# Dell EMC Ready Solutions for AI: Retail Loss Prevention

February 2021

H18172.3

White Paper

## Abstract

This white paper describes a solution that Dell Technologies has developed with hardware and software partners to make it easier and more cost-effective for retailers to employ new technology to reduce inventory loss at the checkout lane in stores.

Dell Technologies Solutions

MALONG

NVIDIA

DELL Technologies

## Copyright


Published in the USA 02/21 White Paper H18172.3.

# Contents

# Executive summary

## Business challenge

Recent data from the retail sector shows that 2019 worldwide annual revenues exceeded $25 trillion dollars[1]. In the United States alone, the retail sector employs over 15 million workers in more than 1 million retail establishments. The 2012 US Census found that retail establishments vary in size and activity from $10,000 in annual sales to $25,000,000 and more. Despite the size differences and diversity of the retail sector, all retail businesses share the impact from lost revenue that occurs when merchandise goes missing. Retail organizations have long recognized the financial impact of inventory that is deemed missing or lost. A 2017 Money magazine article estimated that inventory loss from stores cost the U.S. retail industry nearly $48.9 billion for the previous year[2]. Most of the financial impact of inventory loss in the retail sector occurs in stores at the point of sale (POS).

Preventing scan fraud activities such as scanning errors and UPC barcode switching at the POS is critical to reducing this major source of inventory loss. Dell Technologies in partnership with Malong Technologies and NVIDIA offer a solution that uses artificial intelligence (AI) to reduce inventory loss that is caused by either accidental or intentional behavior during checkout. The solution is powered by an in-store Dell EMC PowerEdge server that runs state-of-the-art product recognition technology from Malong RetailAI Protect. The solution can be used in most retail stores to reduce inventory loss during POS use at both self-checkout (SCO) and employee-staffed checkout lanes. Cameras that are positioned around an SCO kiosk capture real-time video of product items as they are scanned. The products are cross checked with purchases that are recorded by the POS system. NVIDIA T4 GPUs, part of the NVIDIA Tesla product line, perform video decoding and product recognition. The GPUs provide the throughput and fast response times that a retail checkout system requires. PowerEdge servers have a well-integrated management suite that makes them ideal as edge nodes in distributed retail environments. The Malong software offering is designed for distributed deployment, monitoring, and management through Microsoft Azure IoT Hub and the NVIDIA EGX stack.

## Audience

This document is intended for solution architects, system administrators, and others who are interested in using deep learning with advanced computing for loss prevention at retail stores.

[1] https://www.statista.com/statistics/443522/global-retail-sales/

[2] https://money.com/shoplifting-fraud-retail-survey/

## We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email or provide your comments by completing our documentation survey.

**Authors:** Bala Chandrasekaran, Phillip Hummel

**Contributors**: Malong Technologies and NVIDIA Corporation

**Note**: For links to additional documentation for this solution, see the Dell Technologies Solutions Info Hub for AI and Data Analytics Workloads.

# Retail inventory loss

This section describes the categories of retail inventory loss and the traditional methods that are used to prevent or reduce this loss.

## Categories of retail inventory loss

Industry experts recognize five major categories of retail inventory loss:

- Internal theft (employees)
- External theft (shoplifting)
- Paperwork and operational errors
- System issues
- Supplier fraud

Most studies show that total internal losses are slightly larger than the total external loss. Most inventory loss in the retail sector occurs at the POS. The combined loss from the warehouse, from the store and from supplier fraud is less than the loss from the actions of both employees and customers at the POS.

POS systems can be deployed for use by employees in traditional staffed checkout lanes or by customers at SCO lanes. The retail industry has long been aware of the occurrence of internal loss at staffed checkout lanes. This loss is a result of employees servicing friends, family, or coworkers, a practice referred to as “sweethearting.” More recently, the increasing popularity of SCO options in the retail sector is creating an even higher risk for both accidental and intentional inventory loss at the POS.

Whether accidental or intentional, methods that reduce the payment that is owed at checkout include:

| Mis-scanning an item | Ticket switching |
| --- | --- |
|  |  |

- **Mis-scan**—Either an employee or customer bypasses the scan device to avoid adding the item to the bill. Also, the employee or customer might inadvertently scan the wrong area of the packaging and miss the UPC barcode or obscure the UPC barcode with part of their hand.
- **Ticket switching**—A customer scans the UPC barcode of an item of lesser value in place of a more expensive item.

  A UPC code from the lower-priced item must be passed between the scanner and the higher priced item so that to a casual observer the transaction looks normal. A trained security professional might be able to detect the behavior, but the industry sees many benefits if an automated security system is deployed to detect such behavior.

  The three most common approaches to ticket switching are:

  - Removing a UPC barcode from a less expensive item and using it to cover the UPC barcode of an expensive item or placing it on a different side of the item
  - Attaching a UPC barcode from a less expensive item to the hand
  - Positioning a less expensive product under the more expensive product as both products are moved over the scanner
- **Cart-based**—Items can be accidentally or intentionally left on the bottom or inside the cart. It is difficult for scanner and checkout lane equipment alone to detect these unscanned items. Cart-based loss is a common cause of inventory loss and can occur at both staffed lanes and at customer SCO lanes.

  Typically, ticket switching occurs less often than mis-scanning and is used to remove items of higher than average value from the retail store. This type of scan fraud is of higher risk to would-be shoplifters because there is less plausible deniability. Additionally, this technique can become more prevalent as shoppers become more familiar with SCOs and how to exploit them.

## Traditional inventory protection in retail stores

Loss prevention from shoplifting in retail stores has, until recently, depended primarily on personnel and vigilance. From the smallest owner-operated establishments to the largest retail chains, shoplifters and employees have had only to avoid the visual detection of store personnel. Retailers have tried various combinations of uniformed security guards and nonuniformed personnel with mixed results. The widespread adoption of closed-circuit TV (CCTV) allows fewer personnel to monitor a larger sample of the store, but the basic tools are visual detection and questioning.

The development of low-cost radio frequency identification (RFID) tags and detectors added a new level of external loss detection for high value products. While the RFID industry predicted that widespread adoption of the technology in large markets like the retail industry would make it affordable for lower-cost items, the penetration remains limited today. Another limiting factor of RFID loss prevention is that the detection occurs only as the merchandise and customers exit the store. To protect the safety of their employees, many retailers are reluctant to direct employees to follow customers outside their stores.

