

Lab Insight **Dell CPU-Based AI PoC for Retail**

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

APRIL 2024

Introduction

As part of Dell's ongoing efforts to help make industry-leading AI workflows available to its clients, this paper outlines a sample AI solution for the retail market. The PoC leverages Dell™ technology to showcase an AI-powered inventory management application for retail organizations.

AI technology has been in development for some time, but recent technological advancements have greatly accelerated AI's ability to provide value across a wide range of enterprise applications. AI solutions have become a key initiative for many organizations. While the advancement of AI technology provides the basis for a diverse set of AI-powered applications, the specific requirements of different verticals provide distinct hardware and software challenges. IT organizations might be unsure of the technical requirements for deploying such a solution. This uncertainty may be due to unfamiliarity with AI, as well as an expectation that AI applications will require specialized hardware, often with limited availability.

This paper covers a solution specifically designed to capture the requirements of a retail-based AI deployment using a standard AMD™ CPU for AI training and inference. The solution leverages hardware from Dell, AMD, and Broadcom™, to create a solution powerful enough to capture and analyze large-scale video data from cameras in retail environments, as well as flexible enough to scale to the unique needs of individual retail environments. Training of the model was achieved in two days, utilizing the same Dell PowerEdge server that is used for inferencing. The scalability of the solution was tested with up to 20 video streams. The PoC additionally demonstrates AI optimizations for AMD CPUs by utilizing AMD's ZenDNN library. The utilization of the ZenDNN library, along with node pinning, resulted in an average throughput increase of 1.5x.

While the overall applications of AI in retail environments are much broader than the single inventory management solution outlined in this paper, the PoC demonstrates a framework for how IT organizations can quickly deploy an AI solution that delivers practical value in a retail environment by using readily available hardware.



Importance for the Retail Market

As with many other industries, the retail market has become increasingly data driven. Data can provide greater insight into areas such as customer behavior and product demand, as well as assist in optimizing operational areas such as procurement and inventory management. The emergence of AI technology provides even greater opportunity for valuable data-driven insights and optimizations within the retail industry.

Possibilities for retail-focused AI solutions include both customer experience (CX)-driven solutions and operations-focused applications. CX might be enhanced with personalized recommendation systems based on customer purchase trends, or virtual assistants capable of providing product recommendations for online retail experiences. Retail operations may be optimized through solutions such as AI-enhanced surveillance to detect fraud or theft, inventory management systems, or AI-powered product pricing systems.

These examples, as well as the more in-depth PoC study outlined in this paper, are a small subset of possible AI applications that may be implemented by retail organizations. While the exact solution implementations that are most appropriate may vary between organizations based on several factors such as location, size, type of goods sold, and distribution of online versus in-person sales, it is clear that AI applications can provide immense value in retail environments.

While a proactive approach to AI adoption may be beneficial to retail organizations, unfamiliarity with AI technology and the hardware and software components needed to deploy and optimize such solutions act as a barrier to adoption. The following solution demonstrates a PoC for an AI-powered retail inventory management system that can be quickly deployed and further expanded upon by retail organizations using commonly available hardware.





Solution Overview

The retail inventory management solution addresses a common challenge in retail environments of inventory distortion. Without accurate and timely inventory management, retail organizations can be challenged with stock levels that are either too low or too high. Both situations can prove to be costly. Too much inventory requires additional storage, commitment of capital, and potential waste of perishable items. Conversely, too low of inventory can lead to customer dissatisfaction and loss of sales. In many cases, low inventory leads to customers purchasing at competitive retailers and may lead to overall loss of brand loyalty. By utilizing computer vision and object detection AI models to monitor and track inventory, retailers can achieve real-time insights into their stock to balance their inventory more appropriately and provide valuable insights back to suppliers.

To demonstrate a real-world example solution of an AI application that could be deployed to address such retail challenges, Scalers AI™, in partnership with Dell, Broadcom, and The Futurum Group, implemented a PoC solution for a retail inventory management system. The solution was designed to capture data from store cameras and use an object-detection AI model to monitor and manage product stock levels. The solution was capable of detecting products on store shelves, keeping track of inventory, and raising alerts of low or out of stock items.

