Dell™ PowerEdge™ XR EDGE AI SMART CITY SOLUTION

DELL™ POWEREDGE™ XR SERVERS WITH 4TH GEN INTEL® XEON® SCALABLE PROCESSORS OFFERS UNPRECEDENTED PERFORMANCE FOR VIDEO AI BASED APPLICATIONS IN CITIES & MORE

SCALERS AI

Purpose Built Computing is a must for deploying AI at the edge

Edge computing and AI are rapidly evolving technologies that are driving industry transformation across our retailers environments, manufacturing facilities, healthcare services, our cities and more.

By deploying computing closer to the point of data creation or service delivery we can optimize traffic flows, increase manufacturing yields, fast track medical test results, and make sure retail customers can find the right product at the right time.

To enable these use cases and more the compute must sit at the edge to ensure near-real time insights, comply with privacy and regulatory laws, ensure fault tolerance on some of our most trusted systems, and optimize costs by discarding irrelevant data at the edge.

1.9x

IMPROVEMENT on Scalers AI™ Smart City Solution

Gen on Gen Performance Improvement Using Intel® Deep Learning Boost

These environments can have extreme temperatures, dust, heavy vibration, and power limitations. Dell™ PowerEdge™ XR5610 & XR7620 powered by 4th Gen Intel® Xeon® Scalable Processors is purpose built with a rugged compact form factor and rich I/O suited to meet the environmental challenges of industry use cases at the edge.

Scalers AI™ saw 1.9x performance improvement on Dell™ PowerEdge™ XR5610 platform with Intel® DL Boost on our Smart City Solution. Further, we saw near linear per core scale as we added sockets.

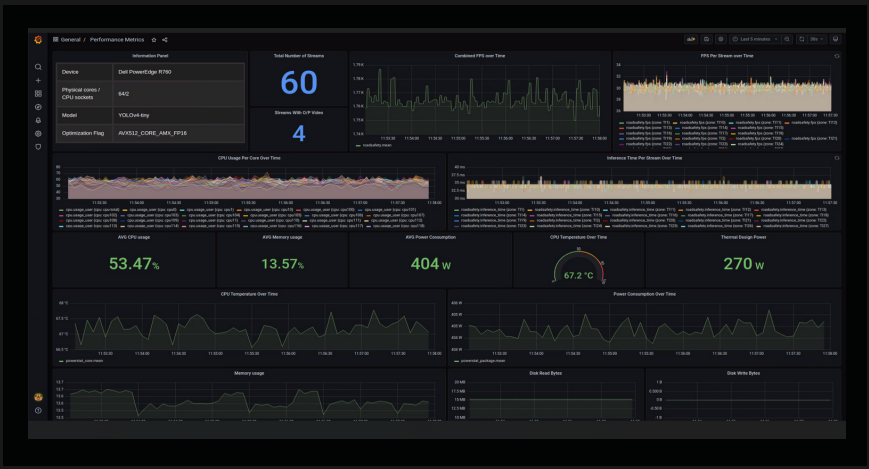
- Steen Graham, CEO at Scalers AI™

CASE STUDY

Scalers AI™ Smart City 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 lighting 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.

