WHITE PAPER

ENABLING INTELLIGENT STORES WITH EDGE AI

Table of Content

The Rise of the Intelligent Store....1

Supermicro and NVIDIA: Accelerating Intelligent Retail Adoption ...2

Challenges in Bringing Edge AI to the Store Floor....2

Intelligence at the Edge: Solving Retail Constraints 2

Intelligent Store Use Cases 3

- Loss Prevention 3
- Store Analytics....4
- In-Store Service Optimization....5
- Edge Management....6
- Making Edge AI solutions real with Supermicro and NVIDIA....7

Supermicro and NVIDIA AI Infrastructure Portfolio....7

- Loss Prevention for Grocery Stores (SYS-111AD-WRN2) 9
- Store Analytics for Gas Stations (SYS-E300-14AR)....9
- In-Store Optimization for Hypermarts (SYS-212B-FN2T) 9

Implementing Edge AI in Retail: Your Path to Smarter Stores....10

Supermicro and NVIDIA Services and Software Support 11

Conclusion: Building the Store of the Future 12

The Rise of the Intelligent Store

Retail is undergoing a profound transformation—one that goes beyond the shift to online marketplaces. Today, retailers are reimagining the physical store as a digitally intelligent environment, powered by AI deployed at the edge. This evolution enables real-time responsiveness to customer behavior, inventory changes, and operational needs, unlocking new levels of personalization and efficiency.

Unlike the centralized infrastructure of e-commerce, Edge AI deployment brings compute power directly into stores. This allows retailers to act instantly on insights from video feeds, sensors, and sales systems, eliminating latency and enabling smarter, faster decisions.

These in-store insights contribute to an omnichannel experience, where customers no longer choose between online and offline channels. Instead, they engage with a blended ecosystem in which physical stores offer digital-like personalization, interactivity, and convenience.

AI also drives operational intelligence, transforming how stores function day-to-day. From loss prevention and store analytics to employee support, AI automates routine tasks, enhances service delivery, and provides actionable insights that improve both customer and staff experiences.

Industry momentum is strong: 51% of surveyed retail organizations are already using AI, with investments spanning physical stores, digital channels, back-office systems, and supply chains. Notably, 94% of adopters surveyed report that AI has helped reduce operational costs by at least 20%, underscoring its tangible business value.[1]

[1] NVIDIA - State of AI in Retail and CPG: 2025 Trends

To realize these benefits, retailers need infrastructure that delivers consistent AI performance across distributed locations. Supermicro's edge AI portfolio, powered by NVIDIA technology, provides the edge AI inferencing power, scalability, and resilience required to run intelligent stores effectively and build a future-proof foundation.

Supermicro and NVIDIA: Accelerating Intelligent Retail Adoption

Supermicro and NVIDIA are transforming retail by delivering edge AI solutions that bring intelligence directly into the store. These systems enable retailers to process real-time data from cameras, sensors, and point-of-sale systems—unlocking capabilities such as personalized customer engagement, predictive inventory management, and loss prevention. Supermicro's modular, high-performance edge servers combined with NVIDIA's AI software stack—including Metropolis microservices and pretrained models—allow retailers to deploy intelligent store features quickly and reliably across thousands of locations, supported by software partners that help manage and scale these applications.

Key Outcomes of the Intelligent Store Transformation

- **Enhanced customer experiences** through personalization and faster service
- **Optimized operations** via smarter merchandising and demand forecasting
- **Improved employee support** through automation and real-time insights

What makes this collaboration especially impactful is its ability to remove traditional barriers to AI adoption. Supermicro's flexible hardware configurations support a wide range of store formats and workloads, while NVIDIA's low-code development tools simplify integration for retailers with limited IT resources. Together, they offer a scalable, resilient foundation that minimizes latency, reduces operational complexity, and accelerates time-to-value—empowering retailers to focus on outcomes like improved customer satisfaction, reduced shrinkage, and enhanced employee support.

Technology Momentum Spotlight: Supermicro delivers infrastructure that can be easily integrated into NVIDIA's Enterprise AI Factory-validated designs. These scalable, edge-ready platforms are purpose-built to support the future of intelligent retail—bringing high-performance AI to physical stores and enabling real-time analytics, personalization, and operational efficiency.

Challenges in Bringing Edge AI to the Store Floor

Retailers are eager to embrace intelligent store models, but scaling from pilot projects to large deployments introduces a series of stubborn challenges. The cost of deploying AI infrastructure across hundreds or even thousands of distributed sites remains a major hurdle, especially when budget constraints are top of mind. While cloud-based models once seemed like a straightforward solution, the high and ongoing cost of running GenAI in the cloud—on top of limitations around latency and bandwidth—has made it clear that processing cannot live solely there. At the same time, integration with existing store systems—often a patchwork of legacy hardware and proprietary platforms—adds another layer of complexity.

