours with a validation accuracy of 85%. The inferencing process achieved a throughput of 337 images per second with a default configuration. By utilizing additional performance optimizations, the deployment achieved a 4X performance increase, resulting in a throughput of 1,390 images per second.



© 2024 The Futurum Group. All rights reserved.

Importance for the Healthcare Market

Technology advances in healthcare drive greater efficiency and accuracy in medical processes and ultimately improve patient outcomes. Such technology advancements are important to moving the healthcare industry forward.

AI in particular is a technology with great potential to create significant value in the healthcare sector. The possibilities for applying AI technology to healthcare may extend to a wide range of healthcare related areas including pharmaceutical research, hospital workflows, and patient experience. Potential uses for AI in healthcare include AI accelerated drug development, predictive analytics for disease prevention, medical focused chatbots, and intelligent hospital staffing systems.

Key to adoption of any technology in healthcare is maintaining data privacy and security due to the handling of sensitive patient information and the heavily regulated nature of the industry. When considering AI, organizations will address this challenge by implementing on-premises solution deployments to maintain control over their private data.

New AI applications must also be capable of integrating with existing technologies, processes, and equipment common to healthcare. Additionally, to adopt AI quickly, healthcare organizations will want to leverage existing hardware or utilize readily available commodity hardware. These challenges may cause uncertainty for organizations who are unfamiliar with AI when planning new deployments, delaying adoption of AI innovation.

The PoC presented in this paper demonstrates an AI implementation that addresses these challenges to assist medical organizations in swift adoption of AI. The PoC serves as a template for using AI to improve diagnoses based on X-ray images. In addition to the benefit of rapid image analysis, the PoC highlights the ability to quickly train an accurate AI model by utilizing a healthcare specific dataset. Notably, the same Dell PowerEdge server was used for both training and inference.

With continued advancements in the field of AI, healthcare executives will recognize the opportunity the technology holds in improving healthcare services and look for ways to leverage it. Developers and IT operations must understand the systems, processes, and effort required for a successful AI deployment, especially when considering vertical specific applications, such as healthcare.

© 2024 The Futurum Group. All rights reserved.

Solution Overview

To demonstrate a practical implementation of a healthcare focused AI solution, Scalers AI™, in partnership with Dell, Broadcom™, and The Futurum Group, implemented an AI-powered system for detecting pneumonia. The solution utilizes the ResNet50 image classification AI model that was fine tuned to recognize pneumonia in X-ray images.

This PoC showcases the ability for AI to assist doctors in patient diagnoses. Relying solely on human evaluation of X-ray results can lead to delays, due to doctors' availability and bandwidth, or misdiagnoses due to human error. Delayed diagnosis, in particular, is a significant issue in healthcare, and it has been found to be a leading cause of patient injury claims concerning medical imaging[1].

These issues are becoming increasingly challenging as the number of patients requiring medical imaging is growing. Meanwhile, hospitals are facing a global shortage of radiologists in the workforce[2]. While AI does not possess the medical expertise required to replace human doctors or radiologists, it provides valuable characteristics that can be leveraged by medical professionals to enhance the efficiency and accuracy of the diagnosis process.

AI image classification models can rapidly detect specific features within images, such as the presence of pneumonia in chest X-rays, and classify them with a high level of accuracy. This approach can be used to quickly identify issues within large amounts of medical images, and in some cases identify issues that may otherwise go undetected. AI provides an efficient and accurate initial classification of images, which will be further analyzed by medical professionals to make a final diagnosis. By leveraging AI to augment the diagnosis processes, medical professionals can provide faster and more accurate diagnoses, ultimately resulting in quicker time to treatment and improved patient results.

Solution Highlights

- AI-powered medical imaging solution to detect pneumonia cases in X-Rays.
- CPU-based AI solution achieved 1,390 images per second throughput.
- Integrated AI pipeline with DICOM server.
- Scalable architecture built on Dell PowerEdge servers connected with high bandwidth Broadcom Ethernet.

1 . Tarkiainen, T., Turpeinen, M., Haapea, M. et al. Investigating errors in medical imaging: medical malpractice cases in Finland. Insights Imaging 12, 86 (2021). https://doi.org/10.1186/s13244-021-01011-8
2 . Radiological Society of North America. Radiology facing a global shortage. https://www.rsna.org/news/2022/may/global-radiologist-shortage.

© 2024 The Futurum Group. All rights reserved.