DASHBOARDS



Performance Insights

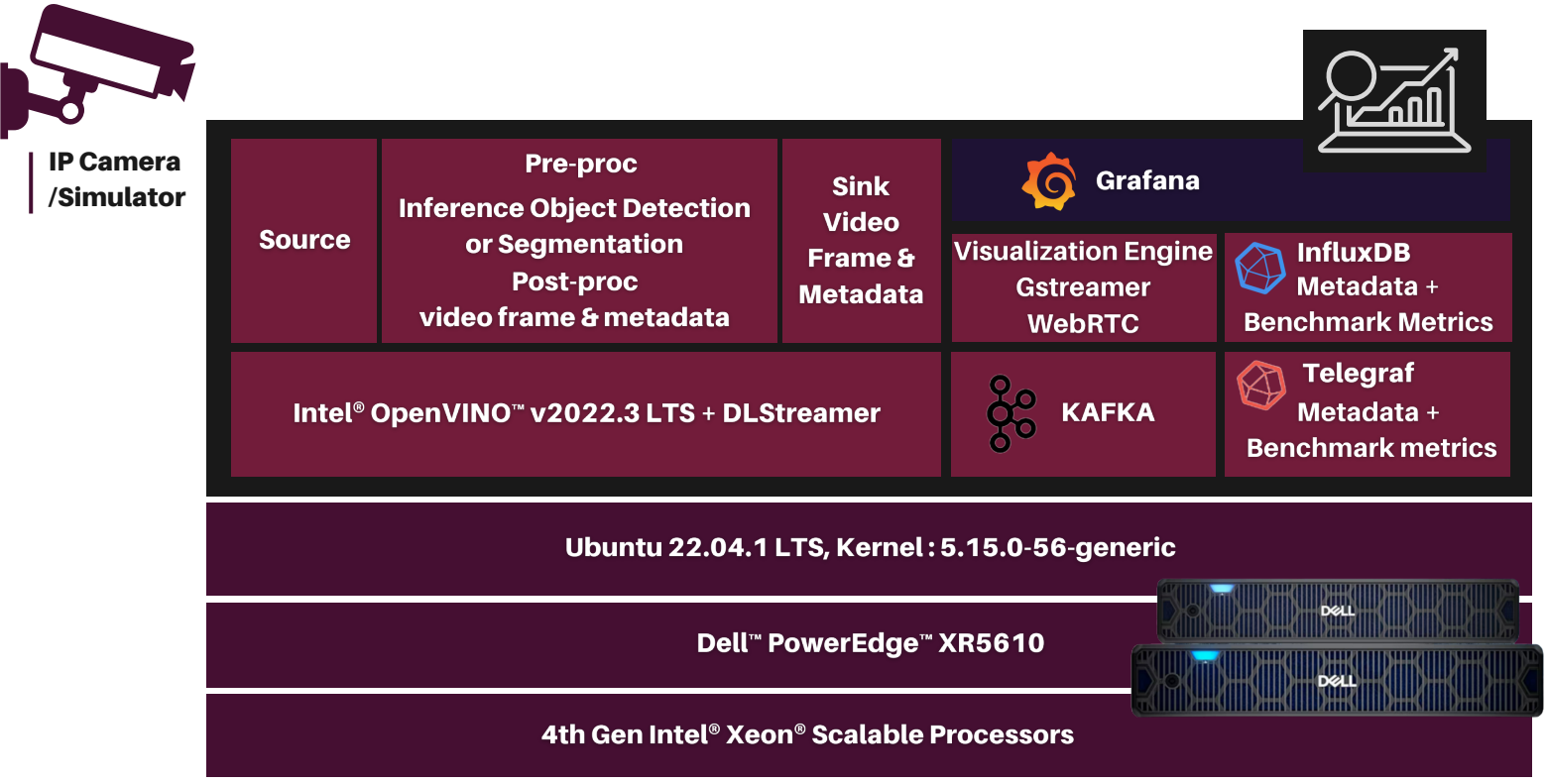
Traffic Safety Solution

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.



Performance Insights

Dell™ PowerEdge™ XR5610 system showed significantly better performance compared to Dell™ PowerEdge™XR11 system for AI inference and decoding tasks using the Tiny YoloV4 model with INT8 precision and the Intel® OpenVINO™ framework. Dell™ PowerEdge™ XR5610 system had a 2.5x improvement in AI inference on images and a 1.77x 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™ XR5610 system had a 1.9x improvement in performance.

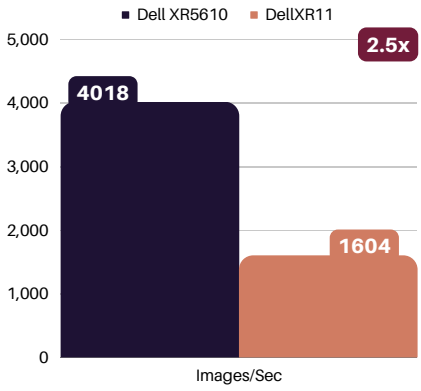
2.5x

IMPROVEMENT ON AI INFERENCE

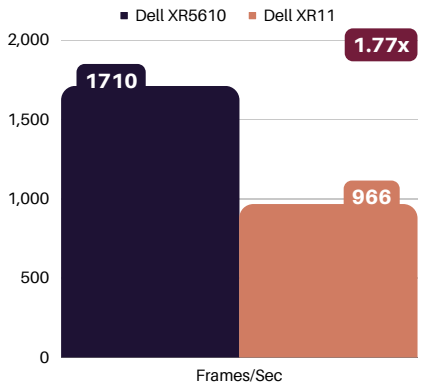
System level performance from Dell™ PowerEdge™ XR11 to Dell™ PowerEdge™ XR5610 using Intel® Deep Learning Boost with AMX instructions on INT8 & Intel® OpenVINO™ Framework Yolov4