All of this was accomplished using standard Dell PowerEdge servers with 32 core 4th Gen AMD EPYC processors and Broadcom networking. No GPUs were required. The CPU-based solution was further optimized with AMD's Zen Deep Neural Network (ZenDNN) library, which provides optimizations for deep learning inferencing on AMD CPU hardware. AMD's ZenDNN optimizations delivered an average of 1.5x increased throughput performance to the PoC. By utilizing modest, CPU-based hardware, this PoC solution demonstrates a clear example of a readily deployable and broadly applicable AI retail solution.

Solution Highlights

- Retail inventory management solution monitors products on store shelves and tracks stock levels.
- Deployed using modest Dell servers with no GPUs.
- AMD ZenDNN optimizations increase throughput by 1.5x
- Tested with up to 64 processes on a 32-core CPU.
- Flexible, dual service architecture separates video inferencing pipeline and visualization process.
- Scalable architecture connected with high bandwidth Broadcom Ethernet.

To achieve the solution, store shelves were configured in zones with the product names and corresponding x,y coordinate pairs that indicated the shelf location. The products, location, and the maximum capacity for each item were stored as JSON objects.

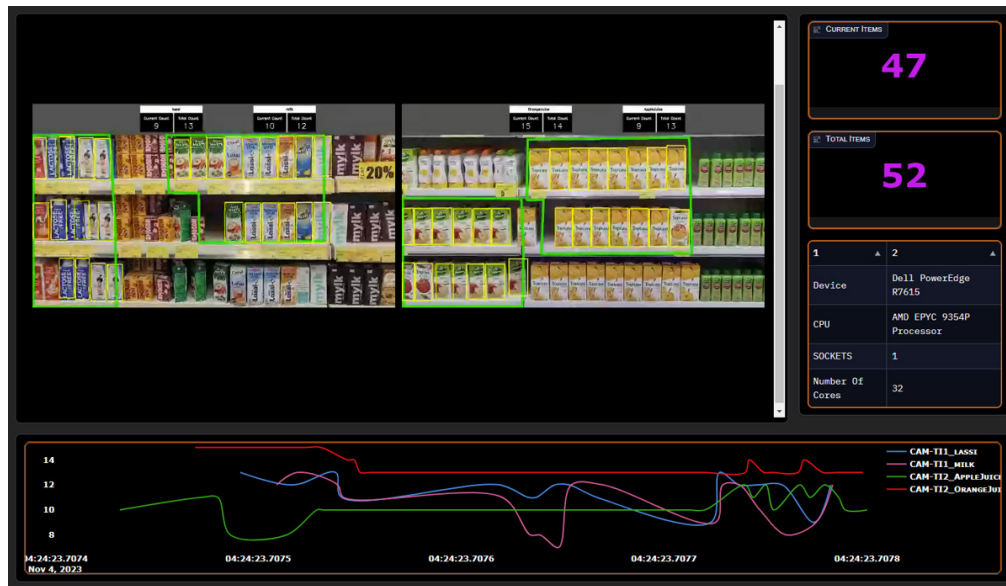


Figure 1: Visualization Dashboard

The identification and monitoring of products in each zone is achieved by capturing video data from store cameras into a video pipeline for processing. The live video stream is captured, decoded, and then inferred using an object-detection AI model. The video pipeline is run on a typical Dell PowerEdge server without requiring any GPUs or specialized accelerators. The video streams can additionally be directed to Dell PowerScale NAS storage for long term retention. Zenoh (Zero Overhead Network Protocol) is then utilized for distribution to an additional Dell server running a visualization process. The visualization engine enables the video stream to be shared over the web for remote viewing and analysis. The visualization dashboard can be seen in Figure 1. Figure 2 depicts a high-level diagram of the solution pipeline.

