WHITE PAPER

AI-Driven Loss Prevention for Retail

Combating Shrinkage at Self-Checkout Systems with Rapidly Deployable and Scalable Solutions

Executive Summary

Retail shrinkage, particularly from self-checkout systems (SCO), is a rapidly growing challenge, with losses in the U.S. alone estimated to be over $100 billion annually[1]. Nearly 7% of all self-checkout transactions contain some form of loss—accidental or intentional—compared to just 0.3% of cashier-led transactions. As self-checkout continues to expand, projected to reach 80% of supermarket transactions[2], the financial impact of shrinkage demands urgent attention.

AI-driven loss prevention presents a transformative solution. By leveraging real-time computer vision, retailers can detect scanning errors and suspicious behaviors as they occur, enabling immediate corrective action and significantly reducing losses. Research indicates most self-checkout errors are unintentional and can be resolved on the spot, converting potential losses into completed sales. Retailers who leverage AI to reduce shrinkage have seen positive ROI in as little as 3 months.

Deploying AI-powered loss prevention requires specialized hardware and software. Supermicro has designed a reference model for loss prevention deployments, combining robust, enterprise-grade edge systems with NVIDIA's library of development tools and pre-trained models. Together with a network of implementation partners, this ecosystem delivers actionable solutions to enable retailers to protect margins, streamline operations, and enhance customer trust.

[1] Source: NRF
[2] Source: ECR Retail Loss

TABLE OF CONTENTS

Highlights

- Retail shrinkage is estimated at $100B annually in the US
- AI Vision based systems enable retailers to detect suspicious behavior in real time
- Blueprints and development tools are available to facilitate the implementation of loss prevention solutions

Retail Loss Prevention Reference Model

To facilitate the implementation of loss prevention in retail organizations, Supermicro has created an AI-based Loss Prevention reference model. This model functions as a blueprint that can be further developed and fine-tuned to meet the requirements of specific deployments.

In this model, AI-powered loss prevention receives input from several devices to identify suspicious behavior:

1. Images captured by security video cameras, cameras in the self-checkout station, and scans by a barcode scanner.
2. The information from these sources is processed simultaneously to detect and recognize retail items.
3. Based on video frames throughout the checkout process, a prediction on the item's barcode is made.
4. If the data from the barcode scanner does not match this prediction, the application labels the item as unmatched.
5. The model then generates an alert to alarm employees about the event in real-time, and a manager can observe events on a dashboard.

When an item is detected with a low confidence level or a match is not found, it is logged for future training, and a warning is sent to the employee to help the customer out. Additionally, all predictions should be logged to be audited when necessary.

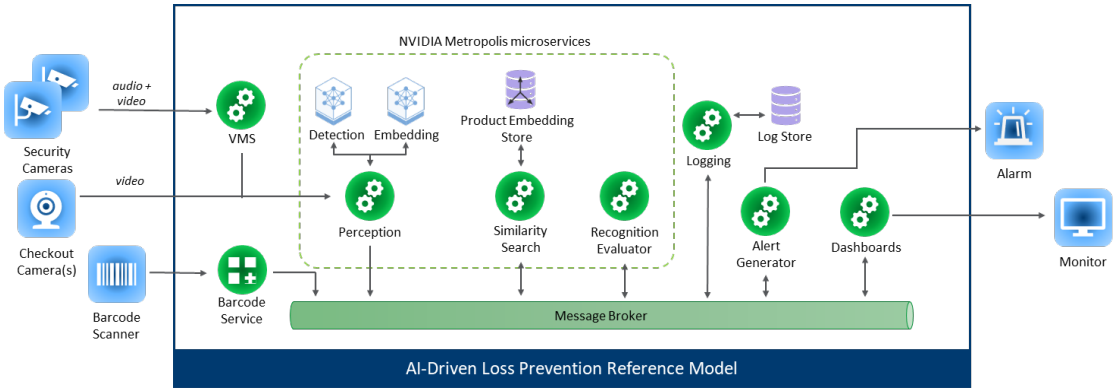

Figure 1: Supermicro Loss Prevention Reference Model

To enable this functionality, the loss prevention model taps into the cloud-native microservices available in the NVIDIA Metropolis framework. This application framework offers an extensive library of predefined AI models and datasets, along with other tools to jump-start development. These assets enable retailers and solution providers alike to speed up the design and implementation of AI-driven applications.