Technology itself has advanced to the point where delivering performance to the edge is less of a barrier. Small form factor, versatile systems from providers like Supermicro, combined with AI acceleration from NVIDIA GPUs, give retailers enterprise-grade compute power that was once too costly or complex to deploy in large numbers. In many cases, these solutions resemble the hardware retailers have already been installing, which helps smooth the adoption curve. The challenge is less about whether the systems can perform, and more about having the correct data, integrations, and investment strategy in place. Even if stores aren't ready to deploy fully today, planning ahead for Edge AI positions them to move quickly when the moment is right.

Intelligence at the Edge: Solving Retail Constraints

An AI app running on an associate's in-store system can instantly recognize products on shelves, flag low stock, and provide real-time product information to customers. By processing data locally, the device delivers fast, reliable support without relying on constant cloud connectivity. If it relies on the cloud and fails, customers face delays in getting answers, uncertainty about product availability, and longer wait times for assistance. This can quickly turn what should feel like a smooth, confident shopping experience into a memorably frustrating one.

Avoiding these customer frustrations requires intelligence at the point of action, which means moving compute directly into the store environment. Edge AI is essential because it enables real-time decision-making where latency, bandwidth, and responsiveness are critical. Unlike traditional cloud-only approaches, which can be slowed by network bottlenecks or unreliable connectivity, Edge AI brings processing directly into the store, ensuring decisions—whether for inventory management, customer experience, or loss prevention—happen instantly.

Deploying AI at the edge introduces its own unique requirements. Retail spaces are often constrained in terms of physical footprint, power availability, and cooling, which means AI infrastructure must be compact, versatile, and energy-efficient. By situating compute at the edge, retailers can overcome traditional hurdles: they achieve reliable performance outside the data center, reduce bandwidth costs by processing data locally, and unlock cost efficiencies that support large-scale deployment across hundreds of sites.

Edge AI not only mitigates the practical limitations of physical retail spaces but also accelerates adoption by transforming previously complex deployments into manageable and repeatable implementations. The combination of edge AI inferencing power, scalability, and operational resilience ensures that retailers can act on insights immediately, expand quickly without prohibitive infrastructure investments, and maintain consistent service levels even in distributed, bandwidth-constrained environments.

Intelligent Store Use Cases

While edge intelligence establishes the technical foundation, its transformative power is best seen through how it addresses core retail challenges. Retail transformation is driven by several use cases, with three of the most common and impactful being loss prevention, store analytics, and in-store optimization. These cases are essential for enhancing operational efficiency, improving customer experiences, and driving better business outcomes.

Loss Prevention

Every retailer, no matter the size or format, experiences shrink age. It shows up in many forms—mis-scans at checkout, damaged goods, stockroom errors, or intentional theft—and together these small leaks silently drain billions in revenue each year. While theft often gets the spotlight, it's only one part of a much wider challenge. The real picture is far more complex, spanning accidental mistakes, process gaps, and systemic blind spots across the store and supply chain.

Edge AI gives retailers the visibility and speed needed to tackle shrink from every angle. By analyzing video and sensor data directly at the store level, it detects issues the moment they occur—helping teams separate accidents from theft, act before problems escalate, and reduce revenue loss in real time. Retailers adopting these systems are seeing measurable gains in accuracy, efficiency, and customer satisfaction. Supermicro's scalable edge platforms and NVIDIA's AI acceleration empower retailers to deploy advanced loss prevention solutions that adapt to any store format, enabling instant, on-premises detection and response to shrinkage events without reliance on the cloud. To fully realize this use case, edge systems can be paired with a software solution expert, such as Everseen.

Making it Real with Everseen

Everseen delivers Vision AI that helps protect profits, improve customer experience, and empower associates. Trusted by the world's leading retailers, their technology powers over 22 million transactions every day across 120,000+ checkouts worldwide. Their solutions span the entire retail ecosystem—from front-of-store and self-checkout to aisles, stockrooms, and supply chains. *Learn more about Everseen*

Evercheck: AI for Loss Prevention

Everseen's Evercheck solution uses Vision AI to help reduce shrink, and improve staff productivity and customer experiences. Built on the Everseen Vision AI platform, Evercheck detects, and deters recovery of unwanted behaviors at checkout, helping retailers recover losses and streamline front-of-store operations. The system is scene-aware, analyzing real-time patterns across shoppers, associates, and scanning behavior to identify issues, whether errors or intentional theft. Patterns include mis-scans, non-scans, product-switching, and basket- and cart-based loss. In self-checkout systems, Evercheck provides a simple on-screen prompt that instructs shoppers on how to correct missed scans themselves; when necessary, the system alerts associates to provide assistance. This helps reduce shrink, resolves many issues quickly and discreetly, and preserves a smoother, more positive experience for both customers and staff.

