Accelerated Retail Analytics on T4 GPUs in Dell PowerEdge Servers

Matt Scott – CTO, Malong

10/24/19

MALONG

DELL EMC

ABOUT MALONG

Malong is a leading AI supplier for the retail industry, solving high-value business challenges with their RetailAI® solutions, primarily for loss prevention, built on world-class scientific research in computer vision.

CO-FOUNDERS : “MA” AND “LONG”

MATTHEW SCOTT
CTO, CO-FOUNDER

- Ex Microsoft Research, CVS
- 15+ yrs. R&D CV/ML, 100+ patents
- IEEE Senior Member
- Fast Company MCP100 Award

DINGLONG HUANG
CEO, CO-FOUNDER

- Ex Google, Microsoft, TripAdvisor
- Doctorate from Tsinghua U.
- World Economic Forum Young Global Leader 2019

DIFFERENTIATION – CORE TECH FOR RETAIL AI

BREAKTHROUGH DEEP LEARNING TECHNIQUE

LEADING RESEARCH IN PRODUCT RECOGNITION ALGORITHMS

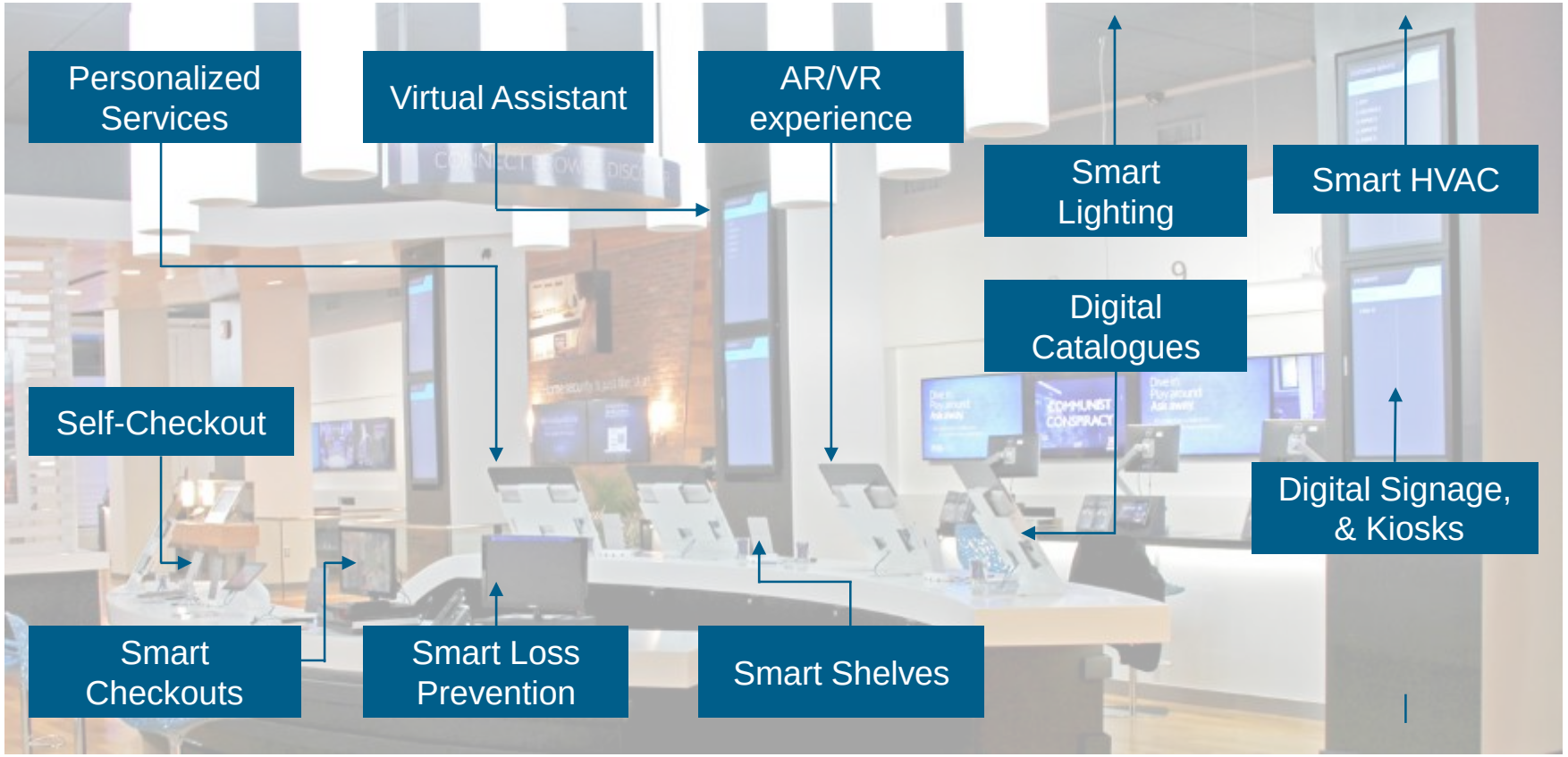
SOTA WEAKLY-SUPERVISED LEARNING CORE TECH

World-leading weakly-supervised deep learning. Won 1st place in Google's 2017 WebVision Challenge by large-margin (2nd place was a relative error rate of nearly 50%; a wide gap in our field). Mitigates the dependency on high-quality human annotated data. Key to win in retail is to be cost efficient on learning new SKUs. With this technology, Malong can learn from natural, noisy data directly, without re-labeling. Years ahead of others in its research and productization.

