MYTH-BUSTING | Are CPUs Powerful Enough to Run AI?

DELL™ POWEREDGE™ R760 SERVERS WITH 4TH GEN INTEL® XEON® SCALABLE PROCESSORS OFFERS UNPRECEDENTED PERFORMANCE FOR APPLIED AI REAL-WORLD APPLICATIONS

SCALERS AI

Can Intel® Xeon® Processors Effectively Run AI Applications?

The myth that specialized hardware is the only type of processor capable of handling AI workloads has been perpetuated by the fact that many popular benchmarks only focus on inference or training performance, while real-world applications require versatile compute resources.

> Dell Technologies™ commissioned Scalers AI™ to evaluate the performance of Dell™ PowerEdge™ R760 with 4th Gen Intel® Xeon® Scalable processors perform in our applied AI application.

Dell™ PowerEdge™ R760 servers powered by 4th Gen Intel® Xeon® Scalable processors are capable of handling a wide range of tasks including AI, and are generally more energy efficient and reliable than specialized hardware. In this whitepaper and accompanying solution code published in a Github repository, we demonstrate that 4th Gen Intel® Xeon® Scalable processors offer impressive performance gains for realistic applications with mixed and versatile workloads.

1.88x

IMPROVEMENT on Scalers AI™ Traffic Safety Solution

| Gen on Gen Performance Improvement Using Intel® Deep Learning Boost

"Scalers AI™ saw 1.88x performance improvement on Dell™ PowerEdge™ R760 platform with Intel® DL Boost on our traffic safety solution"

- Steen Graham, CEO at Scalers AI™

About Scalers AI™

Scalers AI™ specializes in creating end-to-end artificial intelligence (AI) solutions for a wide range of industries, including retail, smart cities, manufacturing, and healthcare. The company is dedicated to helping organizations leverage the power of AI for their digital transformation. Scalers AI™ has a team of experienced AI developers and data scientists who are skilled in creating custom AI solutions for a variety of use cases, including predictive analytics, chatbots, image and speech recognition, and natural language processing.

As a full stack AI solutions company with solutions ranging from the cloud to the edge, our customers often need versatile common off the shelf (COTS) hardware that works well across a range of workloads. Additionally, we also need advanced visualization libraries including the ability to render video in modern web application architectures.

Dell™ PowerEdge™ R760 Powered by 4th Gen Intel® Xeon® Scalable Processors offer Versatile Performance & Scale

Dell™ PowerEdge™ R760 powered by 4th Gen Intel® Xeon® Scalable Processors with Intel® DL Boost offers versatile performance and scale for modern applications. This powerful server features advanced technologies and architectures, such as Intel® Turbo Boost Technology and Intel® Hyper-Threading, which allow it to deliver fast and efficient processing speeds. 4th Gen Intel® Xeon® Scalable processors also provide enhanced security and reliability, making Dell™ PowerEdge™ R760 a secure and robust platform for running mission-critical applications. Additionally, Dell™ PowerEdge™ R760 is designed to be easily scalable, allowing users to add more resources and expand their capabilities as needed. This makes it a versatile and flexible solution for organizations that require high performance and scalability in their applications.

About Intel® DL Boost

Intel® DL Boost with AMX (Advanced Matrix Extensions) is a technology that improves the performance of deep learning tasks on Intel® processors. It uses specialized instructions called "matrix operations" that are optimized for deep learning operations such as convolution and fully connected layers. This allows these operations to be performed more efficiently, resulting in faster and more accurate deep learning models. It can be used with software frameworks like TensorFlow, PyTorch, and Caffe2 and is supported on several types of Intel® processors. It is a useful tool for organizations looking to speed up their deep learning processes and advance their work in artificial intelligence and machine learning.

The Importance of Well-Rounded Performance and Scalability in Modern Applications

Modern applications need fast processing speeds, efficient resource use, and the ability to handle large amounts of data and traffic in order to provide a smooth and efficient user experience. They must also be able to adapt and scale to meet changing needs and operate in a flexible, resilient, and highly responsive manner. Advanced technologies and architectures such as cloud computing and micro-services can help to achieve this. Well-rounded performance and scalability are therefore crucial factors in the real world.