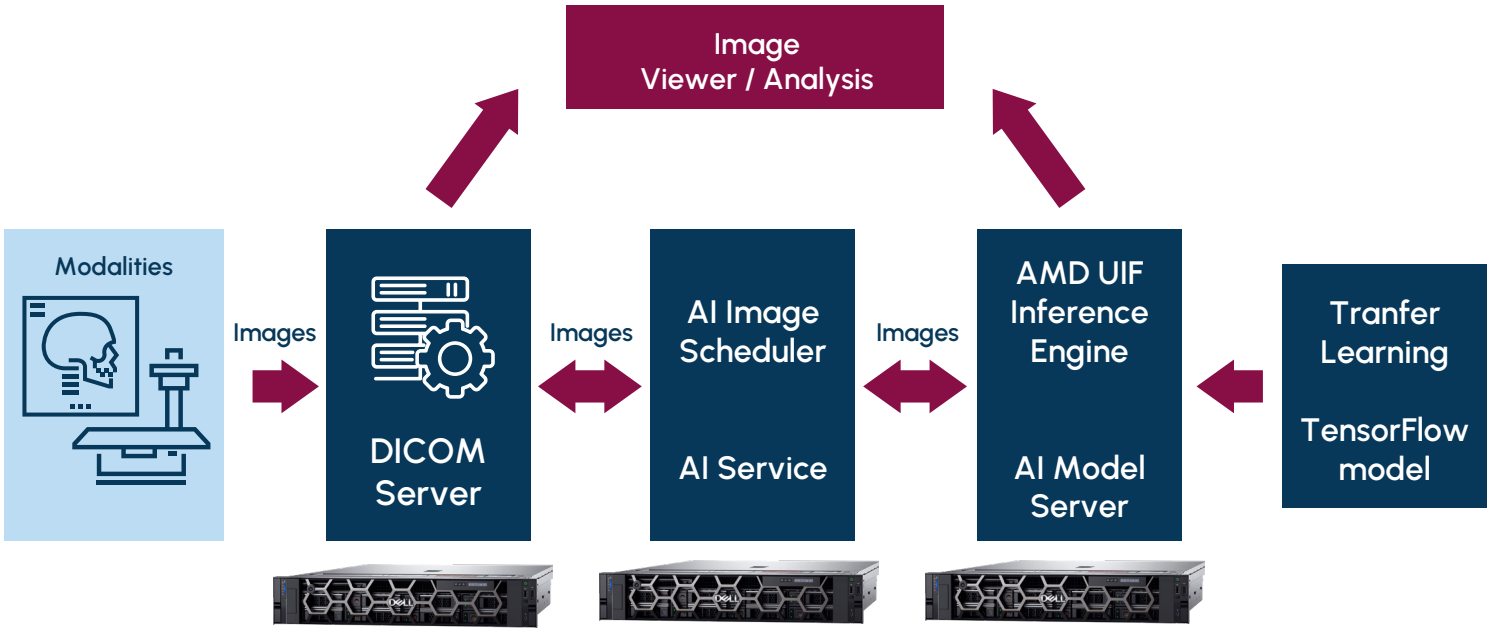


Figure 1: AI Medical Image Analysis PoC Solution (Source: Scalers.AI)

To achieve an AI-powered pneumonia detection system, the PoC integrates an image classification AI pipeline with a standard DICOM server for storing and managing medical imaging data. The DICOM server in the PoC stores X-ray images of potential pneumonia cases. The AI pipeline evaluating the X-ray images consists of two additional components – an AI scheduling service and an inferencing server. The AI scheduling service identifies new images, batches them, and sends them to the inferencing server. The inferencing server utilizes a customized version of the ResNet50 AI model, deployed using AMD's Unified Inferencing Frontend (UIF). X-ray images are inferenced to provide a binary classification regarding the detection of pneumonia. The categorized images are then made available for review with a medical imaging viewer, and returned to the DICOM server.