SOTA PRODUCT RECOGNITION CORE TECH

Product recognition algorithms are the core tech that Retail AI scenarios are built on and Malong is the world leader in this space for 4+ years. This is evidenced by Malong's continual advancement of the **state-of-the-art** on related scientific benchmarks, which are peer-reviewed and published in world-class academic literature. With this portfolio of algorithms, Malong's retail solutions are highly scalable. Related paper list in the appendix for due diligence.

The Intelligent Store



"Stores providing unique in-store experiences will thrive."

Reducing Loss Prevention at scale

Challenge:

- Reduce Shrinkage at checkout

Details:

- Label switching
- Mis-scans
- Sweet Heart deals - Price over-rides by employees

Solution Approach:

- Computer vision based inference of store products
  - Recognize the store products
  - Validate it with PoS Data for discrepancy

Benefits:

- Significant reduction in shrinkage

INTRODUCING
MIS-SCAN

Mis-scan is the **most common form** of scan fraud.

It is the act of **scan avoidance**.

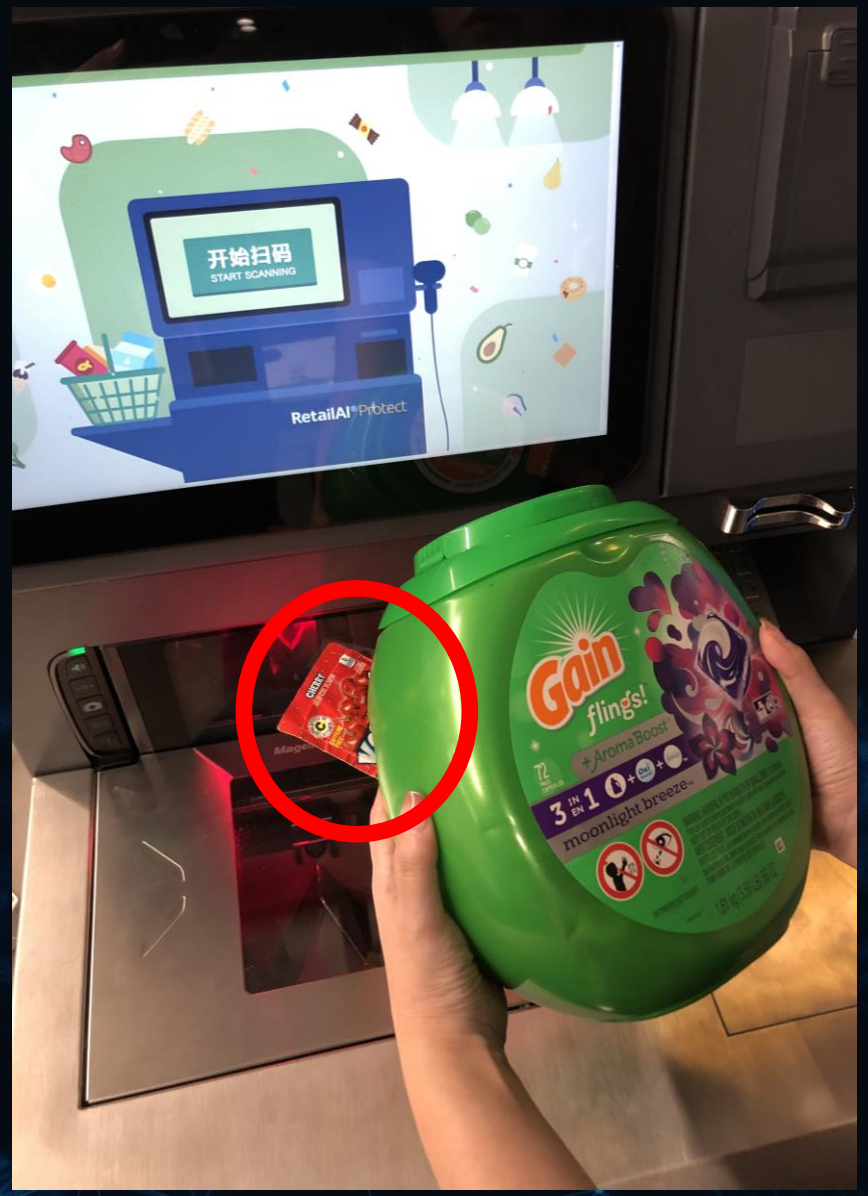
This is when a customer purposely (or accidently) avoids the scanning of their product barcodes. They may move the product in such a way across the scan bed to avoid the scan, or physically cover the barcode.

INTRODUCING
TICKET-SWITCHING

Ticket-switching is the act of scanning a less expensive UPC (the "**ticket**") to apparently scan a more expensive item (the "**switch**").

