

Lab Insight

Dell AI PoC for Transportation & Logistics

AUTHOR

Mitch Lewis

Research Analyst | The Futurum Group

MARCH 2024



Introduction

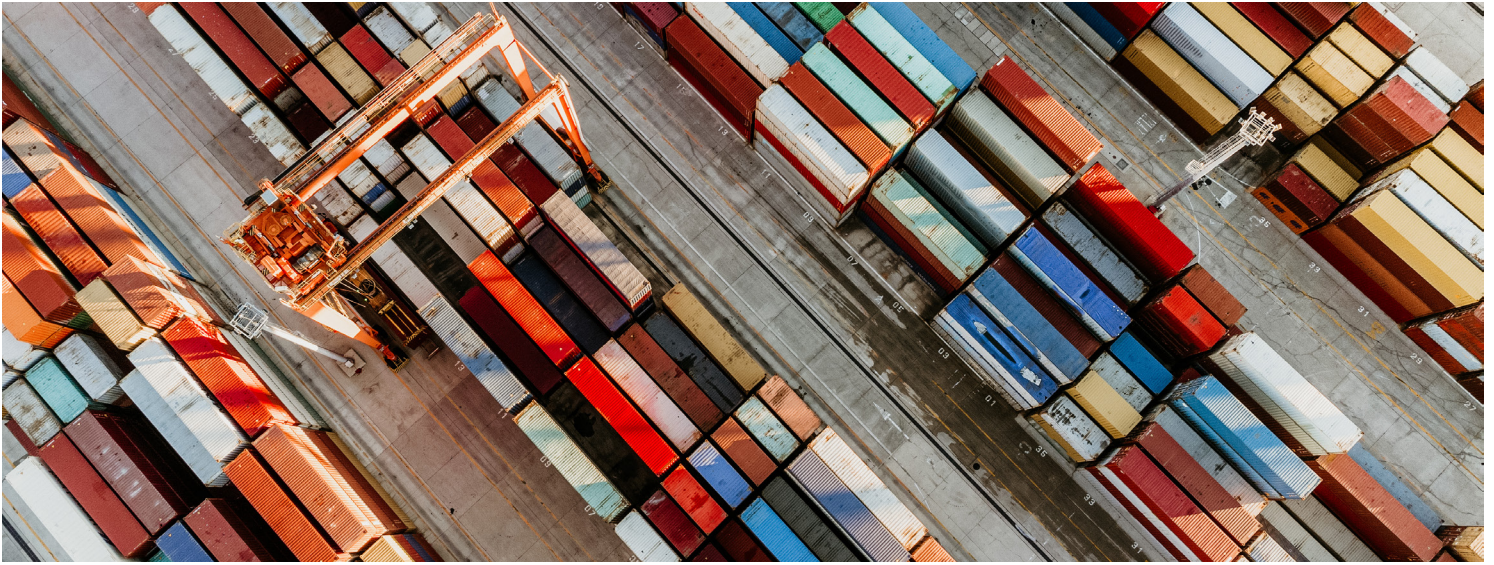
As part of Dell's ongoing efforts to help make industry-leading AI workflows available to its clients, this paper outlines a sample AI solution for the transportation and logistics market. The reference solution outlined in this paper specifically targets challenges in the maritime industry by creating an AI powered cargo monitoring PoC built with Dell™ hardware.

AI as a technology is currently in a rapid state of advancement. While the area of AI has been around for decades, recent breakthroughs in generative AI and large language models (LLMs) have led to significant interest across almost all industry verticals, including transportation and logistics. Futurum intelligence projects a 24% growth of AI in the transportation industry in 2024 and a 30% growth for logistics.

The advancements in AI now open significant possibilities for new value-adding applications and optimizations, however different industries will require different hardware and software capabilities to overcome industry specific challenges. When considering AI applications for transportation and logistics, a key challenge is operating at the edge. AI-powered applications for transportation will typically be heavily driven by on-board sensor data with locally deployed hardware. This presents a specific challenge, requiring hardware that is compact enough for edge deployments, powerful enough to run AI workloads, and robust enough to endure varying edge conditions.

This paper outlines a PoC for an AI-based transportation and logistics solution that is specifically targeted at maritime use cases. Maritime environments represent some of the most rigorous edge environments, while also presenting an industry with significant opportunity for AI-powered innovation. The PoC outlined in this paper addresses the unique challenges of maritime focused AI solutions with hardware from Dell and Broadcom™.

The PoC detailed in this paper serves as a reference solution that can be leveraged for additional maritime, transportation, or logistics applications. The overall applicability of AI in these markets is much broader than the single maritime cargo monitoring solution, however, the PoC demonstrates the ability to quickly deploy valuable edge-based solutions for transportation and logistics using readily available edge hardware.



