SOLUTION BRIEF

Aupera: Smart Retail Video AI Analytics

INTRODUCTION

Retail "shrink", or loss due to theft, is at an all-time high. Shrinkage cost the industry $61.7 billion in 2020, accounting for almost 2% of retailers' bottom lines. AI video analytics applications can provide a powerful anti-shrinkage tool for retailers. However, these applications are complex and require deterministic low latency, requiring an architecture that can deliver performance while maintaining an advantageous TCO.

Aupera's Smart Retail solution, built on Xilinx Alveo® accelerator cards and Zynq™ SoCs, delivers low latency with unsurpassed performance at a low hardware investment, assuring best-in-class TCO.

Built on the integration of the Xilinx Video Machine Learning Streaming Server (VMSS) framework with Aupera's AupXStream video analytics pipeline processing, the solution is delivered on both Boston Stream AI and Aupera appliances.

SOLUTION OVERVIEW

Aupera Smart Retail supports video feeds from multiple full HD cameras for retail object identification and classification purposes for loss prevention applications. The unique proposition of this solution is its ability to run multiple inferencing models concurrently and deliver accurate results at a deterministic low latency, resulting in industry-leading low Total Cost of Ownership (TCO).

Features and Benefits

Full FPGA Accelerated Pipeline

- Real-time, low-latency, high throughput
- Concurrent multi-model inference
- Best-in-class accuracy

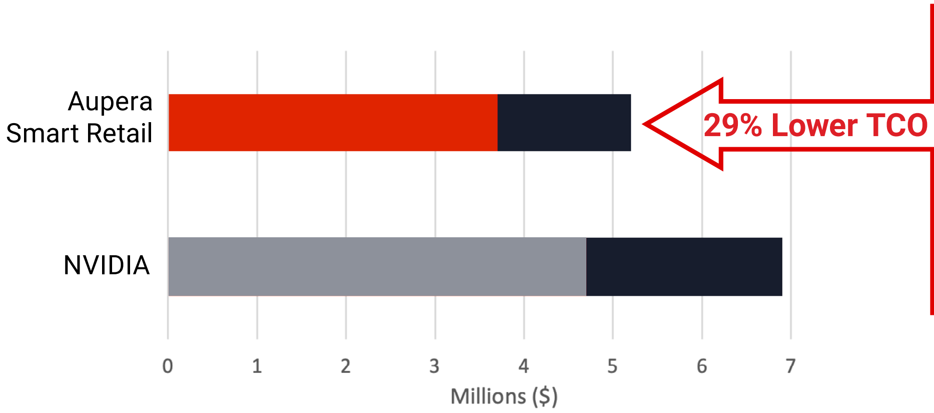
Best-in-Class TCO

- Fewer servers and Lower power
- Fast deploying, turnkey solution

Adaptable & Extensible Framework

- Ready for edge, datacenter, and cloud deployment
- Support for 3rd party plugins
- Standard API interface

TCO ADVANTAGE



- Two Xilinx Alveo cards vs four Nvidia T4s to support 32 cameras
- One server per location vs two required for NVIDIA
- Lower maintenance, power, and cooling costs

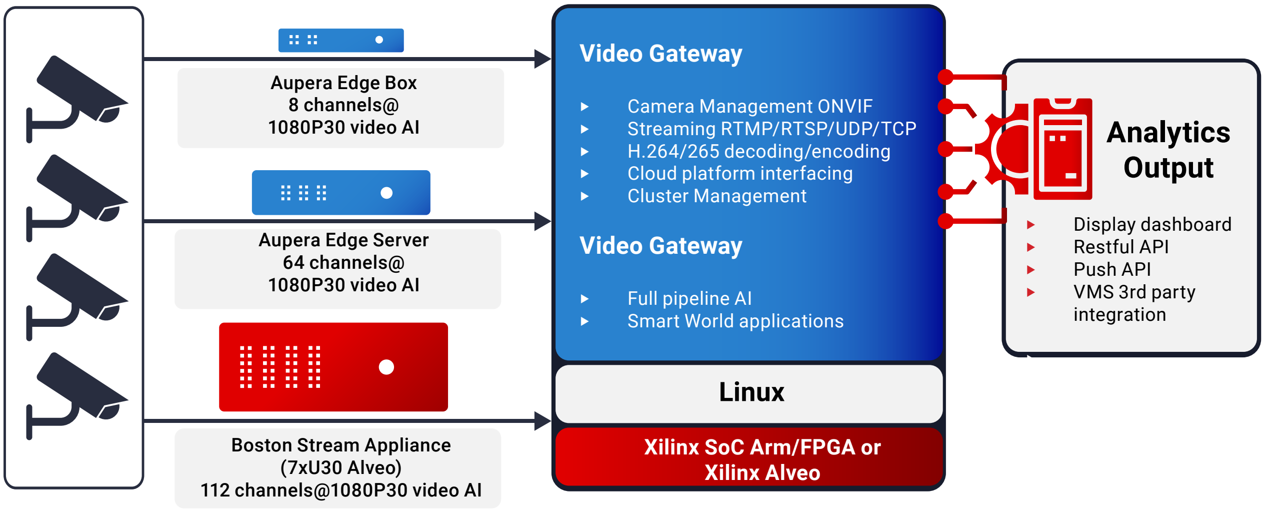
Notes:
Results from PoC at major global retailer
32 cameras/store Res:1080P30fps