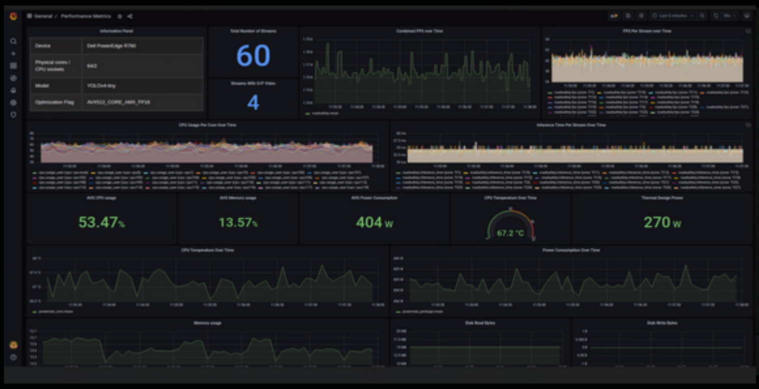
Case Study

Scalers AI™ Traffic Safety Solution

Our smart cities solution uses artificial intelligence and computer vision to monitor traffic safety in real-time. By analyzing video footage from cameras positioned at key locations, the system is able to identify potential safety hazards such as illegal lane changes on freeway on-ramps, reckless driving, and vehicle collisions. When a potential hazard is detected, the system sends an alert to the appropriate authorities, who can then take action to prevent accidents and maintain the flow of traffic. The AI computer vision system is trained on a large dataset of driving scenarios and is able to accurately identify safety hazards even in challenging conditions such as low light or heavy traffic. This solution helps cities improve road safety and reduce the number of accidents, making it safer and more efficient for people to travel within and between urban areas.

Performance Insights

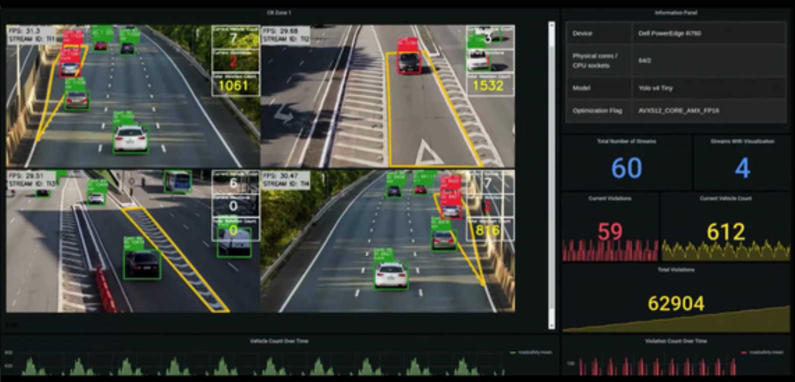
Traffic Safety Solution



General / Performance Metrics

Information Panel

Device	Dell PowerEdge R760
Physical cores / CPU sockets	64/2
Model	YOLOv4-tiny
Optimization Flag	AVX512_CORE_AMX_FP16



Information Panel

Device	Dell PowerEdge R760
Physical cores / CPU sockets	64/2
Model	Yolo v4 Tiny
Optimization Flag	AVX512_CORE_AMX_FP16