Because of the rising cost of labor and the pressure from online retailers, store operators are reviewing staffing levels for dedicated loss prevention activity. The 2017 National Retail Security Survey[3] found that most survey participants are experiencing flat or declining loss prevention budgets, with 8 percent of respondents reporting decreases of 20 percent or more. The retail industry must invest in new methods that use more technology and require less labor for reducing internal and external losses.

Loss prevention at SCO lanes has traditionally been implemented by using scales. This technology attempts to mitigate loss by detecting the mismatch between the expected weight of the item on the sensor and the actual weight of the item. However, these systems have a high rate of false positive results. That is, the system can confuse legitimate buying as potential theft and falsely raise an alarm. This action typically locks the register and requires assistance from an employee. False positive results are the main problem with traditional anti-theft approaches at SCO lanes because they increase operational overhead while frustrating and annoying customers. The industry has spent years trying to address these false positive results with scales but has been largely unsuccessful. Many retail stores have opted to disable the scale-based detection method in their current machines or not purchase this option with new machines. An anti-theft system that can both mitigate inventory loss and provide less false positives is most desirable to the industry now.

## Using real-time video analytics for loss prevention

The increasing accuracy of computer vision technology that is based on deep learning coupled with the declining cost of in-store technology is improving the cost-effectiveness of intelligent retail loss prevention solutions. High-resolution cameras, computers, data storage, and networking together with new AI-powered software can detect both customer and employee behavior that is associated with inventory loss, while keeping false positive results to a minimum. High-resolution cameras give employees a clearer picture over a wider viewing area. However, this solution is still labor-driven.

The real breakthrough technology for loss prevention comes from intelligent software that can be placed in-store. Large retailers and third-party solution providers are investing in the development of computer vision for loss prevention applications. Advances in deep learning models that are trained by using massive amounts of data and large clusters of computers in remote data centers are being applied to retail loss prevention use cases. The outputs from these “trained models” can be used to generate real-time in-store alerts by using less expensive computers that are often found in retail stores for operations such as inventory and POS data aggregation.

High-resolution cameras and software models have been developed exclusively for use in both staffed and SCO lanes. One solution for detecting mis-scans and ticket switching uses data from both the scan terminal and a video stream of the activity at and around the checkout area. The system software attempts to match the items that are being scanned at the terminal with what the camera detects at the checkout lane. Items that are identified in the video stream but do not have a corresponding UPC barcode scan on the terminal are potential mis-scans or ticket swaps. These solutions can also detect items that are left in or under the shopping cart.

[3] https://nrf.com/sites/default/files/2018-10/NRSS-Industry-Research-Survey-2017.pdf

Many other applications are being developed for loss prevention in other areas of the store. For example, a series of video frames that show merchandise being placed in a pocket or concealed beneath a garment easily identifies abnormal shopping behavior. Another possible scenario is alerting security personnel if a person loads a shopping cart with an abnormally high number of expensive items. This behavior might indicate that the customer is preparing to avoid all checkout lanes and proceed directly out of the store in a classic “grab and go” theft.

Solutions that are based on the use of intelligent software models that are developed with deep learning techniques address several key issues for retailers. They reduce the need for additional loss prevention staff by operating largely unattended. These solutions can significantly reduce the false positive rate of traditional weight sensor-based systems. Smart vision systems can notify store personnel while a suspected shoplifter or employee commits theft during the transaction in contrast to systems like RFID that operate only at the door.

# Solution overview

The goal of this Ready Solution is to prevent fraud as it occurs without affecting the customer experience. It is a self-learning computer vision solution that can protect a wide range of stock keeping units (SKUs) by detecting mis-scans and ticket switching in near real time. The main design criterion for choosing solution components is determining if they can co-exist with and augment existing POS scanners.

Two scenarios in which the Malong RetailAI Protect solution can prevent retail loss are ticket switching and mis-scans ( see Traditional methods for reducing inventory loss). An overhead fixed-dome camera captures video of an item that is either not scanned or has a suspect UPC barcode. The video is decoded in the GPU and the decoded video is used as input to the Malong RetailAI model. The model then predicts the item's UPC barcode. If Malong RetailAI Protect determines that:

- The item was never scanned, after a set interval of time, Malong RetailAI Protect notifies the SCO system to raise an alert
- The scanned UPC barcode does not match the correct UPC barcode, Malong Retail AI Protect immediately notifies the SCO system to raise an alert

The retail associate overseeing the SCO lanes can then take the necessary action.

This solution is easy to put into use at scale for thousands of retail stores. The technology for this solution can be used at both SCO and staffed lanes.

## Loss prevention operation

A camera at a POS captures a video stream that is analyzed to detect the presence of products in the scanning area. The Real Time Streaming Protocol (RTSP) is used as the streaming protocol for video in the solution. The RSTP stream is received and decoded using one or more NVIDIA GPUs, as shown in the following figure:

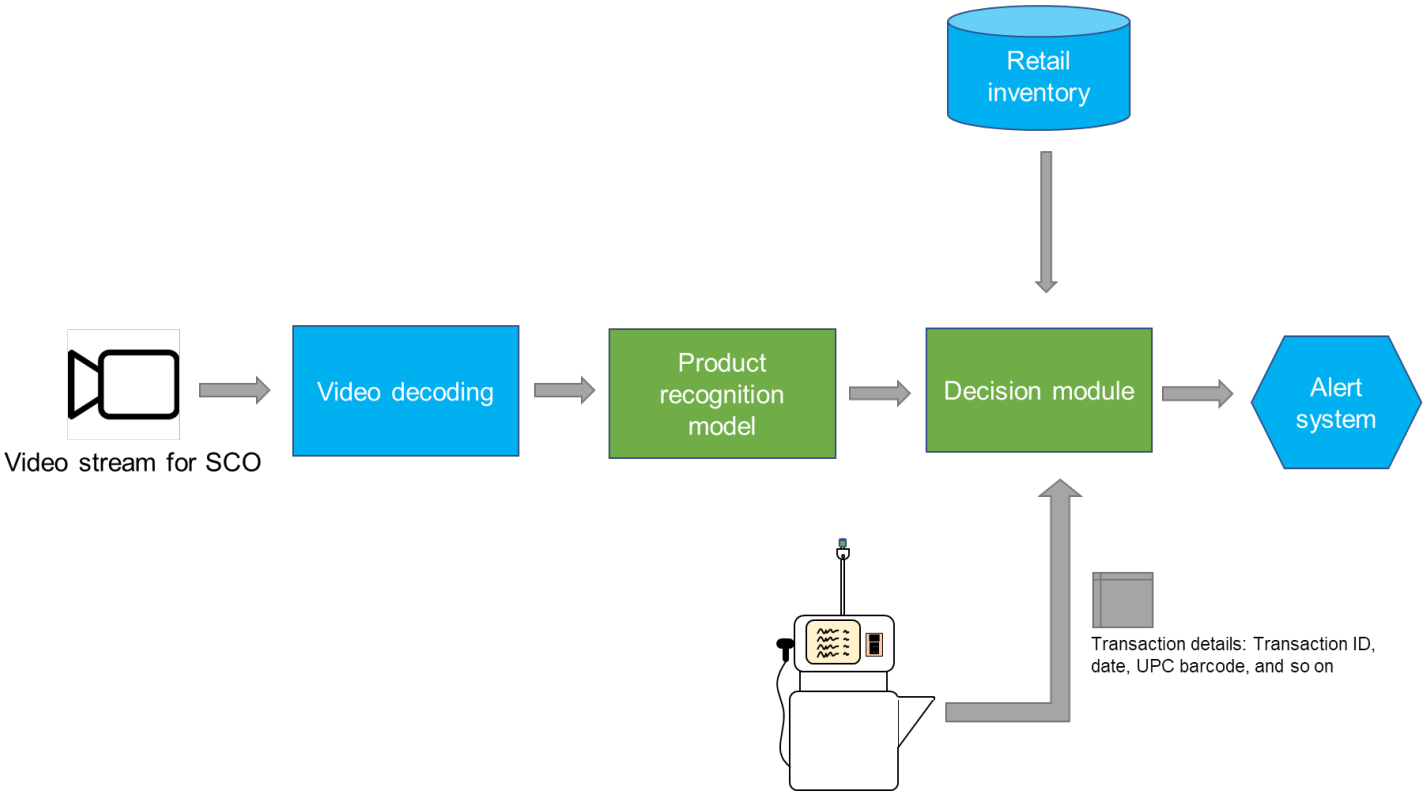


Figure 1. **Simplified operation overview of loss prevention system**

This Ready Solution for Retail Loss Prevention supports cameras with up to 4k resolution. H.264 is a digital video compression standard that uses half the space of the standard for DVDs for equal quality video. H.264 and H.265 are supported for the coding and decoding. NVIDIA DeepStream SDK is used for decoding the video in the T4 GPU. The decoded video stream is then sent to the product recognition model.

In this solution, the product recognition model is a key differentiator from traditional asset protection approaches. A deep learning-based AI model from our software partner Malong Technologies is used for product recognition. The decoded video stream from the T4 GPU is fed into the input layer of the deep learning model. The AI model outputs a list of the most likely product matches in that video frame.

Simultaneously, the retail shopper scans the UPC barcode of the product. The UPC barcode scan is fed into the decision module. The decision model compares the outputs from the deep learning model and the UPC barcode-scanned product from the retail database:

- If there is a match, no action is taken.

- If there is a mis-match between the POS signal and the visual outputs from the deep learning model, a ticket-switching alert is triggered.
- If there are only visual outputs from the deep learning model and no corresponding simultaneous POS signal, an alert for a mis-scan is triggered.

The retail store then takes the appropriate action.

Importantly, the AI algorithm only considers the visual image of the item that is being scanned. It does not take into consideration the image of the shopper, ensuring customer privacy and no bias, which is a fundamental principle in responsible AI.

## Edge architecture for retail loss prevention

IT technology solutions that are deployed outside a traditional enterprise data center are often referred to as edge solutions. “Edge” is a networking term that is used to represent the connections where end-user equipment first links to a communications system. Cell phones, tablets, and portable laptops are some of the most common edge devices. Commercial and industrial applications in stores and plants, which are the furthest away from the corporate data centers and high-speed data networks, are also called edge applications. The term “edge architectures” describes how we design solutions for many uses cases including retail in-store systems. The following figure shows the edge architecture for a retail store:

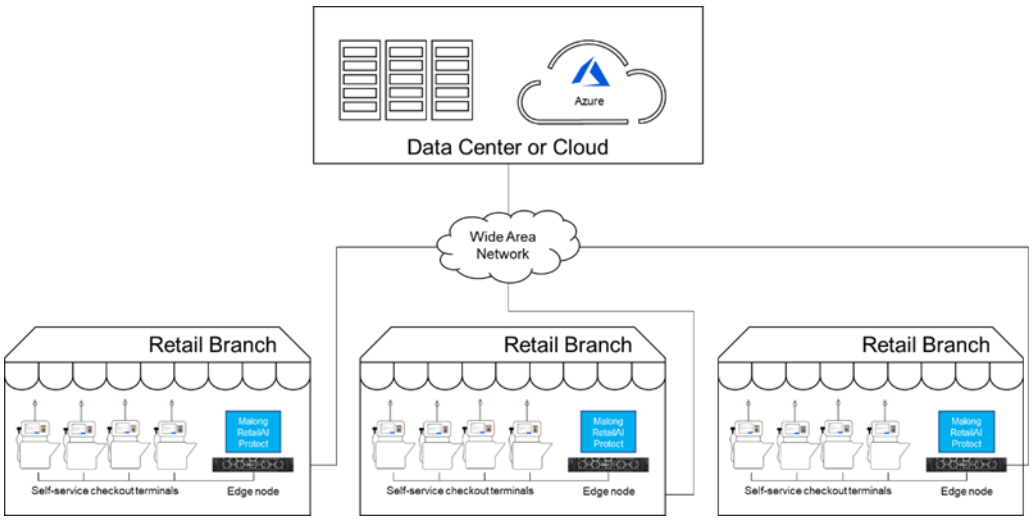


Figure 2. **Edge node components**

# Technology components

This section provides a detailed description of each of the components in this Ready Solution:

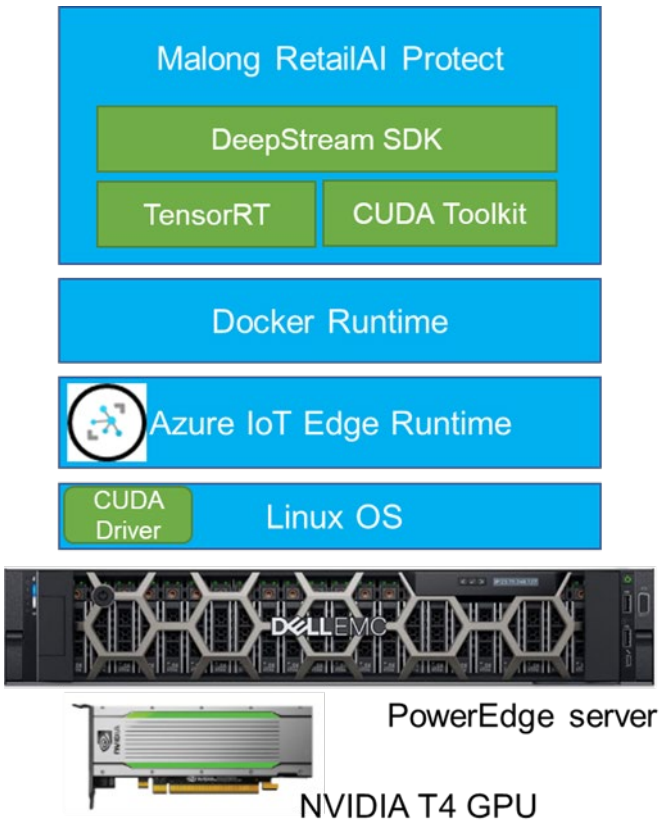


Figure 3. **Edge node components**

An edge node which consists of a single PowerEdge server with the following hardware and software components installed:

- **Linux operating system**—Ubuntu Server LTS is the preferred Linux distribution deployed on the VM server. Red Hat Enterprise Linux 7.5 or later is also supported
- **Docker container runtime—**This Ready Solution uses container technology for ease of packaging and deployment.
- **NVIDIA components**—This solution uses the following NVIDIA components (see NVIDIA Tesla GPU and software components):
  - T4 GPU for video decoding and inference.
  - CUDA driver and toolkit for interfacing with the GPU.
  - DeepStream SDK, which is an analytics toolkit for AI-based video processing and analysis. The toolkit includes TensorRT SDK for high-performance deep learning inference.
- **Malong RetailAI Protect**—The AI model from Malong, which is deployed as a Docker container. This container includes DeepStream SDK and the CUDA toolkit.

Malong RetailAI Protect can also be installed in a VM. VMware ESXi server is installed on the PowerEdge server. Customers can install other applications as virtual machines to increase the server resource utilization.

- **Self-checkout terminals**
- **Microsoft Azure IoT Hub**

## Dell EMC PowerEdge servers

This Ready Solution uses PowerEdge servers for computing at the edge. PowerEdge servers are powered by 2nd Gen Intel Xeon Scalable processors or 2nd Gen AMD EPYC processors. These servers support NVIDIA T4 GPU controllers for video decoding and image recognition during deep learning model inference. For their retail needs, customers can choose one of the following rack server options for an optimum balance of compute, memory, and GPU performance:

**Table 1. Server options**

| Server model | Processor | Memory | Maximum number of T4 cards |
|---|---|---|---|
| PowerEdge R640 | Dual socket 2nd Gen Intel Xeon Scalable processors | 24 DIMMs | 3 |
| PowerEdge R740 | Dual socket 2nd Gen Intel Xeon Scalable processors | 24 DIMMs | 3 FI + 3 APOS on x8 slots |
| PowerEdge R740xd | Dual socket 2nd Gen Intel Xeon Scalable processors | 24 DIMMs | 3 FI + 3 APOS on x8 slots |
| PowerEdge R7515 | Single socket 2nd Gen AMD EPYC processor | 16 DIMMS | 4 |
| PowerEdge R6525 | Dual socket 2nd Gen AMD EPYC processor | 32 DIMMS | 3 |
| PowerEdge R7525 | Dual socket 2nd Gen AMD EPYC processor | 32 DIMMS | 6 |
| DSS 8840 | Dual socket 2nd Gen Intel Xeon Scalable processors | 24 DIMMS | 16 |
| PowerEdge XE2420 | Dual socket 2nd Gen Intel Xeon Scalable processors (maximum 150 W Gold) | 16 DIMMs | 4 |
| PowerEdge XR2 | Dual socket 2nd Gen Intel Xeon Scalable processors | 16 DIMMs | 1 |

For our recommendations for small, medium and large retail stores, see Configuration.

## NVIDIA Tesla GPU and software components

The NVIDIA T4 GPU and software components include:

- **NVIDIA T4 GPU**—A single-slot, low profile, PCIE Express Gen3 Deep Learning accelerator card that is based on the TU104 NVIDIA GPU. The T4 GPU has 16 GB GDDR6 memory and a 70W maximum power limit. It is a passively cooled board.

  The T4 GPU is powered by NVIDIA Turing Tensor Cores to accelerate inference, video transcoding, and virtual desktops.

  This Ready Solution uses T4 cards for video decoding and deep learning inference.

- **NVIDIA DeepStream SDK**—An SDK that delivers a streaming analytics toolkit for AI-based video and image processing, and multisensor processing. DeepStream is part of the NVIDIA Metropolis platform that enables building end-to-end services

and solutions for transforming videos and images to actionable insights. Relevant features for this solution include:

- Reduced memory footprint that results in enhanced stream processing density
- Integration with Microsoft Azure Edge IoT to build applications and services by using the power of Azure cloud
- Containerized deployment
- Plug-in sources for inference, message schema converter, and message broker plug-ins
- Support for heterogeneous cameras, segmentation networks, monochrome images, and hardware-accelerated H.264 and H.265 video decoding
- Support for TensorRT-based inferencing for detection, classification, and segmentation

- **NVIDIA TensorRT**—An SDK for high-performance deep learning inference. It includes a deep learning inference optimizer and runtime that delivers low latency and high throughput for deep learning inference applications. TensorRT can optimize neural network models that are trained in all major frameworks, calibrate for lower precision with high accuracy, and deploy to hyperscale data centers, and embedded or automotive product platforms. It is ideally suited for inference from video streaming, such as retail product identification used in this solution.

  TensorRT is built on CUDA, which is a parallel programming model from NVIDIA. TensorRT enables optimized inference for all deep learning frameworks by using libraries, development tools, and technologies in CUDA-X for artificial intelligence, autonomous machines, high-performance computing, and graphics.

  You can import trained models from every major deep learning framework into TensorRT. After applying optimizations, TensorRT selects platform-specific kernels to maximize performance on T4 GPUs in the data center, Jetson embedded platforms, and NVIDIA DRIVE autonomous driving platforms.

- **NVIDIA CUDA**—A parallel computing platform and programming model that was developed by NVIDIA for general computing on GPUs. With CUDA, developers can accelerate computing applications by harnessing the power of the GPUs. Applications and operations (such as matrix multiplication) that are typically run serially in CPUs can run on thousands of GPU cores in parallel.