Measurable outcomes

Everseen has a proven track record with retailers, reflecting how Edge AI is transforming retail loss prevention and delivering measurable ROI. Retailers face rising shrink age from theft, errors, and back-of-store risks. With real-time video and sensor analytics at the edge, issues are detected and addressed immediately, helping create safer environments for staff and customers.

> *"Evercheck has proven to be a game-changer for loss prevention, recovering more value than we thought possible. Not only are we minimizing shrink, it's helping us boost staff productivity and streamline overall operations."*
>
> General Manager, EMEA grocery retailer

A Forrester Consulting Total Economic Impact™ Study commissioned by Everseen (Sept 2024) reflects the measurable outcomes of Evercheck. The study showed that over three years, a composite organization representative of interviewed customers with experience using Evercheck achieved a 374% ROI, with payback in under six months. Based on in-depth interviews and financial analysis, the composite organization realized more than $37M in shrink reduction, recovering $88,000 per store annually. In addition, the study indicated that retailers realized a 15% improvement to staff productivity and additional cost savings from retiring outdated systems.[1]

> *"Evercheck has transformed our approach to loss prevention— helping us address not only malicious activity but also honest customer mistakes. It also helps us strike the right balance between store efficiency and customer experience."*
>
> VP, U.S. National Grocer

1. The Total Economic Impact™ Of Everseen's Evercheck Solution, Forrester, 2024

Store Analytics

Every retailer wants to understand what's happening in their stores—product availability, promotional compliance and safety—but gaining a complete, real-time picture has historically been difficult. Sales and inventory data only tell part of the story, creating blind spots in the front of store that can lead to missed sales, suboptimal layouts and frustrated shoppers.

Edge AI changes this. By processing video data on-site, retailers can track real-time shelf conditions and promotional execution in order to trigger actions like restocking shelves or ordering product. Real-time awareness of item availability can update online channels to reduce costly substitutions or optimize delivery routes. And busy store associates appreciate the assistance in keeping track of endless details. Together, this creates a smarter, more responsive store environment that aligns products, people, and operations with actual demand, helping retailers maximize sales, reduce operating costs and optimize the customer experience. With Supermicro's modular hardware and NVIDIA's real-time AI processing, retailers can rapidly implement store analytics that deliver actionable insights to associates on the floor, improving inventory accuracy and promotional execution across all locations. To fully realize this use case, edge systems can be paired with a software solution expert, such as Augmodo.

Making it Real with Augmodo

AI-driven solutions help retailers transform their physical stores into intelligent, digitally aware environments. Augmodo specializes in providing a Spatial AI platform that delivers a real-time view of every product in every store. Using wearable Smartbadge™ technology, associates passively capture live data on product availability while completing their normal tasks. This lightweight device with dual 4K lens scales quickly with an affordable price point, easy install and no operational change for the store [Learn more about Augmodo]

Augmodo's solution for Intelligent Store Analytics

An intelligent store leverages Augmodo's Spatial AI to convert live video and sensor data into actionable insights. Throughout the day, retailers can track on-shelf availability, monitor high velocity product zones, and understand planogram and promotional compliance. The system helps optimize staff workflows and associate time with "next best tasks" while reducing tedious audits. With fast deployment and low-capex passive hardware, stores can scale quickly, get immediate insights, and make operational adjustments in real time.

> *"Our hardworking associates spend hours every day completing tedious but critical tasks, like auditing and restocking shelves so consumers can have increased product availability,"*
>
> Mark Finocchiaro, Managing Partner and General Manager IT Retail at Chemist Warehouse.

Measurable outcomes

Retailers using Augmodo have achieved measurable improvements in store performance. Within a month, some clients reported up to a 30% improvement in on-shelf availability, faster restocking, and optimized associate allocation. Edge AI enables immediate responses to changing conditions, improving customer experiences, increasing conversion rates, and enhancing operational efficiency, creating smarter, more adaptable store environments.

> *"With Augmodo's spatial AI assistant, they wear the Smartbadge on a lanyard or clip, and everything is automatically scanned as they walk the floor, providing us with real-time inventory data. The assistant tells them what's out of stock and what needs to be done next, giving them more time with shoppers."*
>
> Mark Finocchiaro, Managing Partner and General Manager IT Retail at Chemist Warehouse.

In-Store Service Optimization

Operational inefficiencies in stores—long queues, inconsistent product quality, under- or over-staffed areas, overlooked cleaning or restocking, and challenges in evaluating final product quality—can frustrate customers and erode margins. For retailers with fresh food operations, including delis, grocery counters, convenience store cases, and in-store kitchens, menu innovation and advertising can create additional headaches as maintaining consistency and quality across multiple locations becomes difficult. Every misstep adds up, impacting service speed, product freshness, and overall shopper satisfaction.