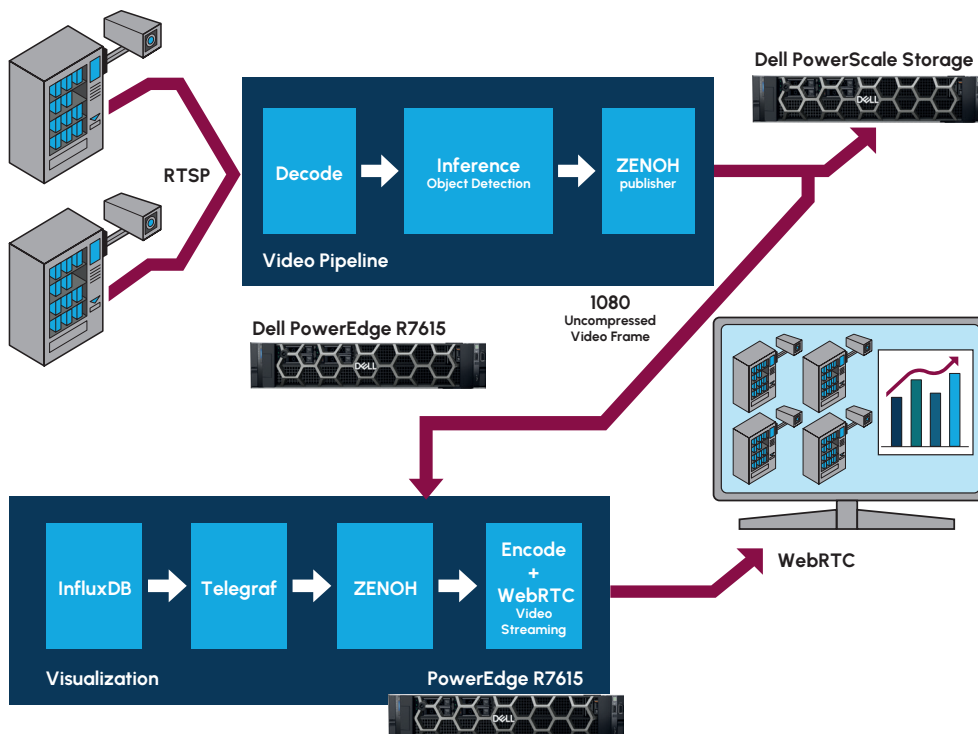


Figure 2: Retail Inventory Management AI Pipeline (Source: Scalera AI)

By separating the architecture into two distinct pieces, with one server powering video decoding and object detection, and a separate server for the visualization process, the PoC provides a framework for a highly scalable solution. Traditional approaches would combine the processes into a single pipeline, however, this architecture can prove challenging to scale due to the different computational needs of the services. Utilizing a dual service approach, provides greater flexibility to scale the processes as needed for retail organizations further expanding upon this PoC. Both the video pipeline and the visualization service can be scaled independently as requirements such as the number of video streams or application logic are adjusted. The dual service architecture and scalability of the overall solution is enabled by utilizing high speed Broadcom NetXtreme-E NICs which maintain high bandwidth between the video inferencing and visualization services.

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 include the following:

- Dell PowerEdge R7615 Servers
 - AMD EPYC 9354P 32-Core Processors
 - 768 GB Memory
 - 1 TB Storage
 - Broadcom BCM57508 NetXtreme-E 200G Ethernet Controller
- Dell PowerSwitch Z9664
- Dell PowerScale Scale-Out NAS Storage
 - Optional for long term retention



Highlights for AI Practitioners

It is notable for AI practitioners that the project was not limited to the deployment and inferencing of the AI model. The solution additionally involved customization of the pre-trained base model using a process known as Transfer Learning. The solution began with the SSD_MobileNet_v2 model for object detection, which was an ideal model for this PoC as it provides a one-stage object detection model that does not require exceptional compute power. The model was then customized via Transfer Learning with the SKU110K image data set. The training process involved 23,000 images and resulted in a mean average precision (mAP) of 0.7. The training process was completed in approximately two days.