## RetailAI Protect by Malong

Malong Technologies specializes in AI that provides computer vision technology for retail applications. Malong has developed a state-of-the-art computer vision system that uses deep neural network technology to identify retail products. This technology has been applied successfully to identify both accidental and intentional product scan errors. Malong RetailAI Protect provides a state-of-the art software solution for inventory loss at the POS, which is among the leading causes of inventory loss in brick-and-mortar retail environments. The technology works for both SCO terminals and staffed lanes.

Malong developed an algorithm called CurriculumNet, which is a breakthrough approach to learning directly from noisy and unbalanced visual data—the kind of data that is found in retail environments. With CurriculumNet, the model starts learning the more clearly defined and easier-to-recognize characteristics of a product. It gradually includes more

complex learning tasks, such as learning rare and hard-to-distinguish items, into the learning process.

The learning model with CurriculumNet follows three main steps:

1. **Initial features generation**—All the data is used to train an initial model. This model is used to understand the underlying structure and relationship of the images for each item.
2. **Curriculum design**—The entire training set is divided into three subsets on a scale from the easiest-to-recognize subset containing clean images with more reliable labels, to the most complex subset containing massively unreliable (noisy) labels.
3. **Curriculum learning**—Curriculum learning is based on human learning theory. This step uses a strategy where learning is ordered by increasing difficulty and training proceeds sequentially from easier tasks to harder tasks. First, the deep neural network is trained with a subset of images (categorized in the preceding step) that have clean images and correct labels. This training allows the model to learn basic but clear visual information from each category, serving as the fundamental features for the following process. The learning process continues by adding the second subset of data, which includes images with more significant visual diversity. This training enables the model to learn more meaningful and discriminative features from harder samples. Finally, model training continues with the addition of the noisiest data that contains many visually irrelevant images (not in one of the output classes) with incorrect labels. By using this method, we found that accurate recognition of images with highly unreliable labels increases and can improve the ability of the model to generalize.

These three steps enable the Malong algorithm to learn during normal operations without the need for additional human supervision for annotation. Manual labeling incurs one of the highest costs involved in implementing a deep learning-based solution and is therefore not feasible in large-scale retail scenarios. The Malong algorithm was tested in 2017 at a worldwide image recognition competition called WebVision at CVPR, the premier conference in computer vision. More than 100 scientific research organizations participated in the competition, which was held by Google Research. The Malong algorithm won first place by a wide margin, outperforming the second-place submission by a reduction in relative error rate of nearly 50 percent, as shown in the Challenge Results.

## Summary

Malong RetailAI Protect is the latest generation of a deep learning model designed for the retail store market. It builds on CurriculumNet to deliver a state-of-the-art algorithm performance. The AI model runs in an Intelligent Video Analytics (IVA) pipeline, using proprietary weakly supervised learning. In this model, noisy and imprecise data is used effectively to develop an AI algorithm that maps visual information about products to their UPC barcodes. As customers scan their items, the AI model compares visual images of the items to the scanned UPC barcodes. If there is a mismatch or if a mis-scan occurs, the AI model raises an alert by using a configured mechanism.

## SCO terminals

SCO terminals provide a mechanism for retail customers to process their own purchases with minimal or no assistance, as shown in the following figure:

Figure 4. **SCO terminal**

The SCO terminal that is used with this Ready Solution includes the following components:

- **Scanner**—Scans the UPC barcode of the product being purchased
- **Touch-screen display**—Provides the customer with a user interface during checkout. The touch screen displays the scanned item, the price of the item, the total cost of all purchases, the payment option, and other necessary information.
- **Video camera**—Monitors and records the items being scanned
- **Weighting and bagging area**—Includes a weight sensor
- **Alerting mechanism** (for example, an overhead LED or cell phone-based notification)—Alerts retail associates to any issues during self-checkout

**Note**: SCOs are not included as part of this Ready Solution. The customer must have SCOs installed at the retail store.

## Cameras

This Ready Solution is agnostic to the camera model. SCOs are typically equipped with fixed-dome cameras. Cameras support H.264 or H.265 video compression and the video stream is decompressed/decoded in the T4 GPUs for further processing and inference. Dell Technologies recommends cameras that require Power over Ethernet (PoE) for ease of consistent deployment across all retail branches. Dell EMC PowerSwitch N3000 series or S3100 series switches can be used to power the camera and provide network connectivity.

If SCOs across various retail branches have similar camera positioning, scanner positioning, lighting, and environmental factors, it is easy to train and deploy the initial model rapidly across all the retail stores, without needing further customization.

At Dell laboratories, we tested this solution with an AXIS P1275 network camera with HDTV 1080p. The camera comes with an adjustable field of view that is enabled by rotating and tilting the camera sensor to the desired viewing direction.

**Note**: Cameras are not included as part of this Ready Solution. The customer must have cameras installed at the retail store.

## Microsoft Azure IoT Hub

Malong RetailAI Protect can use Azure IoT Edge capability to deploy and manage the deep learning model for the solution, as shown in the following figure:

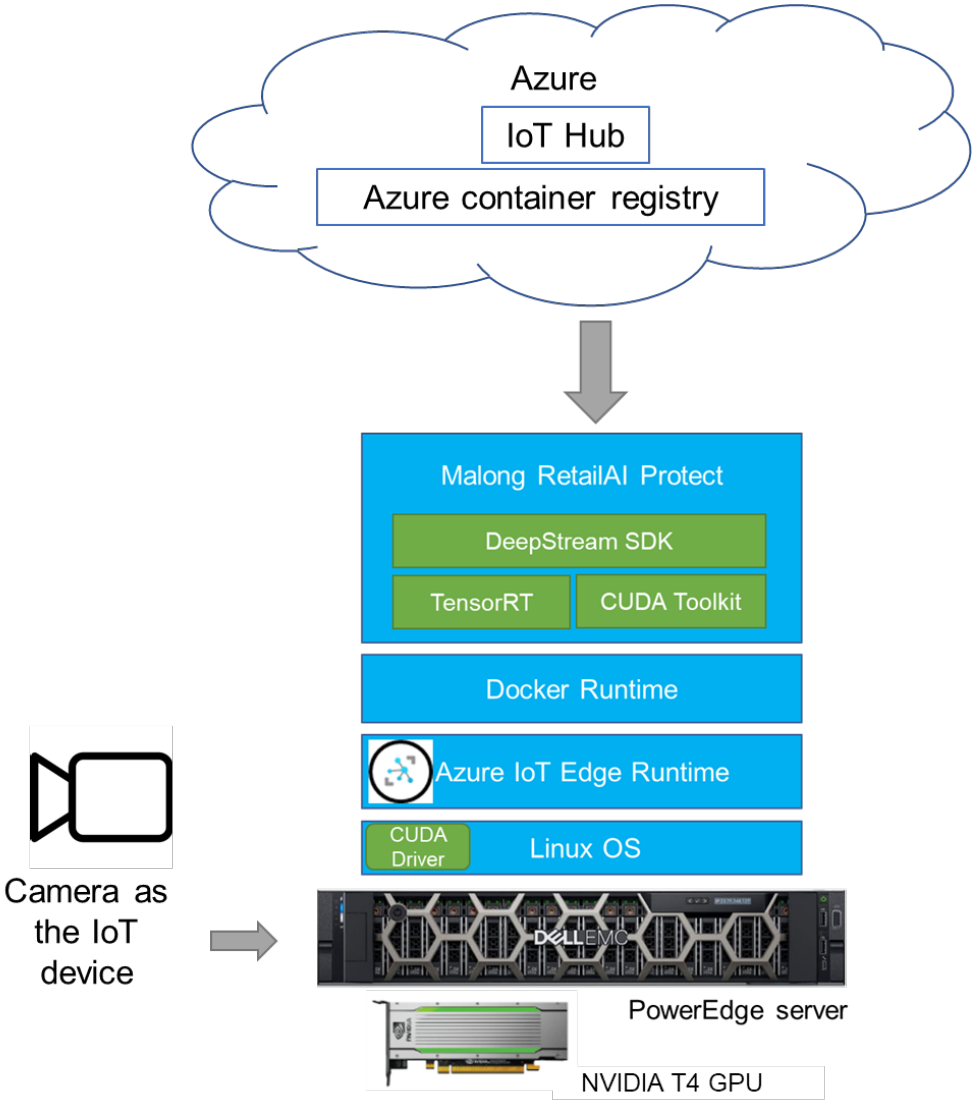


Figure 5. **Managing the solution with Azure**

Azure components include:

- **IoT Hub**—Used to provision, update, and manage IoT devices from the Azure cloud at scale. IoT Hub uses a security-enhanced channel to communicate from the cloud to the registered IoT devices.
- **IoT Edge runtime**—A collection of programs that turn a device into an IoT Edge device. Collectively, the IoT Edge runtime components enable IoT Edge devices to receive code or they enable a model to run at the edge and communicate the results. In this solution, the IoT Edge runtime is deployed in the PowerEdge server. IoT Edge runtime consists of two components (which are not shown in Figure 6).
  - **IoT Edge hub**—Acts as a local proxy for IoT Hub by exposing the same protocol endpoints as IoT Hub. This consistency means that clients (whether devices or modules) can connect to the IoT Edge runtime similarly to connecting to IoT Hub.
  - **IoT Edge agent**—Responsible for instantiating modules, ensuring that modules continue to run, and reporting the status of the modules back to IoT Hub. This configuration data is written as a property of the IoT Edge agent module twin.
- **Container Registry**—Used to store and distribute container images. When container images are registered, they can be deployed across the globe. Container Registry supports mechanisms for version control and for scanning and updating container images.

The Malong AI model is packaged as a container and is hosted on IoT Hub services on Azure using Container Registry services. IoT Edge Runtime is deployed on the PowerEdge server, which also includes Docker Runtime configured on the server. The server is then configured as an IoT Edge device and discovered by IoT Hub. When the PowerEdge server is configured as an IoT Edge device, the container with the AI model is pushed to the server from Azure IoT Hub. Model monitoring, updates, and management can now be performed from IoT Hub.

# Configuration

The following table recommends three configurations that are based on the number of SCOs in the retail store. Dell Technologies recommends one T4 GPU for every four to five SCOs.

**Table 2. Recommended configurations: summary**

| Recommended scenario | Configuration |
|---|---|
| Small configuration for 4 to 5 SCOs or proofs of concept | R7515 server with a single socket AMD processor and 1 NVIDIA T4 GPU |
| Medium configuration for 8 to 9 SCOs | R740 server with dual socket Intel Scalable Xeon processor and 2 NVIDIA T4 GPUs |
| Large configuration for 12 to 13 SCOs | R7525 server with dual socket AMD EPYC processor and 3 NVIDIA T4 GPUs |

The following table provides additional details for each configuration. The camera and customer traffic at the SCOs determine the requirement for resources.

**Table 3. Recommended configurations: details**

| Component | Small | Medium | Large |
|---|---|---|---|
| Server model | R7515 | R740 | R7525 |
| Processor | AMD EPYC 7502P 2.5 GHz, 32C/64T, 128 M Cache (180W) | Intel Xeon Gold 6230 2.1 G, 20C/40T, 10.4 GT/s, 27.5 M Cache, Turbo, HT (125W) DDR4-2933 | AMD EPYC 7702P 2.00 GHz, 64C/128T, 256 M Cache (200W) DDR4-3200 |
| Memory | 8 x 8 GB | 12 x 8 GB | 16 x 16 GB |
| GPUs | 1 NVIDIA T4 GPU | 2 NVIDIA T4 GPUs | 3 NVIDIA T4 GPUs |
| Network adapter | Broadcom 57416 Dual Port 10 GbE BaseT Network LOM Mezz Card | Broadcom 57416 Dual Port 10 GbE BaseT Network LOM Mezz Card | Broadcom 57416 Dual Port 10 GbE BaseT Network LOM Mezz Card |
| Storage | 6 x 2 TB SAS SSDs in RAID 6 | 6 x 2 TB SAS SSDs in RAID 6 | 12 x 2 TB SAS SSDs in RAID 6 |
| Operating system | Ubuntu server | Ubuntu server | Ubuntu server |

## Network layout for the configuration

The following figure shows a sample network layout for the configurations. PowerSwitch N3000 series switches support PoE, which is typically required by the camera. Each SCO requires two network connections, one for the camera and one for the SCO. The PowerEdge server with Malong RetailAI Protect and NVIDIA T4 GPU is also connected to the same network switch, as shown in the following figure:

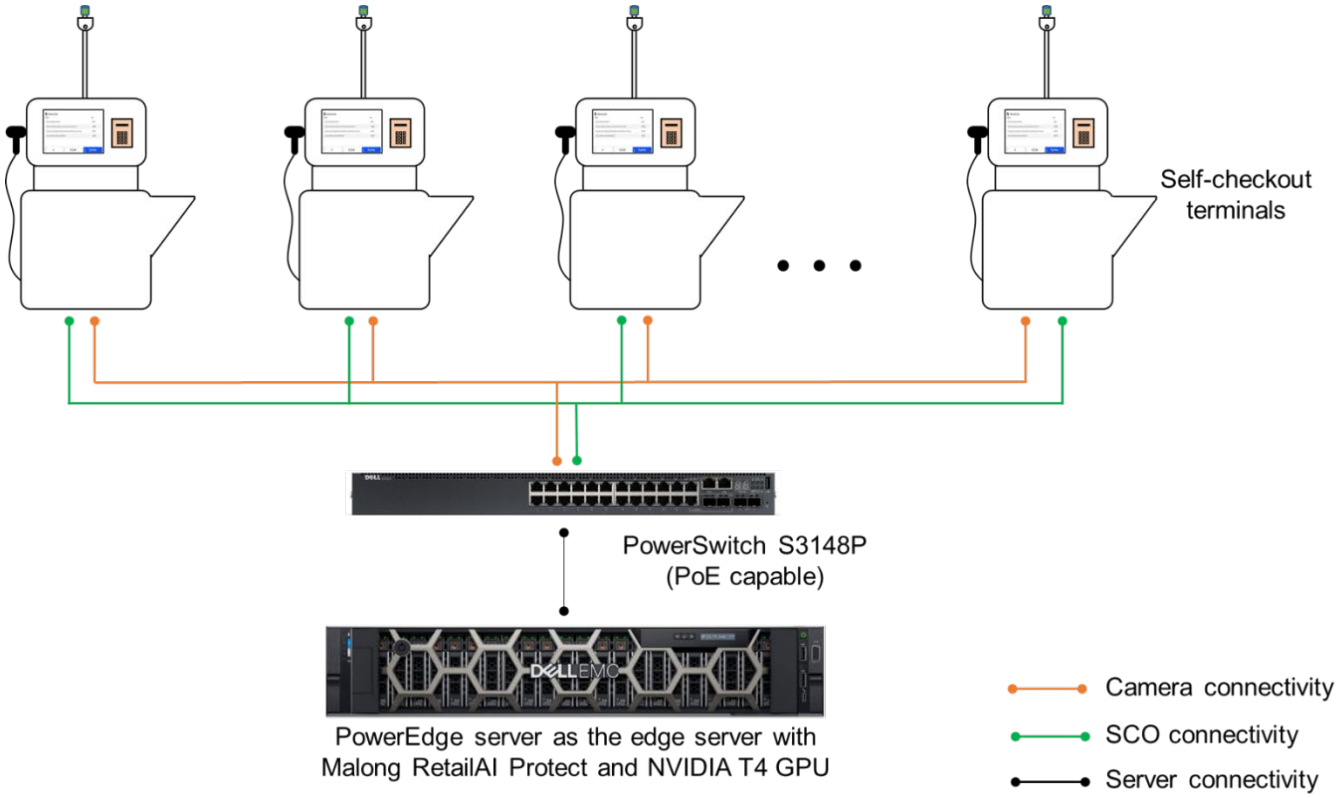


Figure 6. **Network layout**

The following table provides a list of the PowerSwitch switch models and corresponding capabilities that are relevant to this Ready Solution:

**Table 4. PowerSwitch model capabilities**

| Switch model | Capability |
|---|---|
| N3048ET-ON | • 1 GbE port attributes:<br>▪ 12 RJ45 auto-sensing (1Gb/100 Mb/10 Mb) PoE 60W fixed ports<br>▪ 36 RJ45 auto-sensing (1 Gb/100 Mb/10 Mb)<br>• 2 integrated 10 GbE SFP+ dedicated ports<br>• 2 integrated GbE SFP combo ports |
| N3024EP-ON | • 1 GbE port attributes:<br>▪ 12 RJ45 auto-sensing (1 Gb/100 Mb/10 Mb) PoE 60W fixed ports<br>▪ 12 RJ45 auto-sensing (1 Gb/100 Mb/10 Mb) PoE+ fixed ports<br>• 2 integrated 10GbE SFP+ dedicated ports<br>• 2 Integrated GbE SFP combo ports<br>• 60 watts on 12 ports; 30.8 watts on remaining 12 ports (might require a second power supply module) |
| N3048EP-ON | • 1 GbE port attributes:<br>▪ 12 RJ45 autosensing (1Gb/100 Mb/10 Mb) PoE 60W fixed ports<br>▪ 36 RJ45 auto-sensing (1 Gb/100 Mb/10 Mb) PoE+ fixed ports<br>• 2 integrated 10 GbE SFP+ dedicated ports<br>• 2 integrated GbE SFP combo ports<br>• 60 watts on 12 ports; 30.8 watts on remaining 12 ports (might require a second power supply module) |
| S3148P | • 1 GbE Port Attributes:<br>▪ 48x RJ45 auto-sensing (1Gb/100Mb/10Mb) PoE+ fixed ports<br>• Integrated 10GbE SFP+ dedicated ports: 2<br>• Integrated GbE SFP combo ports: 2<br>• Maximum PoE Watts per port: 30.8 watts on 48 ports (may require 2nd power supply module) |

# Deployment and management

## Initial deployment

The initial deployment of this Ready Solution includes the following activities:

- **Initial model training**—Malong has a portfolio of pretrained AI models that can be deployed for retail POS inventory loss detection. Various models can be tested to determine if a model performs well when it is deployed to retail stores. Typically, the base model requires additional training to fit the customer's retail inventory and the local environment, including SCO configuration and camera positioning. If the target retail stores have consistent SCOs and camera positions, the model can be deployed to all retail stores after local training. If retailers have more than one model of SCOs and cameras, RetailAI Protect can adjust and perform well in these scenarios.

- **Integration with customer SCO**—The Malong RetailAI Protect system can be integrated with a POS using APIs from the SCO system. SCO systems, which are either commercially available or custom engineered for large retail chains, typically have an API that sends and receives information through a message brokering service like ActiveMQ. When an item is scanned, the SCO system can return information about the transaction, including the UPC barcode, to the Malong RetailAI Protect solution.
- **Deploying across multiple retail branches**—This Ready Solution can be deployed easily across various retail stores in several geographic locations. Each retail store can have one or more PowerEdge servers as the edge node. The PowerEdge servers are configured and registered in Azure IoT Hub. Then, AI models such as Malong RetailAI Protect can be deployed to the IoT devices by using a few clicks. If SCOs across various retail branches have similar camera positioning, scanner positioning, lighting, and environmental factors, it is easy to train and deploy the initial model rapidly across all the retail stores, without needing further customization. The AI algorithm uses the visual image of the item being scanned to identify mis-scans or ticket switching. The image of the item being scanned depends on the position of the camera and the lighting of the store. If different retail stores have different camera positioning (for example, one store has an overhead fixed-dome camera while another has cameras next to the scanner that point up), the algorithm must be customized for the two scenarios.