NVIDIA's database contains hundreds of thousands of products that can be found in supermarkets and convenience stores, enabling rapid indexing of products and avoiding the need for a lengthy, costly training process. The Retail Object Detection and Retail Object Recognition models offered by NVIDIA are pre-trained to identify products prone to theft and identify them in varying shapes and sizes. Additionally, they are based on a state-of-the-art, few-shot learning technique developed by NVIDIA Research that, combined with active learning, identifies and captures any new products scanned by customers and sales associates during checkout to improve model accuracy further.

Building on the AI workflow, the Loss Prevention model also references functional modules to integrate with existing systems, for example, barcode scanners, and other features such as logging and alerts. These are essential to evolve the AI model into a mature solution that can be deployed in retail organizations, and where the involvement of a trusted advisor or system integrator is invaluable.

Model Optimization and Management

Before being deployed in stores, the AI model used for loss prevention needs to be fine-tuned. This action can take place in the organization's central data center. Here, the cloud-based microservices are customized for a retailer's unique needs, ensuring that all stores operate with the same product data and use the same parameters. This fine-tuning process can be performed through the NVIDIA TAO toolkit.

If there are variations in in-store conditions such as lighting, the model needs to be trained accordingly with additional data. This is where synthetically generating more data based on the different store environments and products can reduce the data collection and labeling cost. Upon completion, the pre-trained model is distributed to all locations in the network.

As the model is operational, data is collected and returned to the central data center. This feedback includes information on unrecognized products, log files, and aggregated data about instances where suspicious behavior was detected. The data collected this way is then used to further improve the model. By creating a continuous improvement cycle, retail organizations ensure optimal ROI from the retail loss solution.

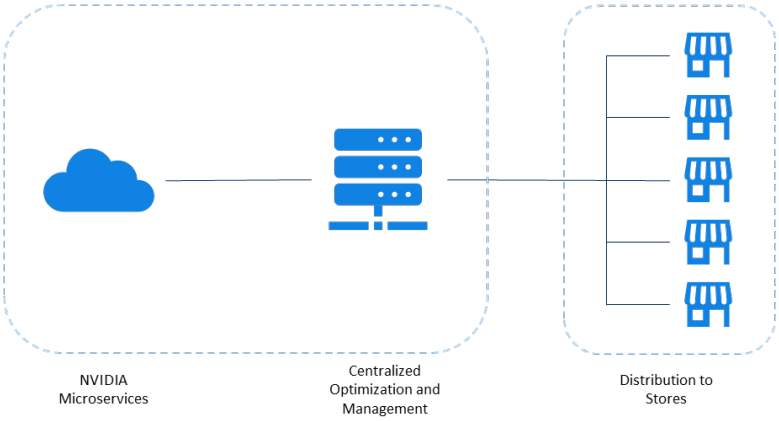

Figure 2: AI Model Management and Distribution

Server Infrastructure

Successfully using AI to combat retail loss requires data from cameras and barcode scanners to be processed in real-time, with as little delay as possible. The best results will be achieved by running the AI model on-premises in the retail store. The alternative – sending the data off-site to a central data center for processing – will inevitably lead to longer response times, lower accuracy (due to video compression), or alerts being sent when it's already too late.

Inferring the data on-premises, at the edge of the network, has several additional benefits:

- No dependency on external connectivity, which can be costly and prone to downtime
- Fewer security and privacy concerns as camera images are not transmitted live over the internet
- Better scalability as the number of SCO lanes or cameras changes
- Faster rollout to new locations

The technical requirements of a system to run AI-based loss prevention at the edge depend on a range of factors – the number of cameras and SCO lanes, the quality and encoding of the video streams, and additional software and AI-based applications that a store is using, to name a few. We recommend working with a trusted implementation partner, such as a system integrator or software developer specializing in AI applications, to determine the exact details of your unique environment. Here are some aspects to consider when selecting the right system model for your deployment.

- Hardware-based decoding capacity for multiple video and audio formats for edge AI inferencing
- Network capacity to support the control plane, the data plane for camera streams, out-of-band (OOB) management, and in-band management connections
- Sufficient CPU cores and memory to run Video Management Systems (VMS), Perception, Logging, Dashboard and Alerts workloads
- Storage capacity for short-term video storage and long-term event logging
- Able to run reliably without requiring the conditioned environments of a data center

In the data center, AI training servers are deployed to fine-tune and update the loss prevention, product detection, and product recognition models. Processing the incredible amounts of data involved in AI training requires specialized hardware. To fine-tune an AI Vision model, the NVIDIA L40S GPU can be used, although a powerful accelerator might be required if multiple AI workloads are processed on the same infrastructure.