Although this form of scan fraud occurs less than mis-scan, the items stolen are often **much higher value**.

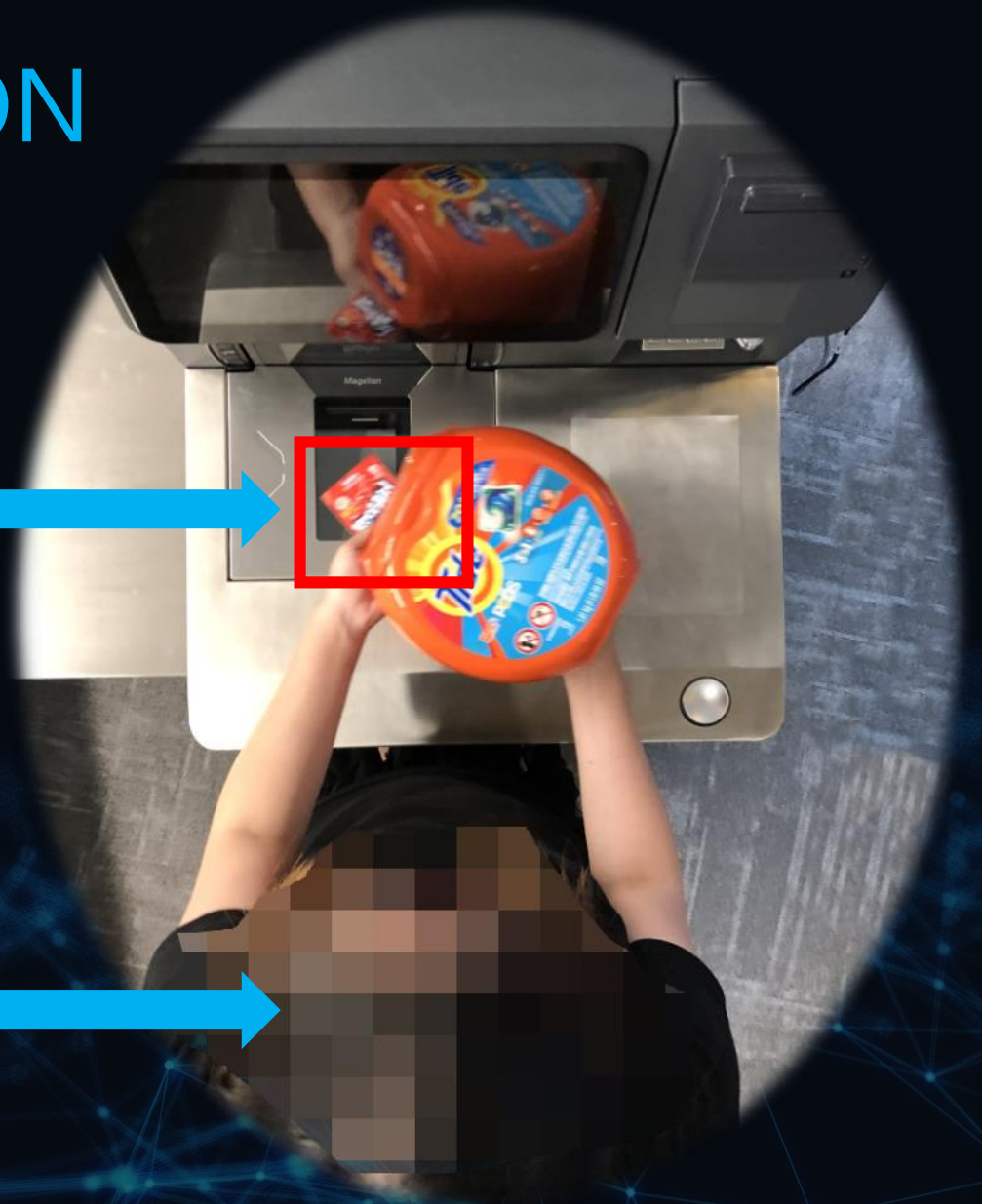
This can be performed in various ways: from physically attaching a barcode to a more expensive item, to holding the inexpensive barcode in hand at time of scan, to holding a less expensive item underneath a more expensive item.



RetailAI® Protect
OUR SOLUTION

Auto scan-fraud detection, triggers transaction pause, sends alert indication to staff

Customer privacy protected via real-time AI-based mosaic overlay (relevant for GDPR)