### Adding products to the inventory

The Malong RetailAI Protect system automatically learns new products as they are scanned through normal use of the SCO system. When new products are introduced to the inventory and as customers scan these products, the algorithm learns about them from the captured images. The AI model learns to associate the product images to the UPC barcode that is scanned. The model learns about new products after a few scans and correctly identifies if there is a ticket switching or mis-scan. The knowledge of the new products is then automatically distributed across all the store branches.

No action is required from the retail store to update the model as the inventory of the store is updated.

### Monitoring and managing the IoT Edge server

IoT Hub is used to monitor and manage the AI model that runs on the IoT Edge server. The model is deployed as a container to the PowerEdge server edge device. From the Azure IoT portal, IT administrators can monitor all the edge devices across all retail branches. Updates to the models can be deployed to all the retail branches with a few clicks from the Azure IoT portal.

## Solution outcome

Solution overview discussed two loss scenarios in which Malong RetailAI Protect can reduce retail loss: mis-scans and ticket switching. A review of the key components of the solution enables us to better understand how the components of the solution work together to detect these scenarios. A high definition camera monitors retail items as they are scanned. The Malong AI model uses this video stream and the UPC barcode for the scanned item to ensure that there is no loss. The AI algorithm tries to predict the UPC barcode that is based on the video feed of the scanned item to verify the transaction. If the scanned UPC barcode does not match the predicted UPC barcode that is associated

with the visual information from the video feed, the transaction is marked as “needs attention.”

The following scenarios help us understand how this process works:

**Solution in action with a valid scan**—When items are scanned, the Malong AI model uses the visual information to predict the correct UPC barcode and compares it with the scanned UPC barcode. When the scanned UPC barcode matches the predicted UPC barcode value that is based on the visual image in the camera, the transaction proceeds without interruption. The customer can then scan the next item or complete the purchase, as shown in the following figure:

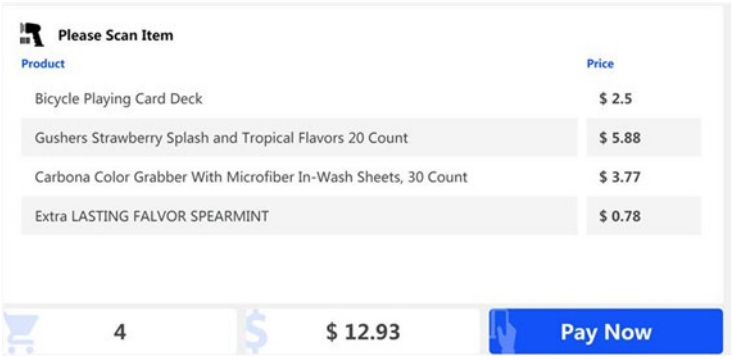


Figure 7. **Valid scan**

**Ticket switching and mis-scan**—Ticket-switching occurs when a customer substitutes a product's barcode with a barcode from a less expensive item. The Malong AI model detects that the predicted UPC barcode is not consistent with the item’s visual information and the scanned barcode. The Malong AI model immediately notifies the alerting system, as shown in the following figure:

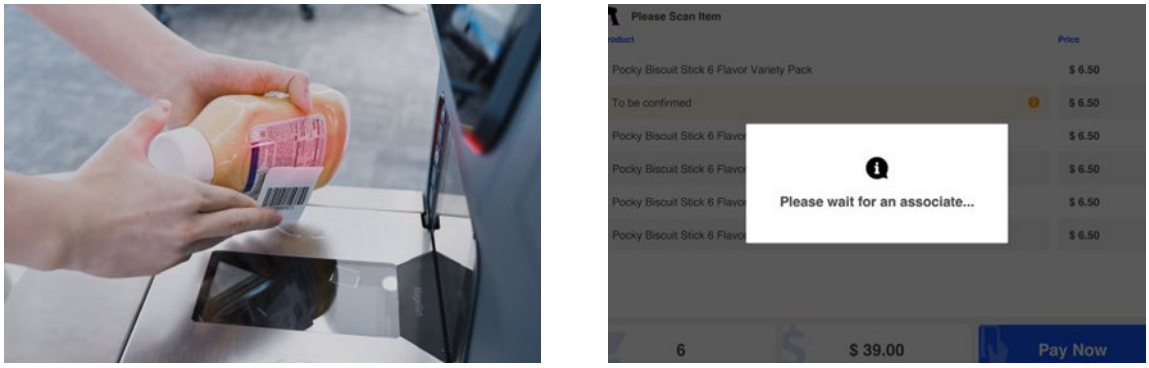


Figure 8. **Ticket-switching scan**

The transaction is paused. Typically, a retail associate is notified to assist the customer to rescan the item correctly. If configured, a recorded video clip of the transaction event is available to the retail associate.

A mis-scan occurs when the scanned item does not register with the SCO scanner because the customer avoided scanning the UPC barcode intentionally or accidently. Accidental mis-scans occur when a part of the hand obscures the barcode or the customer scans the wrong side of the package. In this scenario, the Malong AI model

detects that there is no scanned code to match the visual information of an item and waits for the scan to register. If the model detects that the transaction is being completed or another item is being scanned, the Malong AI model notifies the alerting system and the transaction is paused.

# Conclusion

Most inventory loss in the retail sector occurs in traditional stores at the POS. The increased use of security personnel in stores involves a large cost with limited value. More focus on training is required for retail loss prevention. New AI-based automated systems using Dell EMC hardware and special purpose deep learning models from Malong provide cost-effective tools to reduce inventory loss. NVIDIA's development of lower-cost GPUs for use with pretrained models in retail applications that require high scalability with low latency evaluation also help address inventory loss. The combination of Dell EMC hardware with NVIDIA GPUs and Malong software provide a complete solution that is easy to deploy and maintain. The solution can satisfy the demanding needs of retail loss prevention by improving margins while remaining largely transparent to the customer's shopping experience.