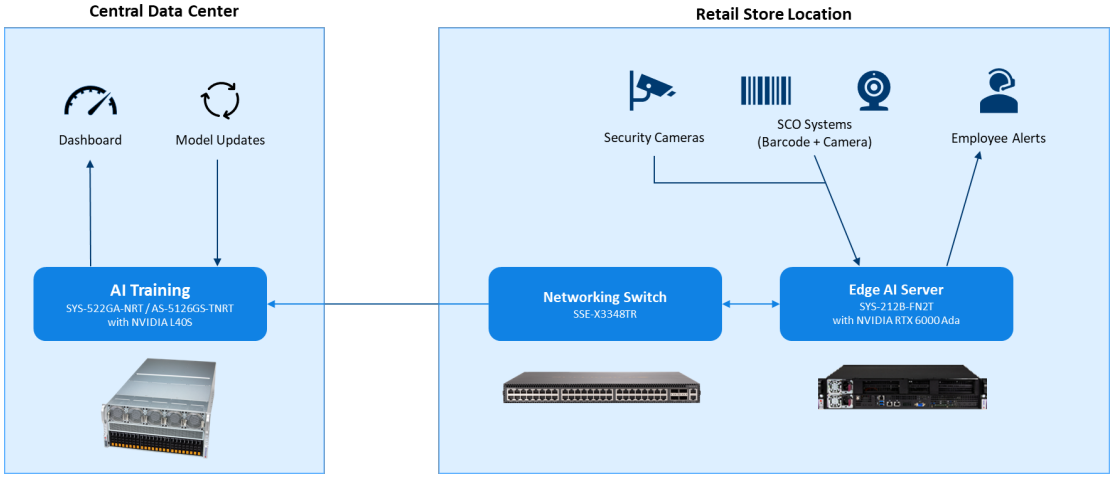

Figure 3: Loss Prevention Server Infrastructure

Featured Systems

Below are some example system configurations that can be used to run loss prevention in retail stores. Consult an expert from Supermicro or one of our ecosystem partners to find out which server is the best match for your requirements.

AS -1115S-FWTRT

The AS -1115S-FWTRT is a compact, 1U system designed to support intelligent workloads at the edge. With a chassis depth of only 429mm (16.9”) the AS -1115S-FWTRT can be installed in space-constrained environments. Combined with an NVIDIA L4 Tensor Core GPU, this server is well-equipped to run loss prevention workloads in small-scale retail locations.

- AMD EPYC™ 8004 series processor
- Room for up to 1 single-width GPU accelerator
- Front I/O interface for easier physical management in constrained spaces
- Operating temperature 0°C to 40°C (32°F to 104°F)
- 2 internal 2.5” storage bays and 2 M.2 NVMe slots

Learn more about the AS -1115S-FWTRT

SYS-E403-14B-FRN2T

Designed for tight spaces, the box-sized SYS-E403-14B combines a high-performance CPU with support for multiple AI accelerator cards in a compact, wall-mountable form factor. These features make it an ideal system for AI workloads in retail stores.

- Single Intel® Xeon® 6700/6500 series processors with P-cores or 6700 series processors with E-cores up to 300W with air cooling
- Up to 1 double-width or 2 single-width GPUs
- Operating temperature 0°C to 45°C (32°F to 113°F)
- Up to 2 front hot-swap 2.5" NVMe + 2 internal fixed 2.5" SATA drive bays

Learn more about the SYS-E403-14B-FRN2T

SYS-212B-FN2T

Supermicro's X14 SYS-212B-FN2T is a 2U, short-depth system ideally suited to run edge AI workloads such as loss prevention for retail. Featuring an Intel® Xeon® 6 series processor, this system delivers data center tier compute performance to a compact, versatile system for remote deployments. It can be equipped with up to two double-width GPU accelerators such as the NVIDIA L40S or RTX 6000 Ada, which means it can easily be scaled to meet the AI requirements of a specific deployment environment. The SYS-212B-FN2T can operate in environments from 0°C to 40°C (32°F to 104°F) and does not require the specialized cooling of a data center.

- Intel® Xeon® 6700/6500 series processors with P-cores or 6700 series processors with E-cores
- Versatile PCIe configuration for up to 2 double-width GPU accelerators
- Front I/O interface for easier physical management in constrained spaces
- Operating temperature 0°C to 40°C (32°F to 104°F)
- 4 hot-swappable 2.5" storage bays and 2 M.2 NVMe slots

Learn more about the SYS-212B-FN2T

SYS-222HE-FTN

For high-end AI inferencing workloads, the 2U Hyper-E is a high-density server with increased performance that can be deployed in micro data centers at the edge. Featuring two Intel® Xeon® 6 processors and up to three double-width GPU accelerator cards, including support for the H100, L40S, and RTX 6000 Ada models, the SYS-222HE-FTN performance complex AI workloads, including large language models (LLM).