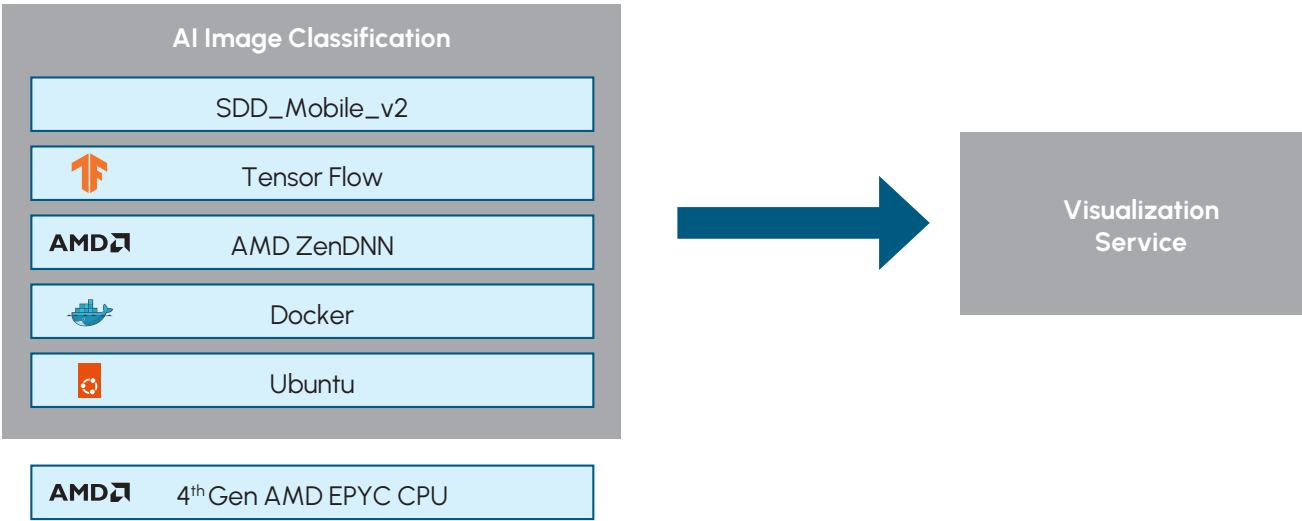


Figure 3: Object Detection Software Overview

It should also be noted that both the model training and deployment of the video pipeline solution were accomplished using the same 32 core Dell PowerEdge R7615 server. The PoC demonstrates the ability to achieve useful AI applications on CPU-based hardware that is commonly found in retail environments. The solution is further optimized for inferencing on AMD CPUs by utilizing AMD’s ZenDNN library and node pinning. The ZenDNN library provides performance tuning for deep learning inferencing on AMD CPUs while node pinning can further optimize the application by binding processes to dedicated compute resources.

The below table shows the ZenDNN parameter configurations used.

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	0	Default Value
ZENDNN_TENSOR_POOL_LIMIT	512	Set to 512 to optimize for Convolutional Neural Network
OMP_NUM_THREADS	32	Sets threads to 32 to match # of cores
GOMP_CPU_AFFINITY	0-31	Binds threads to physical CPUs. Set to number of cores in the system

Figure4: ZenDNN Configurations

Key Highlights for AI Practitioners

- AI powered retail solution deployed on standard Dell PowerEdge hardware with 32 core 4th Generation AMD EPYC processors. No GPUs were required.
- SSD_Mobile_Net_v2 model was used for object detection without high compute requirements. Achieved mAP of 0.7.
- Transfer learning process provides customization of model with relatively small training dataset. Training achieved in 2 days.
- Inferencing tested with up to 64 processes on a 32 core CPU. Optimized with ZenDNN for an average throughput increase of 1.5x.

Considerations for IT Operations

The hardware used in this AI application, including Dell PowerEdge R7615 servers with 4th Gen 32 core AMD EPYC 9354P Processors, Dell PowerScale NAS, Dell PowerSwitch Z9664, and Broadcom NetXtreme-E NICs, is familiar and available to IT operations, yet each component provides valuable characteristics needed to support this type of solution.

The Dell PowerEdge servers provide powerful 4th Generation AMD EPYC processors that are capable of supporting both the AI and application workloads, and the Dell PowerScale NAS provides a high-performance, highly scalable NAS storage system capable of handling large-scale video and image data. The solution is then tied together using Broadcom Ethernet capable of supporting the high bandwidth requirements of video streaming. Most notably, these components all provide scalability for IT organizations to further build out this application with more demanding requirements such as additional video streams or additional application logic.