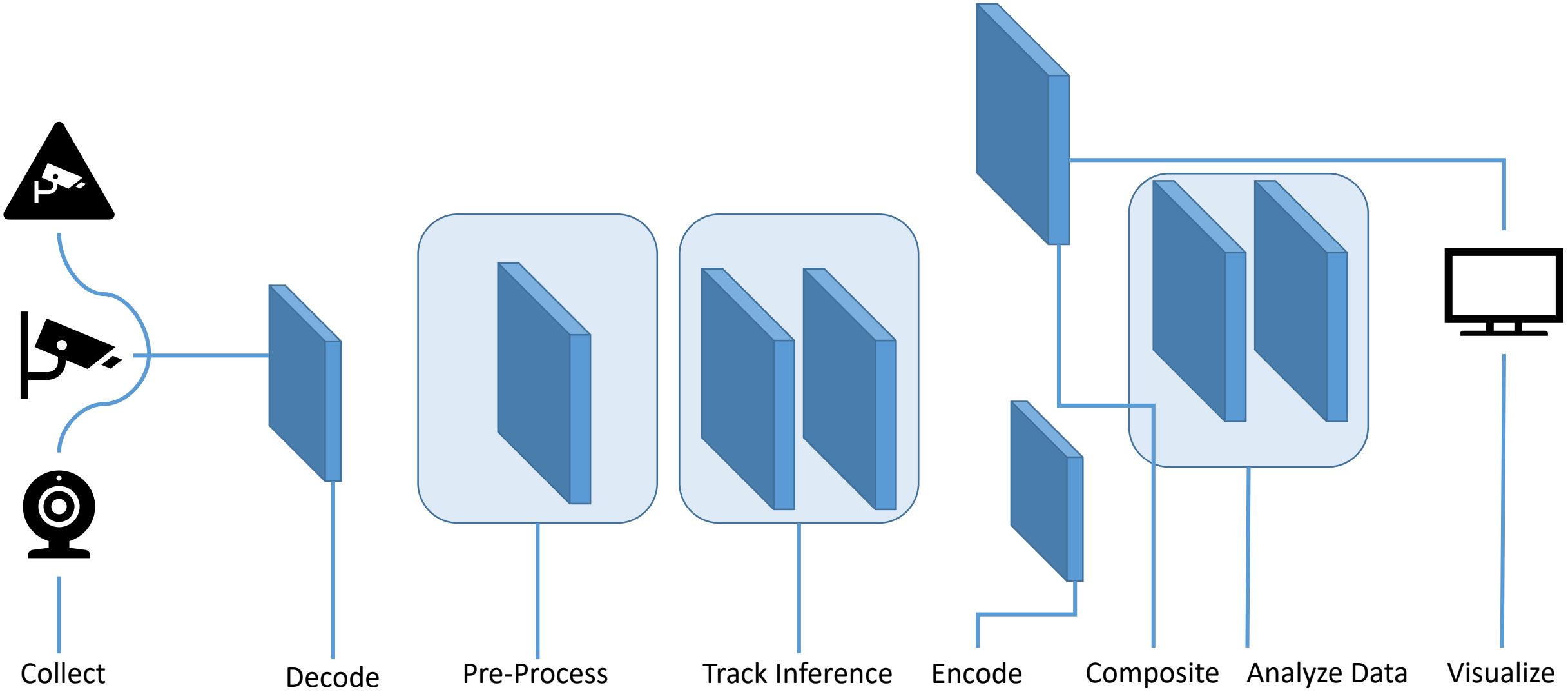


(overhead IP camera view)

- Catches both ticket switching & mis-scan fraud, simultaneously.
- Industry-leading in reducing false positives.
- Highly scalable and generalized - one AI model for all stores.
- Can be retro-fitted to existing store infrastructure, keeping costs low.
- World-class technical support with 24/7 professional services provided by Accenture.

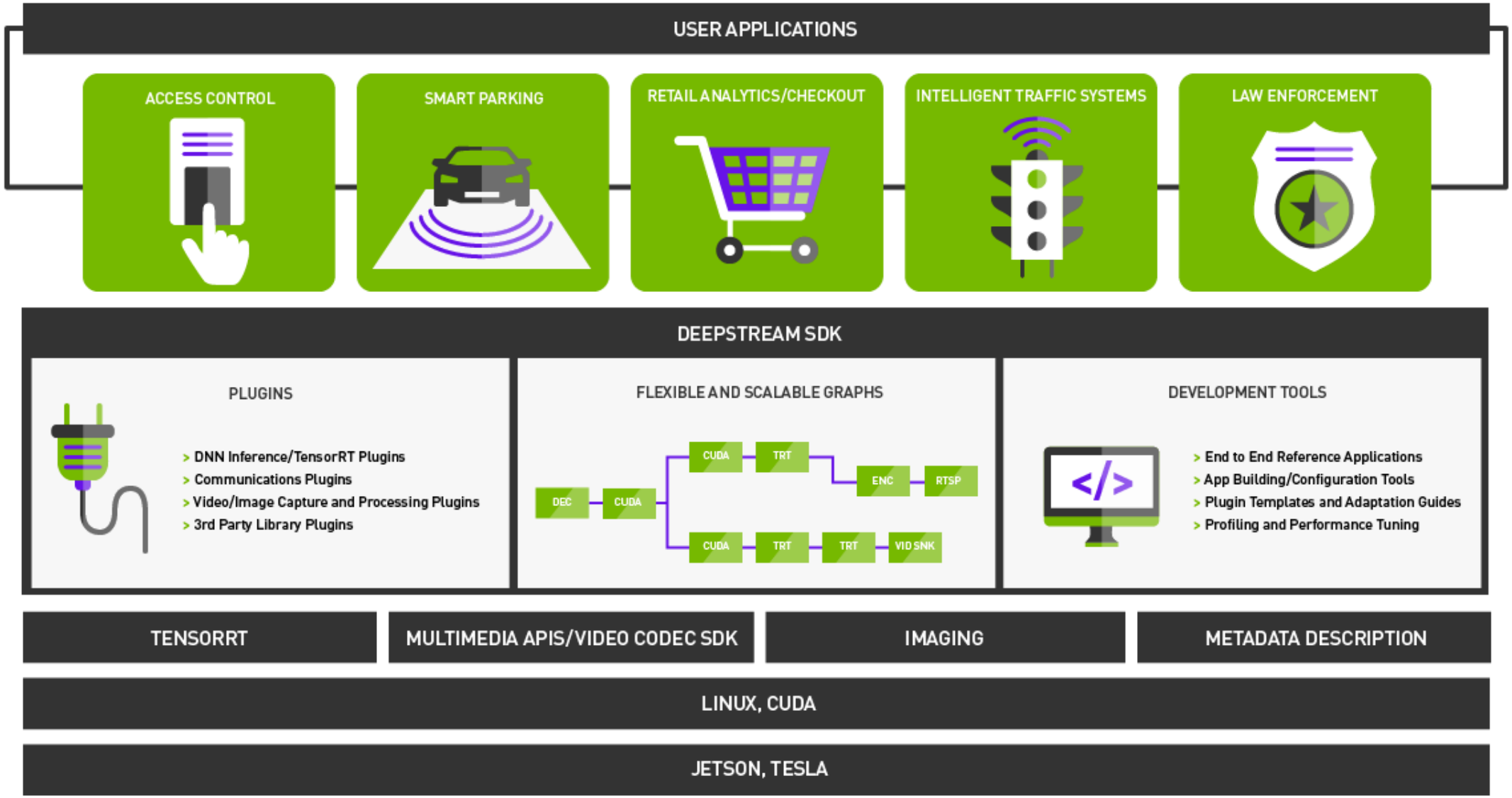
Demo video on following slide.