Importance for the Transportation and Logistics Market

Transportation and logistics cover a broad industry with opportunity for AI technology to create a significant impact. While the overarching segment is widespread, including public transportation, cargo shipping, and end-to-end supply chain management, key to any transportation or logistics process is optimization. These processes are dependent on a high number of specific details and variables such as specific routes, number and types of goods transported, compliance regulations, and weather conditions. By optimizing for the many variables that may arise in a logistical process, organizations can be more efficient, save money, and avoid risk.

In order to create these optimal processes, however, the data surrounding the many variables involved needs to be captured. Further, this data needs to be analyzed, understood, and acted on. The large quantity of data required and the speed at which it must be processed in order to make impactful decisions to complex logistical challenges often surpasses what a human can achieve manually.

By leveraging AI technology, impactful decisions to transportation and logistics processes can be achieved quicker and with greater accuracy. Cameras and other sensors can capture relevant data that is then processed and understood by an AI model. AI can quickly process vast amounts of data and lead to optimized logistics conclusions that would otherwise be too timely, costly, or complex for organizations to make.

The potential applications for AI in transportation are vast and can be applied to various means of transportation including shipping, rail, air, and automotive, as well as associated logistical processes such as warehouses and shipping yards. One possible example is AI optimized route planning which could pertain to either transportation of cargo or public transportation and could optimize for several factors including cost, weather conditions, traffic, or environmental impact. Additional applications could include automated fleet management, AI powered predictive vehicle maintenance, and optimized pricing. As AI technology improves, many transportation services may be additionally optimized with the use of autonomous vehicles.

By adopting such AI-powered applications, organizations can implement optimizations that may not otherwise be achievable. While new AI applications show promise of significant value, many organizations may find adopting the technology a challenge due to unfamiliarity with the new and rapidly advancing technology. Deploying complex applications such as AI in transportation environments can pose an additional challenge due to the requirements of operating in edge environments. The following PoC solution outlines an example of a transportation focused AI application that can offer significant value to maritime shipping by providing AI-powered cargo monitoring using Dell hardware at the edge.

Solution Overview

To demonstrate an AI-powered application focused on transportation and logistics, Scalers AI™, in partnership with Dell, Broadcom, and The Futurum Group implemented a proof-of-concept for a maritime cargo monitoring solution. The solution was designed to capture sensor data from cargo ships as well as image data from on-board cameras. Cargo containers can be monitored for temperature and humidity to ensure optimal conditions are maintained for the shipped cargo. In addition, cameras can be used to monitor workers in the cargo area to ensure worker safety and prevent injury. The captured data is then utilized by an LLM to create an AI generated compliance report at the end of the ship's voyage.

This proof-of-concept addresses several problems that can be encountered in maritime shipping. Refrigerated cargo, known as reefer, is utilized to ship perishable items and pharmaceuticals that must be kept at specific temperatures. Without proper monitoring to ensure optimal temperatures, reefer may experience swings in temperature, resulting in spoiled products and ultimately financial loss. Predictive forecasting of the power requirements for refrigerated cargo can provide additional cost and environmental savings by providing greater power usage insights.

Similarly, dry cargo can become spoiled or damaged when exposed to excessive moisture. Moisture can be introduced in the form of condensation – known as cargo sweat – due to changes in climate and humidity during the ships journey. By monitoring the temperature and humidity of the cargo, alerts can be raised signaling the possibility of cargo sweat and allowing ventilation adjustments to be made which can prevent moisture related damage.

A third issue addressed by the maritime cargo monitoring PoC is that of worker safety. The possibility of shifting cargo containers can lead to dangerous situations and potential injuries for those working in container storage areas. By using video surveillance of workers in cargo areas, these potential injuries can be avoided.

The PoC provides monitoring of these challenges with an additional visualization dashboard that displays information such as number of cargo containers, forecasted energy consumption, container temperature and humidity, and a video feed of workers. The dashboard additionally raises alerts as issues arise in any of these areas. This information is further compiled in to an end of voyage report for compliance and logging purposes, automatically generated with an LLM.

To achieve the PoC solution, simulated sensor data is generated for both reefers and dry containers, approximating the conditions undergone during a real voyage. The sensor data is written to an OPCUA server which then supplies data to a container sweat analytics module and a power consumption predictor. For dry containers, the temperature and humidity data is utilized alongside the forecasted weather of the route to create dew point calculations and monitor potential container sweat. Sensor data recording the temperature of reefer containers is monitored to ensure accurate temperatures are maintained, and a decision tree regressor model is leveraged to predict future power consumption for the next hour.

Solution Highlights

- AI powered cargo monitoring solution for maritime shipping.
- Monitors temperature and conditions of cargo containers. Predicts energy usage.
- Monitors maritime workers in cargo zones to avoid injury.
- Utilizes LLM to automatically generate compliance reports.