- Dual Intel® Xeon® 6700/6500 series processors with P-cores or 6700 series processors with E-cores
- Versatile PCIe configuration for up to 3 double-width GPU accelerators
- Front I/O interface for easier physical management in constrained spaces
- Up to 10 hot-swappable 2.5" storage bays and 2 M.2 NVMe slots

Learn more about the SYS-222HE-FTN

SYS-322GA-NR

Designed to run the most demanding AI workloads at the edge data center, the 3U SYS-322GA-NR offers unparalleled power and flexibility with up to 10 PCIe 5.0 x16 FHFL or 20 PCIe 5.0 x8 FHFL slots. These enable the server to be configured with up to 8 double-width GPU accelerator cards, including support for 8 H100 NVL, L40S, or 4 H200 NVL, RTX Pro 6000 Blackwell Server Edition models, or a combination of GPU and other expansions cards such as video capture cards. The SYS-322GA-NR can run a broad range of different workloads simultaneously, or function as a regional hub between multiple store locations and the central data center.

- Dual Intel® Xeon® 6900-Series Processors with P-cores
- Up to 10 PCIe 5.0 x16 FHFL or 20 PCIe 5.0 x8 FHFL slots
- Up to 8 double-width or 19 single-width GPUs
- Up to 14 E.1 NVMe drive bays or up to 6 2.5” NVMe drive bays

Learn more about the SYS-322GA-NR

SYS-522GA-NRT // AS -5126GS-TNRT

For model training in the data center, Supermicro’s 5U PCIe GPU server delivers a powerful and versatile platform capable of housing up to 10 double-width GPUs. This model provides tremendous acceleration and flexibility, supporting both Intel and AMD CPU families and a broad range of GPU cards, including the NVIDIA H200 NVL, RTX Pro 6000 Blackwell Server Edition, RTX 6000 Ada Generation, or L40S models. When configured with the H200 NVL, the system can be further expanded with an NVIDIA® NVLink® Bridge for faster GPU-to-GPU connectivity.

- Dual Intel® Xeon® 6900 series processors with P-cores (SYS-522GA-NRT) or dual AMD EPYC™ 9005 Series Processors (AS -5126GS-TNRT)
- Up to 10 (SYS-522GA-NRT) or 8 (AS -5126GS-TNRT) double-width PCIe GPU accelerator cards
- Up to 6TB of DDR5 memory
- Additional PCIe slots for DPU/NIC high speed networking
- 6x 2700W Redundant Titanium Level (96%) power supplies

Learn more about the SYS-522GA-NRT

Learn more about the AS -5126GS-TNRT

Solution Partners

Supermicro works with an ecosystem of partners, including software vendors, solution integrators, and consultants, to support the development and implementation of Loss Prevention and other AI-based applications for retail organizations of any kind.

Centific is a Seattle-based tech company specializing in platform-driven solutions that transform raw data into fuel for the future of AI. By elevating data quality and accelerating AI deployments, Centific empowers businesses to build smarter, safer, and more scalable AI solutions.

For more information:

https://centific.com/

RocketBoots transforms video into operational gains for leading global retailers and banks. Store operations, labor planning and loss prevention teams use RocketBoots computer vision AI to reduce shrink at self-checkouts and managed checkouts without adding unneeded friction, while continuously improving front-end productivity in a way that does not risk sales and customer loyalty.

For more information:

https://www.rocketboots.com/

Conclusion

Retail shrinkage is not just a security concern; it's a $100 billion financial opportunity waiting to be reclaimed. With the growing reliance on self-checkout, deploying AI-driven loss prevention becomes essential. Supermicro's reference model, built on enterprise-grade edge systems and powered by NVIDIA's extensive AI toolkit, empowers retailers to detect and prevent losses in real time, turning potential shrinkage into completed sales to deliver ROI in as little as three months.

With ready-to-use development tools, pre-trained models, and integration-ready infrastructure, businesses can accelerate deployment and scale with confidence. Combined with an ecosystem of expert partners, Supermicro makes AI-based loss prevention both accessible and impactful. Retailers now have the tools to protect margins, improve operations, and regain control—at the checkout and across the enterprise.

For more information, visit www.supermicro.com/LossPrevention.

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.