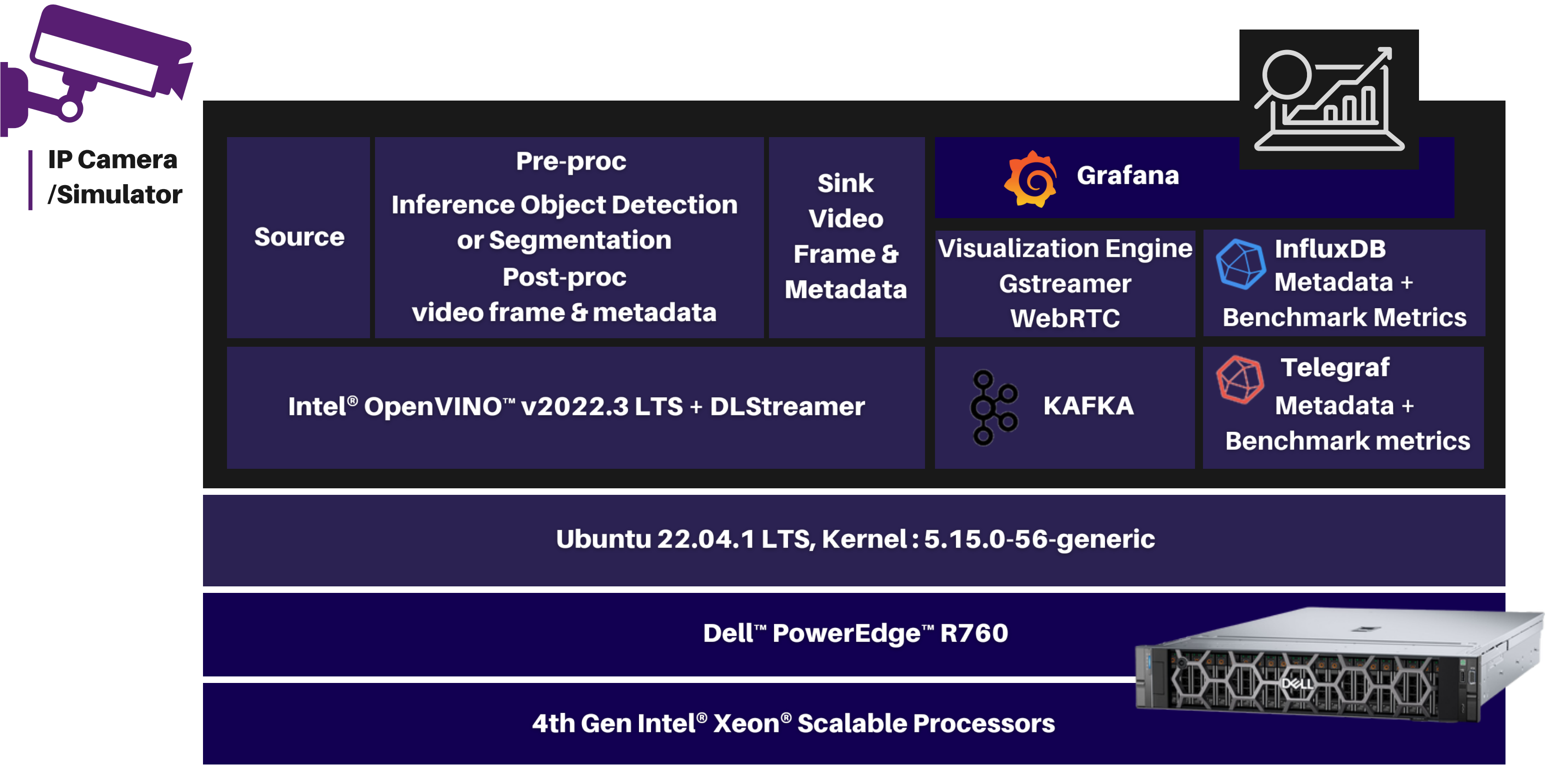
DASHBOARDS

Solution Architecture

In this computer vision AI solution using the RTSP camera, DL Streamer framework, Kafka, Grafana, and InfluxDB, the flow of information would be as follows:

- The RTSP camera captures video footage and streams it to the DL Streamer framework, which is responsible for ingesting and processing the data.
- The DL Streamer framework performs inference on the video data using a trained machine learning model to identify objects or patterns of interest.
- The DL Streamer elements send messages about the inference results to Kafka, a messaging system that acts as a buffer between the DL Streamer and other components of the application.
- Grafana, a visualization tool, retrieves the messages from Kafka and displays them in a dashboard for users to view and analyze.
- InfluxDB, a database for storing time series data, receives the messages from Kafka and stores them for use in application analytics. This data can be used to track trends and patterns over time and to inform decision making within the application.

Overall, this flow of information allows the computer vision AI application to continuously process and analyze video data in real-time, and to present the results to users in a clear and interactive way.