Futurum Group Comment: *The specific use of Dell PowerEdge R7615 servers should be noted, as it demonstrates the ability to run AI workloads on standard hardware, commonly deployed in retail environments. While not considered a high-end compute server, the R7615 servers with mid-range 32 core 9354P Processors proved capable of all processes including model training, inferencing, and the separate visualization engine. This enables retail IT organizations to deploy such solutions without acquiring GPUs or requiring the datacenter level cooling needed for higher end servers. Additionally, by separating the architecture into separate video and visualization pipelines, the solution can be scaled to meet the size and performance requirements of a broad range of retail environments.*

The on-premises deployment of this solution additionally enables IT operations to achieve their data security and data privacy requirements. While public cloud has been utilized for many early iterations of AI applications, data privacy becomes a concern for many organizations as they build further AI applications leveraging private data. By deploying this, or similar, retail solutions on-premises, IT operations have greater control over the privacy of their data, which may include sensitive consumer or product information. The on-premises deployment of this solution also offers a potential economic advantage in its ability to avoid cloud storage costs when storing large capacities of video data. It additionally avoids the high networking requirements of uploading many video streams to the cloud.

Specifications of the Dell PowerEdge servers used in this PoC can be found in Figure 5

PowerEdge R7615		
Device Name		Dell PowerEdge R7615
CPU	Model Name	AMD EPYC 9354P 32-Core Processor
	Number Of Cores per Socket	32
	Number Of Sockets	1
Memory	Size	768 GB
Storage	Size	1 TB
Network		Broadcom NetXtreme-E BCM57508
OS	Name	Ubuntu 22.04.3 LTS
	Kernel	5.15.0-86-generic

Figure 5: Dell PowerEdge Server Details

Key Highlights for IT Operations

- AI solution deployed on readily available Dell hardware commonly found in retail environments.
- PoC built with scalable architecture to handle future development.
- On-premises deployment assists IT in meeting data privacy concerns and economic constraints.

Retail Solution Performance Observations

A key performance metric for the retail inventory management reference solution is the throughput of images per second as they are streamed by the in-store video cameras, decoded, and inferenced by the video pipeline. Video data is a common source for AI applications in the retail market, due to the prevalence of existing cameras deployed in stores, and the value of information that can be obtained by the video data. Because of this, the throughput performance insights gained from this PoC can translate to additional retail solutions that rely on image processing.

To examine the performance of the 32 core AMD EPYC 9354P processor for data capture and inferencing, the video pipeline was tested both with and without [ZenDNN](#) performance tuning, as well as with core pinning and node pinning. ZenDNN is a library that optimizes the performance of AMD processors for deep learning inferencing applications. The node pinning and core pinning are techniques offer optimization by binding processes to specific NUMA nodes or cores. The tests were run with up to 64 processes running on a 32 core server. The results of this testing can be seen in Figure 6.