More information about the specific solution components can be found below:

- **DICOM Server and Storage:** For the PoC, Orthanc open source DICOM server software was used to manage the medical images.
- **AI Service and Scheduler:** The AI service and scheduler identify new images, groups them for individual patient evaluation, and send them over Ethernet for analysis. The AI Service then handles the analysis results and sends annotated images back to the DICOM server.
- **AI Model Server:** The AMD UIF inference engine on the AI model server is an open-source tool that deploys models and makes them available for inferencing. The server is compatible with specific models designed for AMD CPUs.
- **AI Pneumonia Classification Model:** The ResNet50 model serves as the foundation for the AI image classification. The model was customized through a transfer learning process using chest X-ray images from the NIH Clinical Center. Transfer learning involves using pre-trained models as a starting point for creating new models, which are further trained for different tasks. In this case, ResNet50 was used to train the model to recognize pneumonia in DICOM images. ResNet50 was chosen for its support of the AMD UIF with ZenDNN Model Zoo. Additionally, ResNet50 is well suited for image detection in critical medical use cases as it is a highly accurate model.

© 2024 The Futurum Group. All rights reserved.

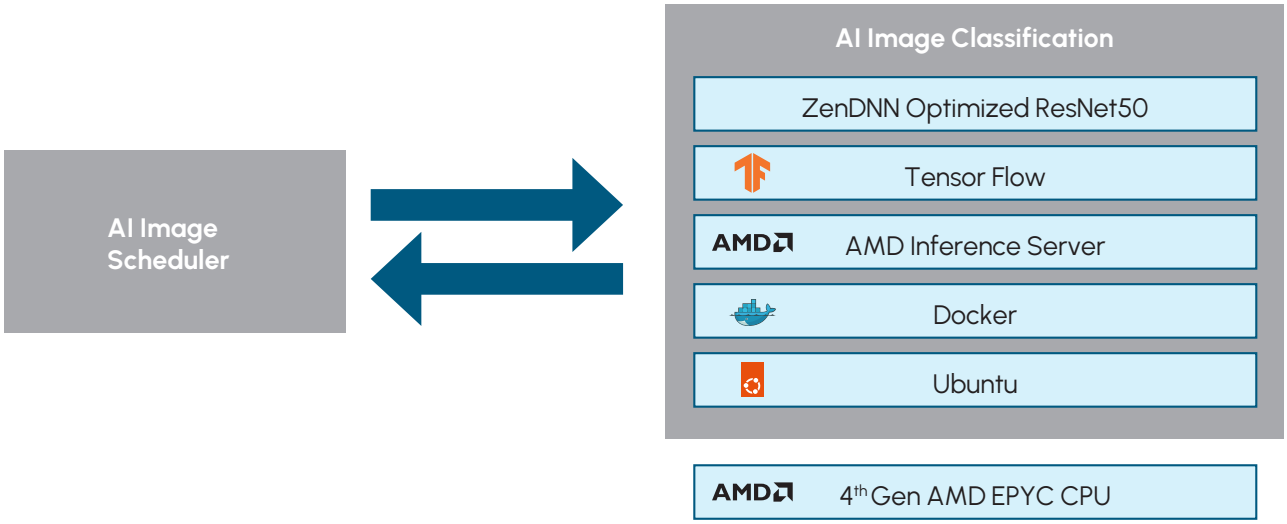


Figure 2: AI Image Classification Software Overview

The AMD ZenDNN library is used to provide performance optimizations. ZenDNN is a library with APIs designed to accelerate deep learning inference applications on AMD CPUs, aiming to improve performance. ZenDNN performance guide recommendations, along with node pinning and core pinning, were used to optimize the performance of AMD EPYC processors used in the PoC.

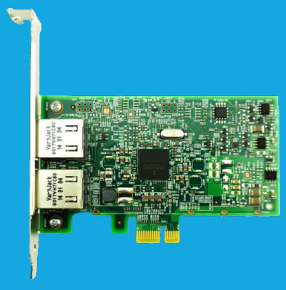
- **Medical Image Viewer:** The PoC used an Open Health Imaging Foundation (OHIF) viewer for viewing the DICOM images. Although radiologists might have distinct preferences for viewers, the PoC specifically used the OHIF viewer as a it is an open source, web-based medical imaging viewer, however, other DICOM viewers could be used.

Additional details about the implementation and performance testing of the PoC have been made available by Dell on GitHub.

The key hardware components used in the solution included the following:

Dell PowerEdge R7625 servers

AMD EPYC 9554 64-core processors



Broadcom NetXtreme NICs

© 2024 The Futurum Group. All rights reserved.