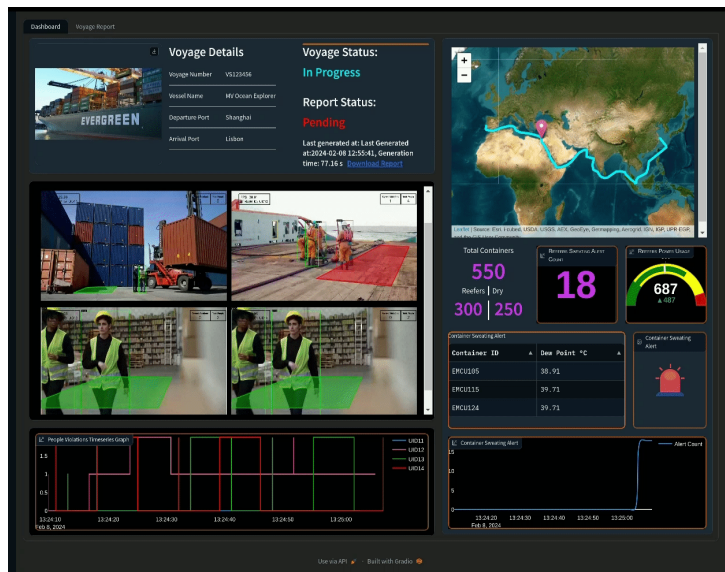


Figure 1: Visualization Dashboard

For monitoring worker safety, RTSP video data is captured into a video analytics pipeline built on NVIDIA™ DeepStream. Streaming data is decoded and then inferred using the YoloV8s model to detect workers entering dangerous, restricted zones. The restricted zones are configured as x,y coordinate pairs stored as JSON objects. Uncompressed video is then published to the visualization service using the Zero Overhead Network Protocol (Zenoh).

Monitoring and alerts for all of these challenges is displayed on a visualization dashboard as can be seen in Figure 1, as well as summarized in an end of voyage compliance report. The resulting compliance report that details the information collected on the voyage is AI generated using the Zephyr 7B model. Testing of the PoC found that the report could be generated in approximately 46 seconds, dramatically accelerating the reporting process compared to a manual approach.