IP Camera /Simulator

Source

Pre-proc
Inference Object Detection or Segmentation
Post-proc
video frame & metadata

Sink Video Frame & Metadata

Grafana

Visualization Engine
Gstreamer
WebRTC

InfluxDB
Metadata +
Benchmark Metrics

Intel® OpenVINO™ v2022.3 LTS + DLStreamer

KAFKA

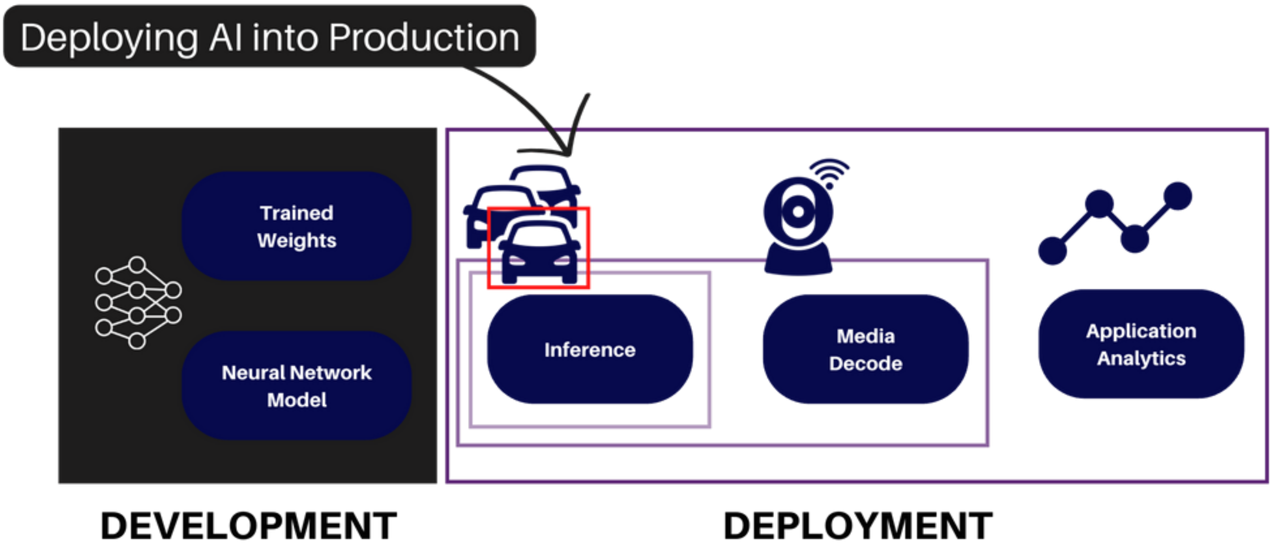
Telegraf
Metadata +
Benchmark metrics

Ubuntu 22.04.1 LTS, Kernel : 5.15.0-56-generic

Dell™ PowerEdge™ R760

4th Gen Intel® Xeon® Scalable Processors

AI Model Deployment



*Performance varies by use case

In order to deploy a trained AI model into a "live" solution, you will need to integrate with existing and new IP Cameras to ingest each stream of video data and then apply inference and use case specific logic on top of the inference results. This additional (real-world) workload is typically substantial and hence part of our solution testing methodology for realistic understanding of the system performance characteristics.

Our Solution Testing Methodology

- The workload and test cases were designed to maximize CPU utilization, ensuring that it was at least 90% throughout the scenario.
- Two Dell™ servers with different CPU models were used in the testing: Dell™ PowerEdge™ R750 with Dual Socket Ice Lake CPUs and Dell™ PowerEdge™ R760 with Dual Socket 4th Gen Intel® Xeon® scalable processors CPUs.
- The testing was done using the Intel® OpenVINO™ benchmark and DLStreamer benchmark, and system performance was monitored with Linux System tools.
- The AI model used in the testing was YOLOv4 Tiny from the Intel® Model Zoo and computation was in int8 format.
- The testing included scenarios with and without the use of AMX optimization software.
- The tests were run using 128 streams in parallel, with a source video resolution of 1080p and a bitrate of 8624 kb/s.

Performance varies by use case, model, application, hardware & software configurations, the quality of the resolution of the input data, and other factors. This performance testing is intended for informational purposes and not intended to be a guarantee of actual performance of an AI application.