Dell™ PowerEdge™ XR5610 vs XR11

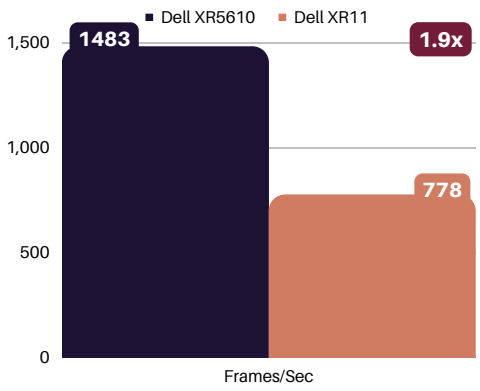
Inference Performance



Decode + Inference Performance

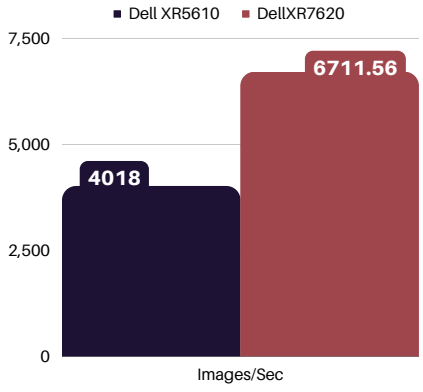


Decode + Inference + Application Logic

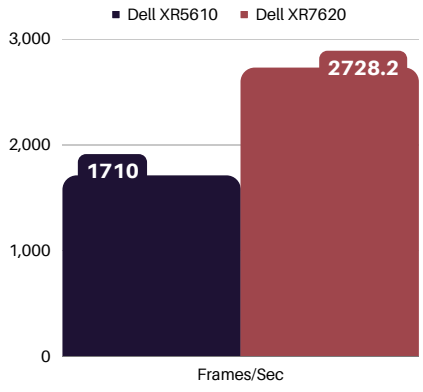


Dell PowerEdge XR5610 (32 cores) vs XR7620 (48 cores)

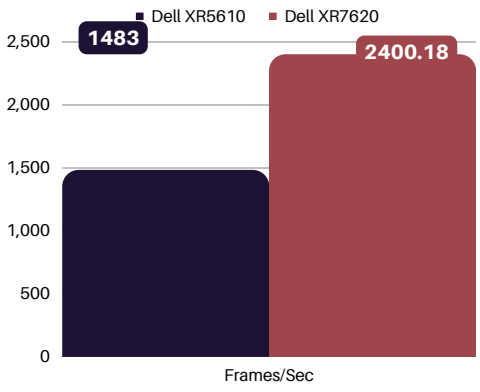
Inference Performance



Decode + Inference Performance



Decode + Inference + Application Logic



Results are show near linear scale on a per core basis. XR7620 available with 64 core as well.

Conclusion

Dell™ PowerEdge™ XR5610 server, 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™ XR5610 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. Importantly, Dell™ PowerEdge™ XR5610 has us covered for real-world deployments in harsh conditions requiring compact form factors that can withstand dust, vibration, extreme temperatures and humidity. For more compute Dell™ PowerEdge™ XR7620 offers two 4th Gen Intel® Xeon® Scalable Processors and we saw near-linear scale in our AI and video decode workloads making Dell™ PowerEdge™ XR ideal for high performance AI application deployment in harsh environments.

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.

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.

APPENDIX

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™ XR11 with Single Socket Ice Lake CPUs and Dell™ PowerEdge™ R5610 with Single Socket 4th Gen Intel® Xeon® scalable processors CPUs.
- The testing also included the Duel Socket Dell™ PowerEdge™7620 for scalability insights. PowerEdge™ XR also includes dual socket 64 core total systems.
- 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.

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.