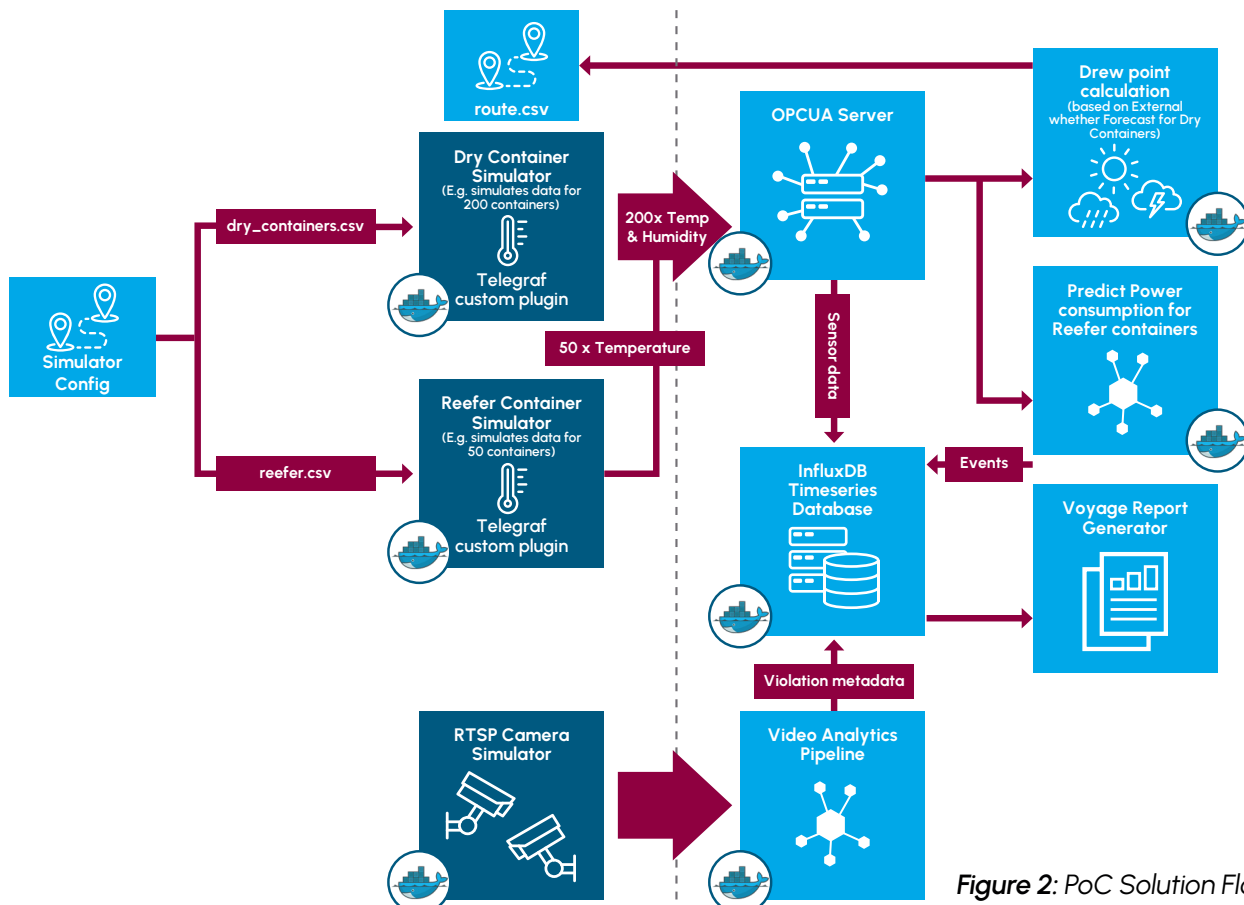


Figure 2: PoC Solution Flow

To achieve the PoC solution in-line with the restraints of a typical maritime use case, the solution was deployed using Dell PowerEdge servers designed for the edge. The sensor data calculations and predictions, video pipeline, and AI report generation were achieved on a Dell PowerEdge XR7620 server with dual NVIDIA A100 GPUs. A Dell PowerEdge XR12 server was deployed to host the visualization dashboard. The two servers were connected with high bandwidth Broadcom NICs.

An overview of the solution can be seen in Figure 2

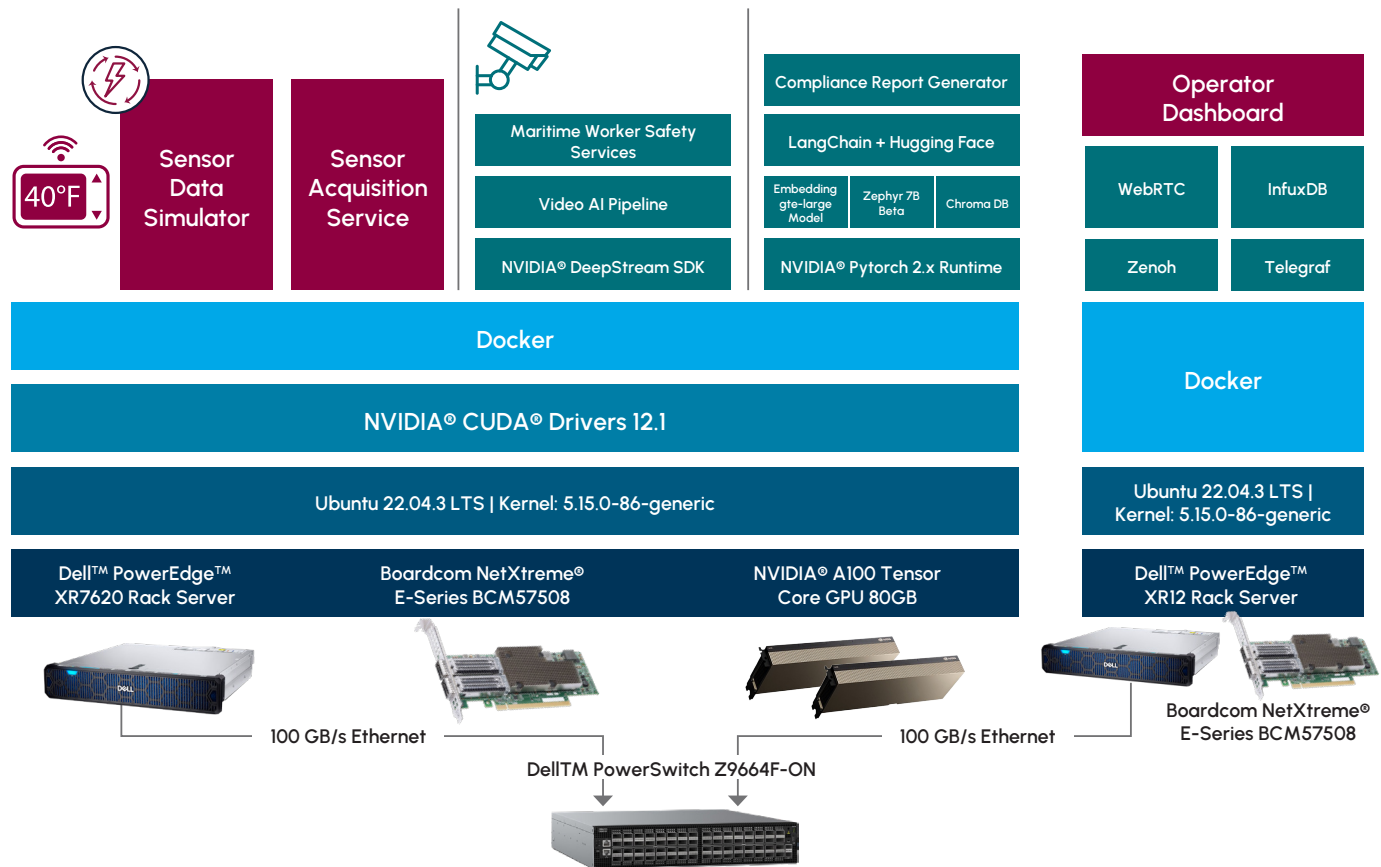


Figure 3: PoC Architecture Overview

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



Additional details about the implementation and performance testing of the PoC on GitHub, including:

- Configuration information including diagrams and YAML code
- Instructions for doing the performance tests
- Details of performance results
- Source code
- Samples for test process

<https://github.com/dell-examples/generative-ai/tree/main/transportation-maritime>

Highlights for AI Practitioners

The cargo monitoring PoC demonstrates a solution that can avoid product loss, enhance compliance and logging, and improve worker safety, all by using AI. The creation of these AI processes was done using readily available AI tools. The process of creating valuable, real-world solutions by utilizing such tools should be noted by AI practitioners.

The end of voyage compliance report is generated using the Zephyr 7B LLM model created by Hugging Face's H4 team. The Zephyr 7B model, which is a modified version of Mistral 7B, was chosen as it is a publicly available model that is both lightweight and highly accurate. The Zephyr 7B model was created using a process called Distilled Supervised Fine Tuning (DSFT) which allows the model to provide similar performance to much larger models, while utilizing far fewer parameters. Zephyr 7B, which is a 7 Billion parameter model, has demonstrated performance comparable to 70 Billion parameter models. This ability to provide the capabilities of larger models in a smaller, distilled model makes Zephyr 7B an ideal choice for edge-based deployments with limited resources, such as in maritime or other transportation environments.

While Zephyr 7B is a very powerful and accurate LLM model, it was trained on a broad data set and it is intended for general purpose usage, rather than specific tasks such as generating a maritime voyage compliance report. In order to generate a report that is accurate to the maritime industry and the specific voyage, more context must be supplied to the model. This was achieved using a process called Retrieval Augmented Generation (RAG). By utilizing RAG, the Zephyr 7B model is able to incorporate the voyage specific information to generate an accurate report which detailed the recorded container and worker safety alerts. This is notable for AI practitioners as it demonstrates the ability to use a broad, pre-trained LLM model, which is freely available, to achieve an industry specific task.

To provide the voyage specific context to the LLM generated report, time series data of recorded events, such as container sweating, power measurements, and worker safety violations, is queried from InfluxDB at the end of the voyage. This text data is then embedded using the Hugging Face LangChain API with the gte-large embedding model and stored in a ChromaDB vector database. These vector embeddings are then used in the RAG process to provide the Zephyr 7B model with voyage specific context when generating the report.

AI practitioners should also note that AI image detection is utilized to detect workers entering into restricted zones. This image detection capability was built using the YOLOv8s object detection model and NVIDIA DeepStream. YOLOv8s is a state of the art, open source, AI model for object detection built by Ultralytics. The model is used to detect workers within a video frame and detect if they enter into pre-configured restricted zones. NVIDIA DeepStream is a software development toolkit provided by NVIDIA to build and accelerate AI solutions from streaming video data, which is optimized for NVIDIA hardware such as the A100 GPUs used in this PoC. It is notable that NVIDIA DeepStream can be utilized for free to build powerful video-based AI applications, such as the worker detection component of the maritime cargo monitoring solution. In this case, the YOLOv8s model and the DeepStream toolkit are utilized to build a solution that has the potential to prevent serious workplace injuries.

Key Highlights for AI Practitioners

- Maritime compliance report generated with Zephyr 7B LLM model
- Retrieval Augmented Generation (RAG) approach used to provide Zephyr 7B with voyage specific information
- YOLOv8s and NVIDIA DeepStream used to create powerful AI worker detection solution using video streaming data

Considerations for IT Operations

The maritime cargo monitoring PoC is notable for IT operations as it demonstrates the ability to deploy a powerful AI driven solution at the edge. For many in IT, AI deployments in any setting may be a challenge, due to overall unfamiliarity with AI and its hardware requirements. Deployments at the edge introduce even further complexity.

Hardware deployed at the edge requires additional considerations, including limited space and exposure to harsh conditions, such as extreme temperature changes. For AI applications deployed at the edge, these requirements must be maintained, while simultaneously providing a system powerful enough to handle such a computationally intensive workload.