Video Analytics Flow Diagram

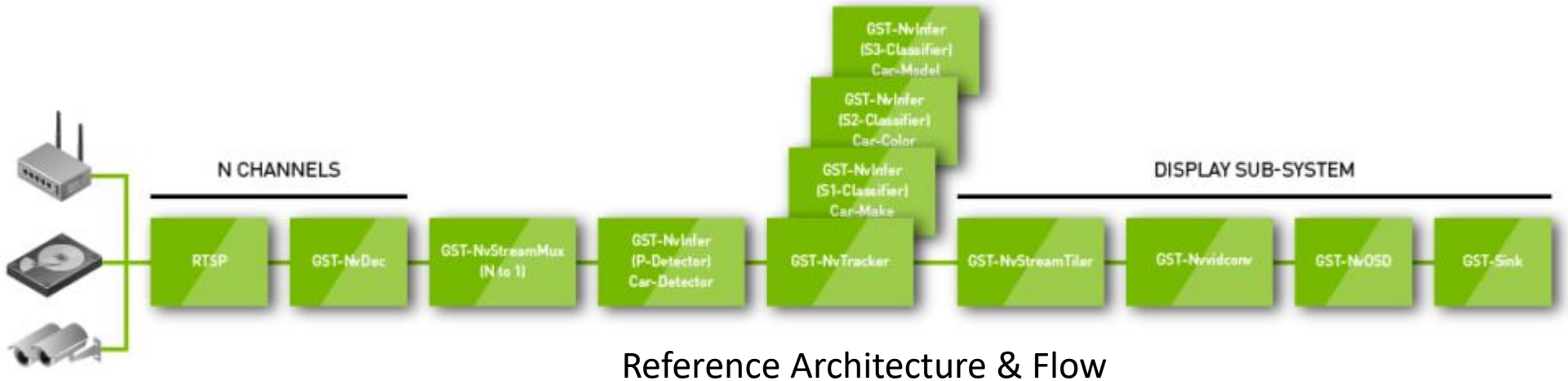


Video Analytics Software stack

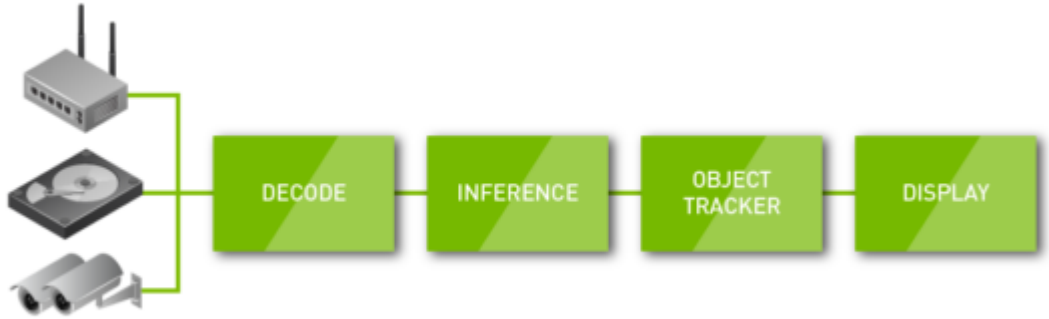
DEEPSTREAM SDK



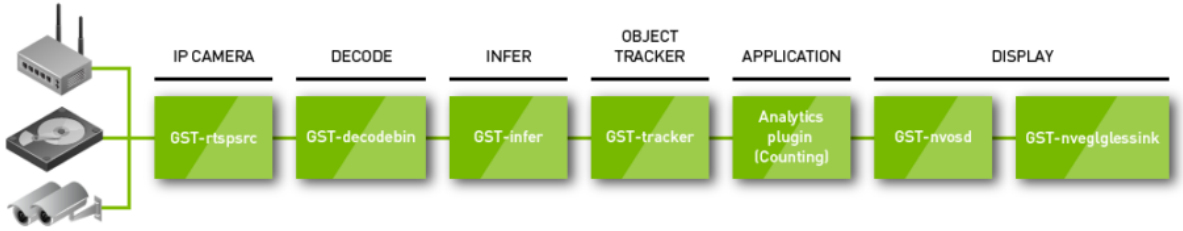
Reference Application Architecture & Flow



Reference Architecture & Flow



Logical flow for count application



Logical flow for tracking application

Source: https://devblogs.nvidia.com/accelerate-video-analytics-deepstream-2/

Accelerators for Inference

- There are wide variety of computing elements that can be targeted for Inferencing but there is a difference in terms of which computing elements are well suited for Inferencing.

Lots of different types

1) CPUs
2) GPUs
3) FPGAs
4) Specialized Accelerators

graphcore | intel nervana SYSTEMS | groq | Tenstorrent | habana

Specialized Accelerators

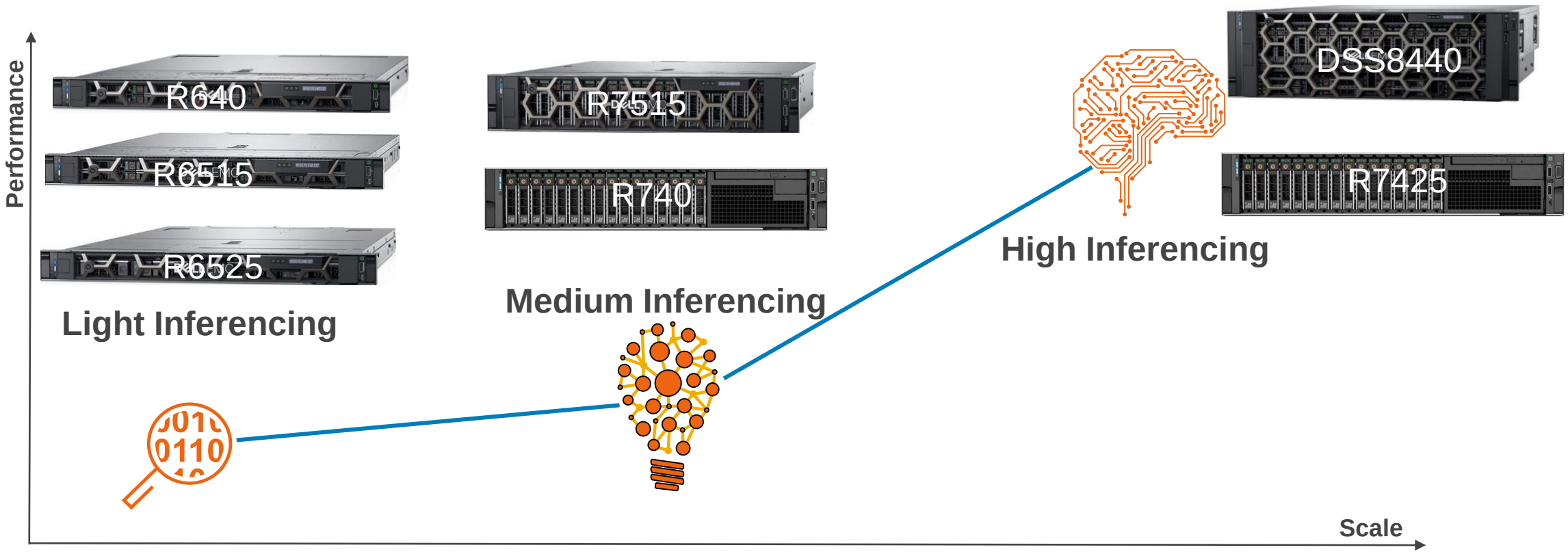
Why T4?

- The NVIDIA® Tesla® T4 is single-slot, low profile, PCIE Express Gen3 Deep learning accelerator card based on the TU104 NVIDIA graphics processing unit (GPU). The NVIDIA T4 has 16GB GDDR6 memory and a 70W maximum power limit. It is a passively cooled board.

- Tesla T4 is powered by NVIDIA Turing™ Tensor Cores to accelerate inference, video transcoding and virtual desktops. Turing Tensor Core technology with multi-precision computing for AI powers breakthrough performance from FP32 to FP16 to INT8, as well as INT4 precisions. It delivers up to 9.3X higher performance than CPUs on training and up to 36X on inference.