Highlights for AI Practitioners

This PoC is notable for AI practitioners as it demonstrates a practical AI application that can be used to enhance healthcare environments. Key to the practicality of the solution is that it utilizes readily available, CPU based hardware, rather than relying on GPUs. A core component of achieving this type of CPU-based approach is utilizing software libraries to simplify and optimize the deployment. The AMD Unified Inference Frontend (UIF) was utilized to easily deploy a model that was optimized to run on AMD EPYC CPUs. While this PoC intentionally utilized a CPU-based deployment to demonstrate running useful AI applications on easily accessible hardware, the use of a model from the UIF model zoo is notable, as the UIF models are transportable across AMD technology stacks. This provides flexibility for organizations who may incorporate GPUs in future deployments as they further expand their use of AI.

AI practitioners should additionally note the performance enhancements that were achieved by utilizing the ZenDNN library, along with core pinning and node pinning configurations. These configurations demonstrated up to a 4X throughput increase, showcasing how the use of software optimization libraries can be leveraged to provide significant inferencing performance without hardware acceleration. Figure 3 shows the ZenDNN parameter configurations utilized.

Variable	Value	Notes
TF_ENABLE_ZENDNN_OPTS	0	Sets native TensorFlow code path
ZENDNN_CONV_ALGO	3	Direct convolution algorithm with blocked inputs and filters
ZENDNN_TF_CONV_ADD_FUSION_SAFE	1	Modified to 1 to enable Conv, Add Fusion.
ZENDNN_TENSOR_POOL_LIMIT	512	Set to 512 to optimize for Convolutional Neural Network
OMP_NUM_THREADS	128	Sets threads to 128 to match # of cores

Figure 3: ZenDNN Configurations

AI practitioners should note that the CPU-based deployment was not only utilized for inferencing. The same Dell PowerEdge server and AMD processor was used for model training. The solution utilizes a pre-trained base model, ResNet50, customized with a transfer learning process. Transfer learning utilizes the foundation of a pre-trained model's capabilities, and provides further customization to support a new, specific task. In this case, transfer learning was used to teach the ResNet50 image classification model to detect pneumonia in X-ray images. This was done by training the model with a dataset of 29,687 X-ray images. The total training process was completed in around 9 hours, and resulted in 99% training accuracy and 85% validation accuracy. The accuracy of the model is especially critical in this type of medical deployment, as the model is responsible for assisting in the diagnosis of medical patients, and can have a direct impact on patient outcomes. The PoC demonstrates the ability to utilize common CPU-based infrastructure along with pre-trained models for efficient, yet accurate AI model training.

Key Highlights for AI Practitioners

- AI powered retail solution deployed on standard Dell PowerEdge hardware with AMD EPYC processors. No GPUs were required.
- AMD UIF utilized to deploy model optimized for AMD EPYC CPUs. ZenDNN delivered further performance optimizations, resulting in 4X increased throughput.
- Transfer Learning process provides customization of pre-trained ResNet50 model. Training was completed in 9 hours with 99% training accuracy and 85% validation accuracy.

© 2024 The Futurum Group. All rights reserved.

Considerations for IT Operations

This AI implementation is notable for those working in IT operations because it demonstrates an achievable AI deployment that utilizes familiar, readily available hardware for both model training and production. IT operations staff will be very familiar with deploying Dell PowerEdge servers and Broadcom networking, and this PoC provides an example for organizations to understand how these familiar solutions can be leveraged for AI workloads.

The PoC leverages three Dell PowerEdge servers powered by 4th Gen AMD EPYC CPUs to deploy the Orthanc DICOM server, the AI scheduler, and the AI model server. The powerful AMD processors alongside large memory capacity make these servers well suited for AI workloads. This PoC leveraged Dell PowerEdge R7625 servers with AMD EPYC 9554 64-core processors, and 2.95 TB of memory. Additional server specifications can be found in figure 4 below.

PowerEdge R7625		
Device Name		Dell PowerEdge R7625
CPU	Model Name	AMD EPYC 9554 64-Core Processor
	Speed Max (MHz)	3762
	Number Of Cores	64
	Number Of Sockets	2
Memory	Size	3.07 TB
	DIMMs	24 x 128 GB
Storage	Size	1.4 TB
Network		Broadcom NetXtreme BCM5720
		Broadcom NetXtreme BCM57414
OS	Name	Ubuntu 22.04.3 LTS
	Kernel	5.15.0-87-generic

Figure 4: Server Details

The Dell PowerEdge R7625 server provides a powerful platform that showcases the ability to run AI on CPUs. For IT operations, this lowers the barrier of entry for supporting AI, allowing them to utilize readily available hardware or leverage their existing infrastructure.