For the maritime cargo monitoring PoC, Dell PowerEdge XR7620 and PowerEdge XR12 servers were chosen for their ability to meet both the most demanding edge requirements, as well as the most demanding computational requirements. Both servers are ruggedized and are capable of operating in temperatures ranging from -5°C to 55°C, as well as withstanding dusty or otherwise hazardous environments. They additionally offer a compact design that is capable of fitting into tight environments. This provides servers that are ideal for a demanding edge environment, such as in maritime shipping, which may experience large temperature swings and may have limited space for servers. Meanwhile, the Dell PowerEdge XR7620 is also equipped with NVIDIA GPUs, providing it with the compute power necessary to handle AI workloads.



Dell PowerEdge XR7620

NVIDIA A100 GPUs were chosen as they are well suited for various types of AI workflows. The PoC includes both a video classification component and a large language model component, requiring hardware that is well suited for both workloads. While there are other processors that are more specialized specifically for either video processing or language models, the A100 GPU provides flexibility to perform both well on a single platform.

The use of high bandwidth Broadcom NICs is also a notable component of the PoC solution for IT operations to be aware of. The Broadcom NICs are responsible for providing a high bandwidth Ethernet connection between the cargo and worker monitoring applications and the visualization and alerting dashboard. The use of scalable, high bandwidth NICs is crucial to such a solution that requires transmitting large amounts of sensor and video data, which may include time sensitive information. Detection of issues with either reefer or dry containers may require quick action to protect the cargo, and quick detection of workers in hazardous environments can prevent serious harm or injury. The use of a high bandwidth Ethernet connection ensures that data can be quickly transmitted and received by the visualization dashboard for operators to respond to alerts as they arise.

Key Highlights for IT Operations

- AI solution deployed on rugged Dell PowerEdge XR7620 and PowerEdge XR12 servers to accommodate edge environment and maintain high computational requirements.
- NVIDIA A100 GPUs provide flexibility to support both video and LLM workloads.
- Broadcom NICs provide high bandwidth connection between monitoring applications and visualization dashboard.



Solution Performance Observations

Key to the performance of the maritime cargo monitoring PoC is its ability to scale to support multiple concurrent video streams for monitoring worker safety. The solution must be able to quickly decode and inference incoming video data to detect workers in restricted areas. The ability for the visualization dashboard to quickly receive this data is additionally critical for actions to be taken on alerts as they are raised. The solution was separated into a distinct inference server, to capture and inference data, and an encode server, to display the visualization service. This architecture allows the solution to scale the services independently as needed for varying requirements of video streams and application logic. The separate services are then connected with high bandwidth Ethernet using Broadcom NetXtreme®-E Series Ethernet controllers. The following performance data demonstrates the ability to scale the solution with an increasing number of data streams. Each test was run for a total of 10 minutes and video streams were scaled evenly across the two NVIDIA A100 GPUs. Additional performance results are available in the appendix.

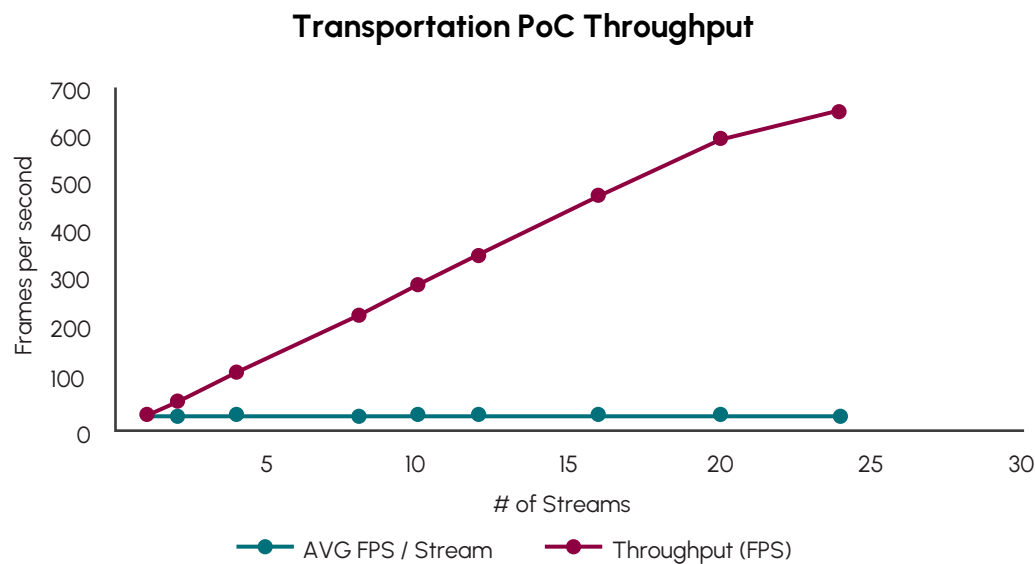


Figure 4: Transportation PoC Throughput

Figure 4 displays the total throughput of frames per second as well as the average throughput as the number of streams increased. The frames per second metric includes video decoding, inference, post-processing, and publishing of an uncompressed stream. The PoC displayed increasing throughput with a maximum of 653.7 frames per second when tested with 24 concurrent streams. Notably, the average frames per second remained steady at approximately 30 frames per second for up to 20 streams, which is considered an industry standard for video processing workloads. When tested with 24 streams, the solution did experience a slight drop, with an average of 27.24 frames per second. Overall, the throughput performance demonstrates the ability of the Dell PowerEdge Server and the NVIDIA A100 GPUs to successfully handle a demanding video-based AI workload with a significant number of concurrent streams.