- Our opinion: T4 is best suited for Retail IVA because of optimal trade-off on teraflops, wattage, developer productivity, model future proofing, support, supply chain scalability.

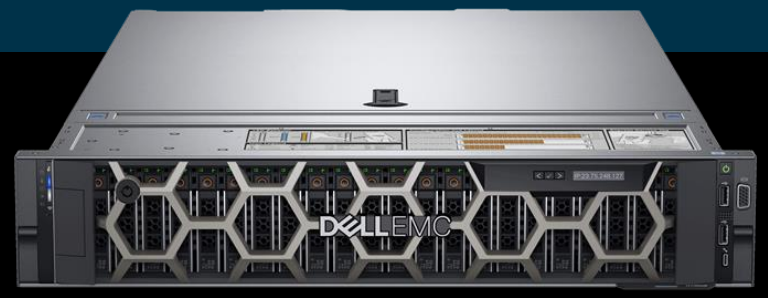
Inference Scaling – Based on # of cameras



Light Inferencing	Medium Inferencing	High Inferencing
Compute: **CPU+ low-profile GPU** **Qty :** 1-3 accelerators per node **Requirement :** Ability to handle ~15-45 cameras @1080p 30fps	Compute: **CPU & low power GPU or FPGA** **Qty :** 3-4 accelerators per node **Requirement :** Ability to handle ~45-60 cameras @1080p 30fps	Compute: **Multiple GPU or FPGA** Qty: 6 – 12 per node Requirement: Ability to handle ~30-90 cameras @1080p 30fps

PowerEdge R7425

The PowerEdge R7425 platform delivers outstanding TCO for data analytics, hybrid cloud and scale up software-defined deployments. Easily add extreme memory and storage capacity for low latency, data intensive workloads.



Targeted Workloads

- HPC
- VDI cloud client computing
- Database/in-line analytics
- Scale up software-defined environments