Another notable takeaway of the PoC is its ability to maintain data privacy and security, which are major concerns for IT organizations in the healthcare sector, due to the sensitive nature of medical data and regulations such as HIPAA and HITECH. Dell PowerEdge servers feature a cyber resilient architecture for zero trust IT environments with capabilities such as silicon-based root of trust, multi-factor authentication (MFA), and role-based access controls (RBAC).

The DICOM server, the AI scheduler, and the AI model server are connected with scalable, high bandwidth, Broadcom Ethernet. This high bandwidth connection is crucial to the solution's ability to support the transmission of medical images, especially as the solution scales. While this PoC demonstrated image classification capabilities using relatively small X-ray images, by implementing a scalable connection, the PoC can be further extended to support larger image files such as MRIs or CT scans.

© 2024 The Futurum Group. All rights reserved.

In addition to providing insight into AI hardware requirements, the PoC provides IT professionals with an understanding of software packages that can be utilized to build a healthcare focused AI solution. The PoC primarily utilized easily accessible, open-source software tools.

Key to deploying the AI model is the AMD Inference Server, which provides an open-source tool to easily deploy AI solutions on AMD hardware. The PoC additionally utilized open-source tools to support the medical imagery workflow, include Orthanc DICOM server and OHIF Viewer. Details of key software utilized, including version and licensing information can be found in figure 5 below.

Component	Description	Version	License
AMD Inference Server	Open-source tool to deploy machine learning tools on AMD hardware.	0.4.0	Apache License 2.0
Orthanc	Open-source, lightweight DICOM server.	1.12.1	GNU General Public License v3.0
OHIF Viewer	Open-source medical image viewer from Open Health Imaging Foundation.	v4.12.51.21579	MIT License
pydicom	Python package for reading and writing DICOM data.	2.4.3	MIT License
requests	Python package for sending HTTP requests.	2.31.0	Apache License 2.0
schedule	Python package for job scheduling.	1.2.1	MIT License
pillow	Python Imaging Library for image processing.	10.0.1	HPND License
pyyaml	YAML processing framework for Python.	6.0.1	MIT

Figure 5: Software Packages

Key Highlights for IT Operations

- AI solution deployed on familiar, readily available Dell PowerEdge hardware.
- High bandwidth Ethernet supports transfer of medical imagery.
- Secure deployment designed with data privacy and medical regulations in mind.

© 2024 The Futurum Group. All rights reserved.

Healthcare Solution Performance Observations

Key to the performance of this PoC is the throughput of images per second. Quick processing of X-ray images is vital to the solutions overall ability to accelerate patient diagnosis, leading to quicker treatment.

To demonstrate the performance of the PoC, the throughput of images per second were measured with an increasing number of processes, both with default settings as well as with configurations that optimized the performance of the AMD EPYC processor. The optimized variations included the use of the ZenDNN library alongside use of core pinning and node pinning. ZenDNN is a library that optimizes deep learning inferencing for AMD processors. Core pinning and node pinning are configurations that bind processes to specific cores or NUMA nodes. The performance of each configuration can be seen in Figure 6.