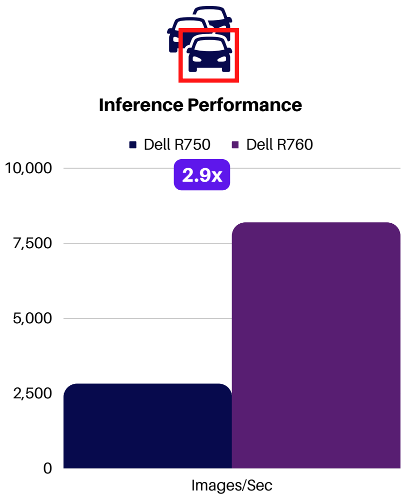
Performance Insights

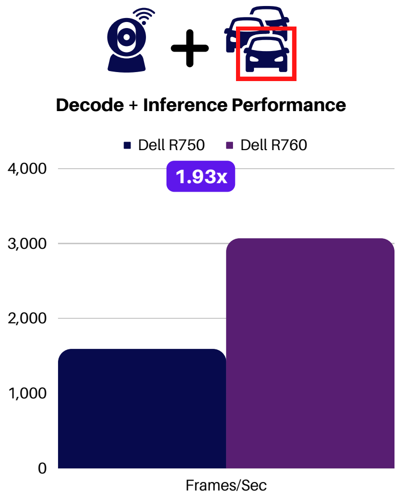
- The PowerEdge™ R760 system showed significantly better performance compared to Dell™ PowerEdge™ R750 system for AI inference and decoding tasks using the Tiny YoloV4 model with INT8 precision and the Intel® OpenVINO™ framework. Dell™ PowerEdge™ R760 system had a 2.9x improvement in AI inference on images and a 1.93x improvement in AI inference and decoding on video at 1080P resolution. When running Scalers AI application, which includes AI inference, decoding a video stream, and application services, Dell™ PowerEdge™ R760 system had a 1.88x improvement in performance.
- Dell ™ PowerEdge™R760 system showed significantly better performance than Dell™ PowerEdge™ R750 system in AI inference tasks: 2.9x improvement on image inference
- 1.93x improvement on combined 1080p video decode + inference
- In an AI application that combines inference, decoding, and application services, Dell™ PowerEdge™ R760 also had a 1.88x improvement in performance over Dell™ PowerEdge™ R750.

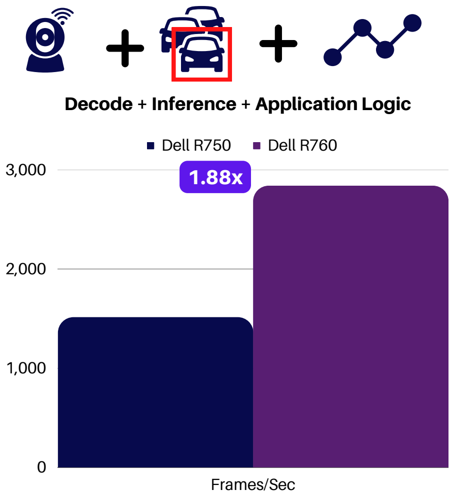
2.9x

IMPROVEMENT ON AI INFERENCE

System level performance from Dell™ PowerEdge™ R750 to Dell™ PowerEdge™ R760 using Intel® Deep Learning Boost with AMX instructions on INT8 & Intel® OpenVINO™ Framework Yolov4


Inference Performance


Decode + Inference Performance


Decode + Inference + Application Logic

> *"We recommend a maximum of up to 90 streams (1080p) running AI on Dell™ PowerEdge™ R760 and up to 50 streams (1080p) on Dell ™ PowerEdge™ R750 for optimal performance of a similar application on a single host with a similar configuration"*
>
> \- Chetan Gadgil, CTO at Scalers AI™

Conclusion

Dell™ PowerEdge™ R760 servers, equipped with 4th Gen Intel® Xeon® scalable processors , are up to the task of handling a wide range of our applications, including AI workloads that we would normally leverage discrete AI accelerators. Dell™ PowerEdge™ R760 is ideal for real-world applications that demand top-notch performance across a variety of workloads, from AI and complex analytics to user interfaces and experiences. Dell™ PowerEdge™ R760 has us covered for our industry specific solutions for both development including transfer learning and deployment with the need to handle AI, data management, web applications, media processing or even something else.

Fast track development with access to the solution code

Save hundreds of hours of development with the solution code. As part of this effort Scalers AI™ is making the solution code available.

Reach out to your Dell™ representative or contact Scalers AI™ at contact@scalers.ai for access.

This project was commissioned by Dell Technologies™ and conducted by Scalers AI, Inc.