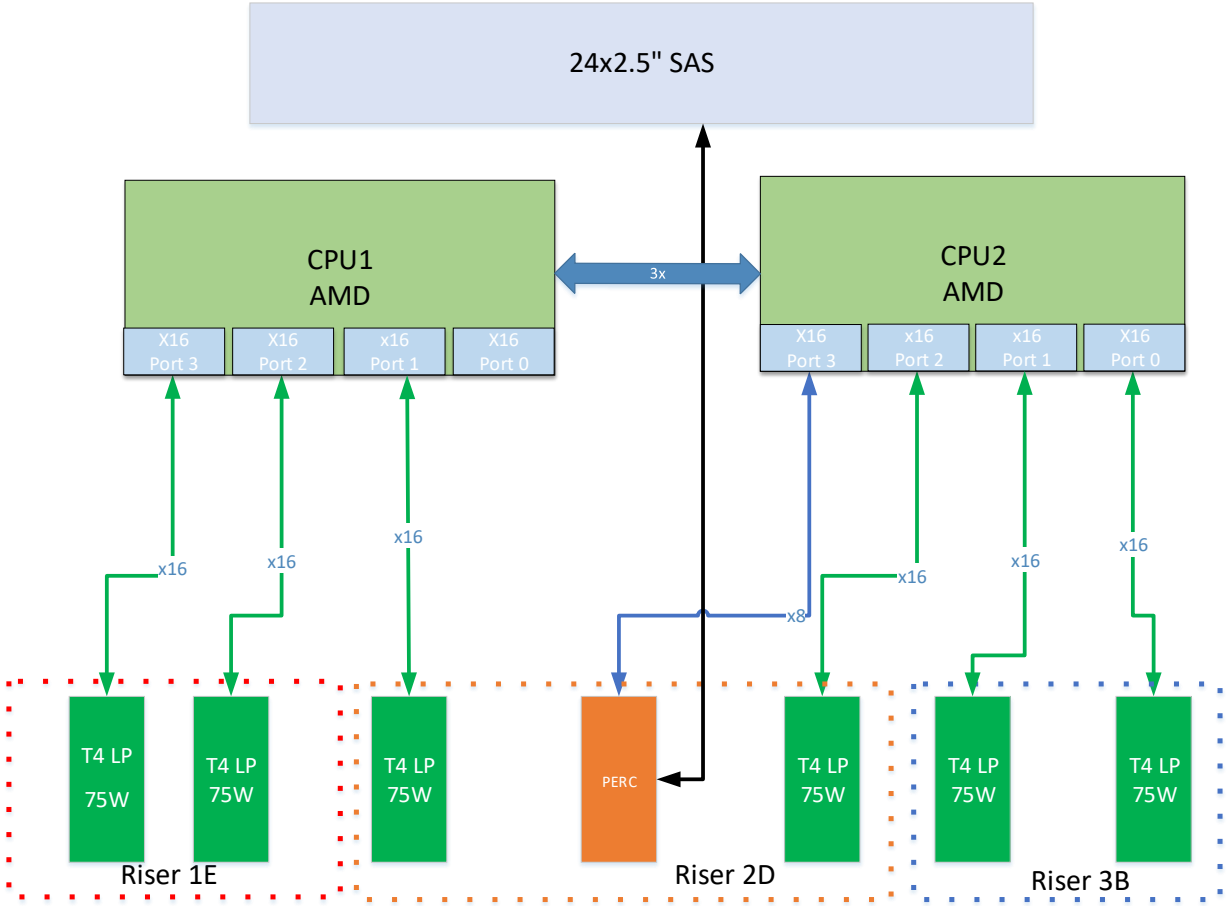
Key Capabilities

- 32 DIMMs
- Up to 24 NVMe drives
- Up to 64 cores
- Up to 128 lanes of PCIe

HIGHLIGHTS

- More cores for today's data intensive workloads
- NVMe for extreme performance storage requirements.
- High bandwidth, dense GPU/FPGA capability
- More number of memory channels optimized for data base and analytics workload flexibility

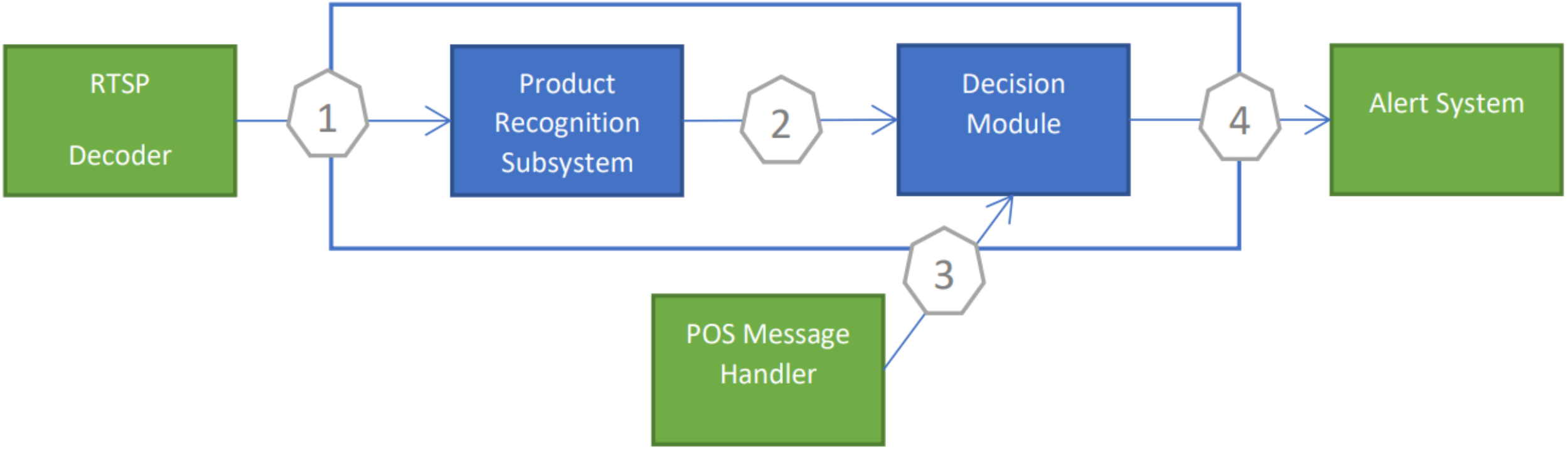
Inference 'Single Node' – R7425 'Inference' – Factory Install



Platform	R7425
Processor	2 x AMD Naples 7281 2.1GHz TDP 155W
Memory	16x 16GB:256GB DDR4 @ 2666 MT/s
Storage	4x 2.5"SAS Mix-use 400GB 512e
PERC	H740p
Networking	Broadcom 57414 Dual port 25Gb, SFP28 rNDC
GPU	Nvidia 6x T4 (75W Single wide)
Riser Config	Config 6 (1E+2D+3B) (6x16 + 1x8)
Power Consumption	Nom: Max: watts

- Extremely high I/O performance

Logical System Design



Optimization

- All models in the pipeline are heavily optimized with NVIDIA TensorRT to best leverage the NVIDIA T4 Turing architecture Tensor Cores with mixed precision accelerated inference, which significantly increases throughput and efficiency on the Dell EMC PowerEdge R7425.

- These optimizations provide for a 480%+ speed up when compared to not using TensorRT on the same hardware. Compared to running on CPU only, the difference with an optimized GPU version is a 99%+ reduction in processing time.

Processor	Speed of Core Model	% Difference to Optimized
CPU	4.7 seconds	-99.36%
NVIDIA T4 GPU (without optimizations)	0.175 seconds	-82.8%
NVIDIA T4 GPU with TensorRT optimizations	0.03 seconds	N/A

Summary

- NVIDIA T4 GPUs and technology stack (Metropolis, EGX) combined with the PowerEdge R7425 is a powerful platform that provides high performance, end-to-end intelligent retail analytics capability that can scale.

- Leveraging this software + hardware as an edge node with Malong RetailAI®, brings to bear industry-leading, advanced computer vision technology to help solve retail business problems in real-time right there in a store.

- With Malong RetailAI®, retailers stand to benefit from significant reductions in shrink loss and minimized inventory data accuracy issues.

Thank you.