Transportation PoC Throughput

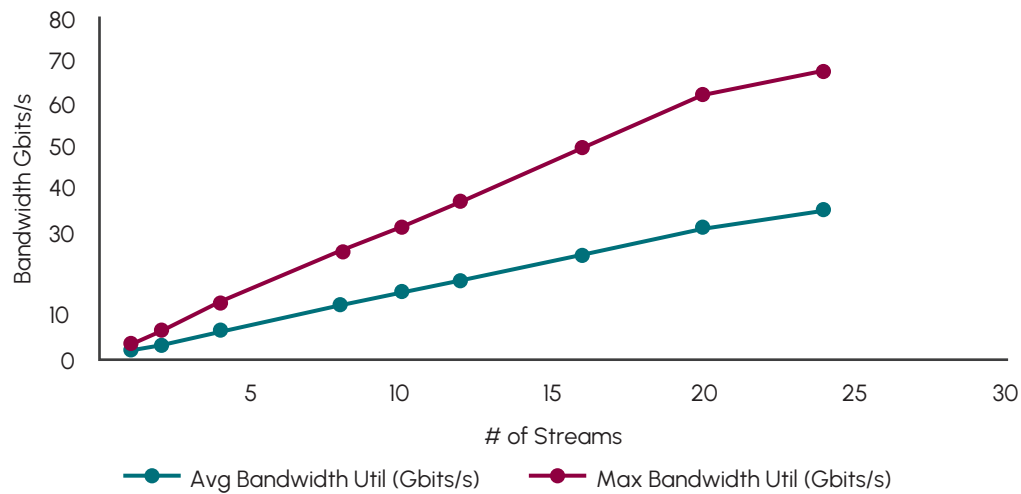
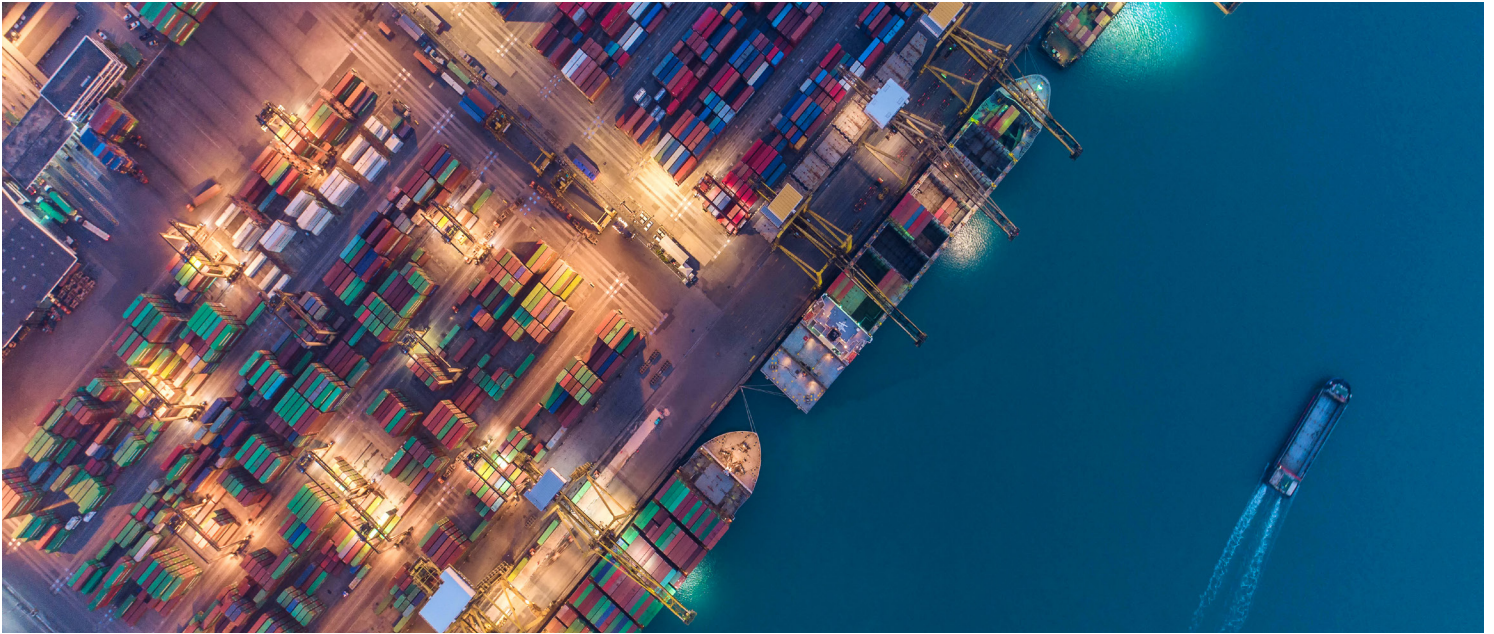