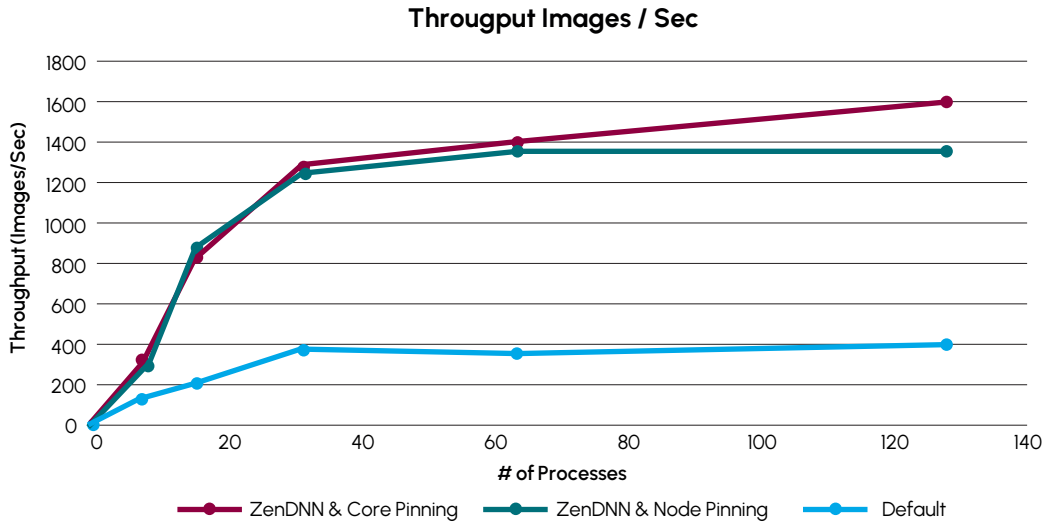


Figure 6: Throughput Performance

The test results demonstrate the ability to significantly improve throughput performance by utilizing ZenDNN with either core pinning or node pinning. When running 64 processes the default configuration achieved a throughput of 337 images per second. Meanwhile, the configuration with ZenDNN and node pinning achieved a 3.9x improvement with 1,338 images per second, and the configuration with ZenDNN and core pinning configuration achieved a 4.1X improvement with 1,390 images per second. Figure 7 includes full testing results of the pneumonia detection PoC throughput performance testing.

Processes	Throughput Images/sec – ZenDNN			Throughput Images/sec – ZenDNN OFF		
	Core Pinning	Node pinning	CPU utilization	Default	CPU utilization	Difference ZenDNN vs Native
1	34.05	37.9	4.85	22.41	7.233333333	1.69
8	281.77	306.51	40.01770833	127.45	45.109375	2.41
16	797.75	845.95	54.74583333	212.65	58.59479167	3.98
32	1282.96	1231.3	78.97604167	355.71	81.08958333	3.46
64	1,390.85	1,337.61	89.60026042	337.09	86.09674479	3.97
128	1,574.28	1,309.06	91.7375651	363.49	87.89980469	3.6

Figure 7: Throughput Testing Results

© 2024 The Futurum Group. All rights reserved.

For most hospitals, 1,390 images per second is likely well beyond their typical X-ray image processing requirements. This level of throughput is notable, however, because it provides flexibility for future adaptation of the solution to support more demanding data such as 3D images or other large data formats.

The performance improvements achieved by the ZenDNN configurations are also quite notable because they demonstrate the ability to optimize AI inferencing performance on a CPU. AI performance is often thought of as a hardware problem that requires GPUs or other specialized hardware to solve. This testing showcases the impact that software libraries, such as ZenDNN, can have in dramatically improving performance, even when using off-the-shelf CPU-based hardware. This type of optimization allows organizations to deploy powerful, high performance AI applications with either their existing hardware or readily available hardware, removing the barrier of acquiring GPUs and facilitating quick AI innovation.

Final Thoughts

Strategically applying AI in healthcare has great potential to enhance medical processes and improve patient outcomes, as demonstrated in this successful pneumonia detection PoC. While the PoC example is focused specifically on pneumonia detection from X-ray images, the solution can be further expanded to analyze patient data from various modalities, allowing trained models to detect a wider range of conditions. The potential for AI to enhance the healthcare industry extends far beyond this type of AI-assisted diagnosis use case.

This PoC demonstrates an AI deployment that utilized off-the-shelf, CPU-based hardware, while providing impressive performance, and meeting the unique requirements of a medical-focused application. The results of this PoC, including the performance details, not only demonstrate a successful example of a healthcare-oriented AI application, but it also emphasizes the broader opportunity for AI to have an immediate impact on improving healthcare processes. AI will prove to be an innovative technology across many areas in healthcare, and healthcare providers should be motivated to adopt the technology quickly, both to maintain competitive advantage in the market and to improve overall patient treatment. By leveraging readily available hardware from Dell and Broadcom, along with the concepts demonstrated in this PoC, healthcare organizations can quickly deploy powerful, innovative new AI solutions.

© 2024 The Futurum Group. All rights reserved.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis
Research Analyst | The Futurum Group

PUBLISHER

Daniel Newman
CEO | The Futurum Group

INQUIRIES

Contact us if you would like to discuss this report and The Futurum Group will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "The Futurum Group." Non-press and non-analysts must receive prior written permission by The Futurum Group for any citations.

LICENSING

This document, including any supporting materials, is owned by The Futurum Group. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of The Futurum Group.

DISCLOSURES

The Futurum Group provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT THE FUTURUM GROUP

The Futurum Group is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION

The Futurum Group LLC | futurumgroup.com | (833) 722-5337

© 2024 The Futurum Group. All rights reserved.