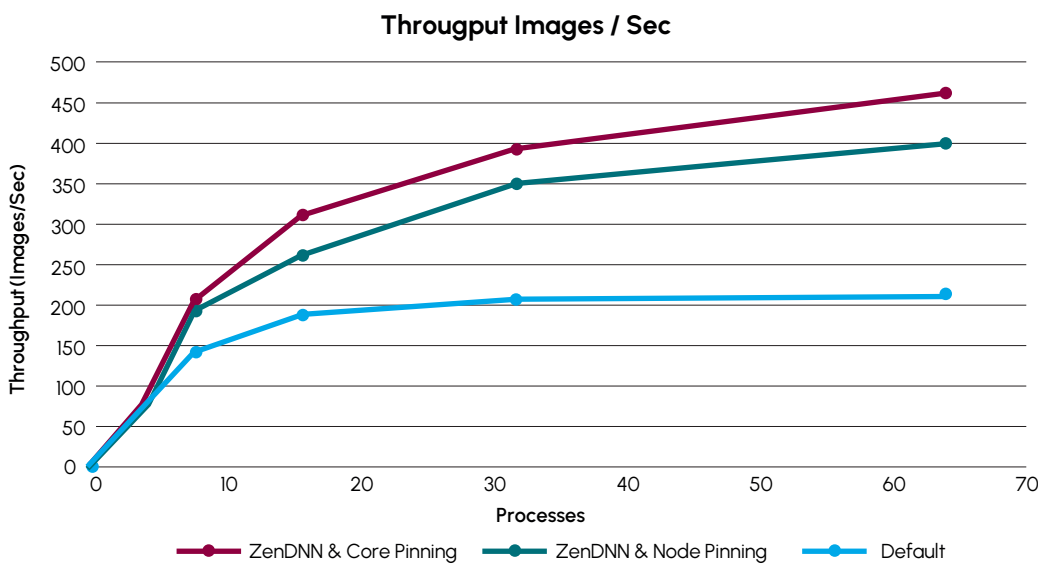


Figure 6: Throughput Performance

The performance results demonstrate that the use of ZenDNN with node pinning can provide a dramatic increase in throughput, with mostly lower CPU utilization. On average, ZenDNN with node pinning achieved a throughput increase of approximately 1.5x. Further throughput increases were additionally achieved by utilizing core pinning. Full results can be seen in Figure 7.

Processes	Throughput Images/sec - ZenDNN			Throughput Images/sec - ZenDNN OFF		Difference ZenD-NN(Node pinning) vs Default
	Core Pinning	Node pinning	CPU utilization	Default	CPU utilization	
1	29.86	31.72	7.808695652	25.06	10.75217391	1.27
8	195.7	188.26	46.27717391	125.02	59.36684783	1.51
16	305.06	264.24	62.7548913	176.99	75.2388587	1.49
32	389.1	347.58	78.978125	204.98	83.00978261	1.7
64	460.88	392.32	93.09952446	214.43	91.55903533	1.83

Figure 7: Video Pipeline Throughput Test

The performance gains achieved with ZenDNN, core pinning, and node pinning demonstrate the ability to optimize CPUs for AI applications. Commonly, computationally demanding AI processes, such as the computer vision and object detection utilized in this PoC, are expected to require GPUs. Hardware alone, however, is not the only component that affects performance. Software such as ZenDNN plays a key role in optimizing the performance of the chosen hardware, as does configuration details such as utilizing core pinning or node pinning. By utilizing these configurations, organizations can achieve AI applications that meet their performance needs with a CPU-based solution utilizing readily available hardware.

The PoC solution was additionally tested with an increasing number of video streams to assess the bandwidth of the networked video pipeline and visualization service. 1080p video was streamed to the video pipeline where it was decoded and inferred. It was then transmitted and received by the visualization pipeline to be encoded and shared. The number of video streams was increased incrementally between 1 and 20 which resulted in an increasing bandwidth utilization. The bandwidth scaled from an average utilization of 1.65 Gbits/s and a max utilization of 3.4 Gbits/s with 1 stream, to an average utilization of 13.9 Gbits/s and a max utilization of 27.4 Gbits/s with 20 streams. An overview of the results can be seen in Figure 8.