Figure 5: Transportation PoC Bandwidth Utilization

Figure 5 displays the solution's bandwidth utilization as the number of streams increased from 1 to 24. The results demonstrate the increase in required bandwidth, both at a maximum and on average, as the number of streams increased. The average bandwidth utilization scaled from 1.56 Gb/s with a single video stream, to 34.6 Gb/s when supporting 24 concurrent streams. The maximum bandwidth utilization was observed to be 3.13 Gb/s with a single stream, up to 67.9 Gb/s with 24 streams. By utilizing scalable, high bandwidth 100Gb/s Broadcom Ethernet, the solution is able to achieve the increasing bandwidth utilization required when adding additional video streams.

The performance results showcase the PoC as a flexible solution that can be scaled to accommodate varying levels of video requirements while maintaining performance and scaling bandwidth as needed. The solution also provides the foundation for additional AI-powered transportation and logistics applications that may require similar transmission of sensor and video data.





Final Thoughts

The maritime cargo monitoring PoC provides a concrete example of how AI can improve transportation and logistics processes by monitoring container conditions, detecting dangerous working environments, and generating automated compliance reports. While the PoC presented in this paper is limited in scope and executed using simulated sensor datasets, the solution serves as a starting point for expanding such a solution and a reference for developing related AI applications.

The solution additionally demonstrates several notable results. The solution utilizes readily available AI tools including Zephyr 7B, YOLOv8s, and NVIDIA DeepStream to create valuable AI applications that can be deployed to provide tangible value in industry specific environments. The use of RAG in the Zephyr 7B implementation is especially notable, as it provides customization to a general-purpose language model, enabling it to function for a maritime specific use case. The PoC also showcased the ability to deploy an AI solution in demanding edge environments with the use of Dell PowerEdge XR7620 and XR12 servers and to provide high bandwidth when transmitting critical data by using Broadcom NICs.

When tested, the PoC solution demonstrated the ability to scale up to 24 concurrent streams while experience little loss of throughput and successfully supporting increased bandwidth requirements. Testing of the LLM report generation showed that an AI augmented maritime compliance report could be generated in as little as 46 seconds. The testing of the PoC demonstrate both its real-world applicability in solving maritime challenges, as well as its flexibility to scale to individual deployment requirements.

Transportation and logistics are areas that rely heavily upon optimization. With the advancements in AI technology, these markets are well positioned to benefit from AI-driven innovation. AI is capable of processing data and deriving solutions to optimize transportation and logistics processes at a scale and speed that humans are not capable of achieving manually. The opportunity for AI to create innovative solutions in this market is broad and extends well beyond the maritime PoC detailed in this paper. By understanding the approach to creating an AI application and the hardware components used, however, organizations in the transportation and logistics market can apply similar solutions to innovate and optimize their business.

Appendix

Figure 6 shows full performance testing results for the cargo monitoring PoC.

Number of Streams	AVF FPS / Stream	Through put (FPS)	Avg Bandwidth Util (Gbits/s)	Max Bandwidth Util (Gbits/s)	Avg GPU Util (%)	Avg GPU Memory Util (%)
1	29.97	29.97	1.56	3.13	4.5	0.48
2	29.99	59.98	3.07	6.25	9.43	7.5
4	29.64	118.56	6.25	12.25	20.8	1.77
8	29.35	234.8	12.5	25	43.4	3.45
10	29.67	296.7	15.3	31.2	55.9	4.33
12	29.72	356.6	18.4	37.4	66.5	5.3
16	29.82	477.1	24.6	50	88.7	6.91
20	29.78	595.6	30.6	62.4	99.2	8.67
24	27.24	653.7	34.6	67.9	98.5	10.44

Figure 6: Full Performance Results

Server	CPU	RAM	Disk	GPU
DellTM PowerEdge™ XR7620	Intel® Xeon® Silver 4416+	128 GB	500 GB	2xNVIDIA A100
DellTM PowerEdge™ XR12	Intel® Xeon® Gold 5318Y	256 GB	10 TB	

Figure 7: Dell PowerEdge Server Configurations

Software	Version
Docker	24.0.7
NVIDIA CUDA	v12.1
Deepstream	6.4
Zenoh	0.10.1rc0
opcua	1.0.6
influxdb	1.7.10
telegraf	1.22.3
transformers	4.37.2
langchain	0.1.5

Figure 8: Software Components

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 |