AI model: Resnet 50 & TinyYoloV3
HW: (1 x U30 + 1 x U50) vs 4 x Nvidia T4

Adaptable. Intelligent.

SOLUTION DEPLOYMENT

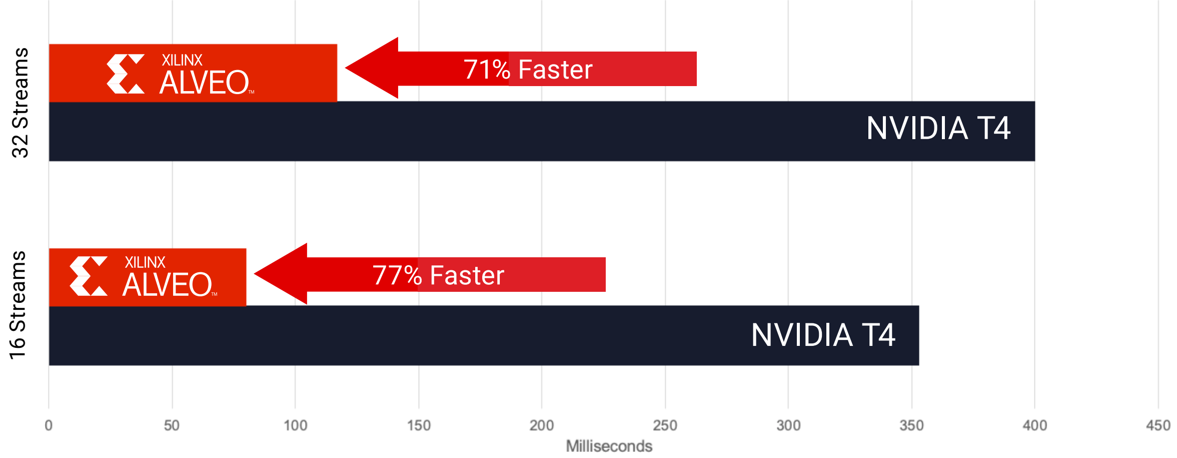
Aupera's video AI solution provides a full software stack that encompasses video gateway (video decoding/encoding/streaming & processing), ready to deploy AI models for inference acceleration and a standard API for easy integration with 3rd party software platform. The solution can be deployed on the edge or a regional data center to cater to multiple retail stores.



The Aupera Smart Retail solution supports up to 32 channels of 1080p at 30 frames per second (1080p30) on a Boston Stream AI appliance populated with one Alveo U30 and one Alveo U50 accelerator card. The Alveo U30 executes the video decode, scaling and preprocessing and the Alveo U50 accelerator runs Resnet 50 & TinyYoloV3 models needed for AI inference. The number of channels (video streams) and concurrent AI models can be increased by adding more Alveo accelerator cards (up to seven) to a single Stream AI appliance.

LOWER DETERMINISTIC LATENCY

The VMSS framework delivers end-to-end pipeline acceleration by running pre- and post-processing as well as inferencing on the FPGA fabric, delivering deterministic latency about 75% lower than the GPU-based competition in a typical large retailer 32-camera store deployment.



Notes:
Results from PoC at major global retailer
32 cameras/store Res:1080P30fps

AI model: Resnet 50 & TinyYoloV3
HW: (1 x U30 + 1 x U50) vs 4 x Nvidia T4

TAKE THE NEXT STEP > Visit www.xilinx.com/smartworld

Adaptable. Intelligent.