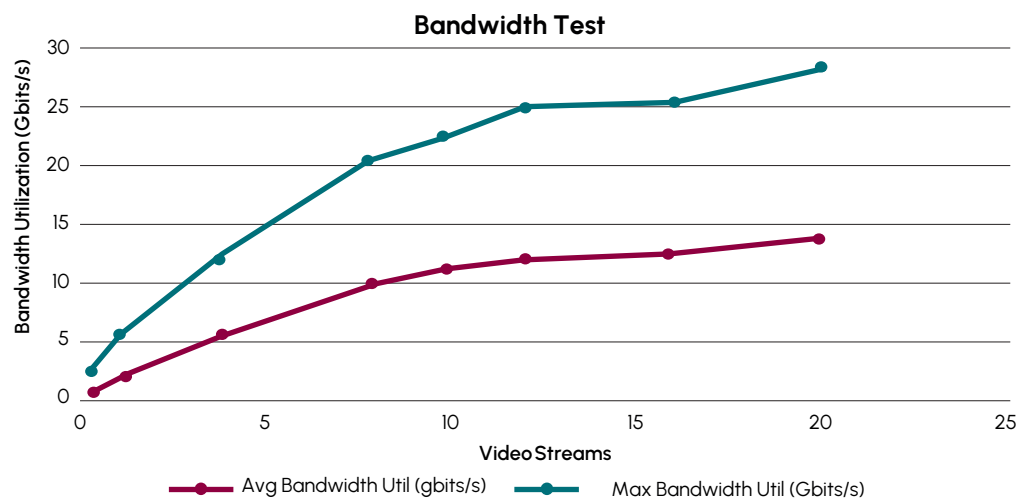



Figure 8: Inventory Management System Bandwidth

Notably, the bandwidth does not increase linearly in relation to the number of streams, allowing the solution to scale as additional streams are needed. As the number of streams increases, however, the solution does experience a decrease in frames-per-second. While frames-per-second decreases, the overall utility of the solution is not significantly impacted. Higher frame rates are of greater importance when considering video with large amounts of motion, or when viewing quality is a major priority. In this particular solution, lower frame rates are acceptable as the focus is stationary store shelves, and real time viewing is not the primary use case. Full results of testing the networked solution, including both bandwidth utilization and frames per second, can be seen in Figure 9.

Number of Streams	AVG FPS / Stream	Throughput (FPS)	Avg Bandwidth Util (Gbits/s)	Max Bandwidth Util (Gbits/s)	Avg CPU Util (%)	Avg Memory Util (GB)
1	31.14	31.14	1.65	3.4	12.61	6.5
2	30.92	61.84	3.2	6.7	21.8	7.27
4	28.78	115.12	6.2	12.2	41.38	9.2
8	22.17	177.36	9.86	20.5	65.06	13.9
10	20.53	205.3	11.2	22.4	73.18	16.4
12	18.8	225.6	12.1	24.7	78.76	18.2
16	13.97	223.52	12.6	25.6	81.39	22.2
20	11.7	234	13.9	27.4	84.1	26.7

Figure 9: Inventory Management System Bandwidth Test



The results of this performance testing demonstrate that the bandwidth of the networked servers is capable of scaling alongside more demanding video requirements. The separation of the video pipeline and the visualization service onto distinct servers allows the architecture to independently scale the compute resources for the two services. To capitalize on this architecture however, the networking between the servers must be capable of providing adequate bandwidth between the services. To do so, the PoC solution utilizes Broadcom BCM57508 NetXtreme-E Ethernet controllers capable of supporting up to 200GbE. By utilizing a modular architecture that's connected with scalable, high bandwidth networking, the retail inventory management PoC provides a flexible starting point for retail organizations to scale to their individual needs, including the number of video streams, FPS requirements, and additional application logic.

Final Thoughts

With the rapid development of AI technology, the retail market presents many opportunities to deploy valuable new AI-powered applications. With the broad range of value that AI can bring to retail environments, both in improving CX and optimizing store operations, retail organizations should look to be proactive in adopting the emerging technology.

As a new technology, there are many unknowns and misconceptions for those in IT who may be unfamiliar with AI deployments, complicating and delaying new AI applications. A common challenge faced by IT is the expectation that AI applications will require specialized hardware solutions that are inaccessible. The AI-powered retail inventory management solution outlined in this paper serves as a demonstration of a broadly applicable AI solution for retail that can be deployed on off-the-shelf hardware solutions. The Dell hardware solutions used in the PoC deployment were demonstrated to handle the high-bandwidth video requirements as well as the AI modeling and inferencing requirements without the use of purpose-built accelerators, GPUs, or custom hardware.

The PoC solution outlined in this paper additionally serves as a reference for retail organizations to quickly deploy their own inventory management solution. While the solution discussed in this paper is limited to a PoC, it was designed with scalability in mind for organizations to further develop and scale a solution for their needs.

The use of an AI-powered inventory management system can provide real value and cost savings to organizations by avoiding over- or under-stocking products. By using readily available hardware and reference solutions, the barrier of entry for deploying such an AI solution is dramatically lowered, allowing retail organizations to achieve quicker deployments of new AI applications and quicker time to value.

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 |