Edge AI addresses these challenges by turning real-time video and sensor data into actionable insights that empower employees rather than replace them. The system monitors queues, guides staffing, evaluates food preparation and final product quality, and alerts staff when cleaning, restocking, or other attention is needed. Productivity is the key: automation is a helpful tool, but the real value comes from enabling the human behind the system to work more efficiently and effectively. By acting on these insights immediately, stores and fresh-food operations reduce waste, maintain high standards, and deliver smoother, more satisfying experiences—aligning people, processes, and operations with real customer demand. By leveraging Supermicro's

edge-ready systems and NVIDIA's AI software stack, retailers gain the ability to monitor and optimize in-store operations dynamically, supporting managers and staff with data-driven recommendations that enhance service quality and operational consistency. To fully realize this use case, edge systems can be paired with a software solution expert, such as Kwali.

Making it Real with Kwali

Kwali provides an AI-powered platform that helps quick-service restaurants and franchise operators optimize in-store operations and food production. Proven in the field with widespread retail deployments, Kwali leverages cameras, sensors, and edge computing to give managers real-time visibility into operations while empowering employees to be more productive and focused on delivering high-quality service. Marketing data suggests that operators using Kwali can experience benefits such as up to 200% improvement in Taste metrics and Customer Overall Satisfaction, 15% lower food costs, and 10% faster food production, illustrating the potential for significant operational impact across a broad portfolio of locations. Learn more about Kwali

Kwali's Solution for In-Store Optimization

Kwali transforms real-time video and sensor data into actionable insights and reports that directly address operational pain points. The system monitors food production according to franchise standards, evaluates final product quality, tracks shifts performance, and highlights areas where training or attention is needed. Managers can view all stores from a responsive dashboard, while store-specific reporting provides guidance on where and how to improve. Recognition features, such as performance-based leaderboards, help identify top talent and boost employee morale, addressing high turnover in the industry. By giving employees time back to focus on producing quality food and improving customer experiences, Kwali enhances both operational efficiency and workforce engagement.

> *"That was the biggest issue, people not saucing correctly, and we fixed that in every store; the product got drastically better."*
>
> Director of Operations, Restaurant Franchise

Measurable outcomes

Operators using Kwali's platform report measurable improvements in quality and operational performance. Within two months of pilot programs, overall product quality increased by 88%, and 75% of stores achieved a monthly quality average above 70%. Real-world deployments also improved operational consistency and employee focus, enabling teams to act immediately on production issues and maintain high standards. Collectively, these improvements demonstrate how Kwali can give teams more time to focus on quality, customer satisfaction, and employee recognition while aligning operations with real-world demand.

> *"The store no one expected was #1, it made me view the stores differently and lends credibility to a store flying under the radar."*
>
> Director of Operations, Restaurant Franchise

Edge Management

As retailers continue to expand and operate across distributed networks of stores, restaurants, and branches, the need for robust, real-time digital infrastructure continues to grow. Edge management solutions play a key role in enabling centralized deployment, monitoring, and control of applications and services across thousands of geographically dispersed locations. By positioning compute resources closer to where data is generated and used, these solutions reduce latency, maintain continuous operations, and accelerate ongoing digital transformation. Edge management platforms provide a consistent foundation for

both legacy and next-generation applications, supporting virtualization, containerization, and IoT device integration. Centralized management tools allow IT teams to orchestrate updates, monitor system health, and enforce security policies from a single interface—whether in the cloud, on-premises, or at third-party data centers. This approach ensures that critical applications remain available and secure, even during connectivity disruptions. Supermicro and NVIDIA provide unified edge management tools that simplify IT operations, allowing centralized orchestration, automated updates, and robust security for AI-powered applications deployed at scale across distributed retail environments.

Making it Real with Edge AI Management

Retailers leveraging edge management solutions can rapidly deploy new digital experiences—such as AI-powered analytics, dynamic signage, and interactive content—without the delays or risks associated with centralized cloud-only models. Edge management ensures that applications and updates are delivered reliably to every location, while IT teams maintain full visibility and control. By consolidating operations and automating routine management tasks, retailers reduce operational overhead and minimize downtime. The result is a more resilient, scalable, and responsive store environment, capable of supporting innovation at the pace of changing consumer expectations.

Measurable Outcomes

Retailers adopting edge management have achieved measurable improvements in operational efficiency and customer experience. Centralized orchestration and monitoring enable immediate response to issues, while robust security and compliance features protect sensitive data. Over time, this approach empowers retailers to scale digital initiatives, adapt quickly to new business needs, and deliver consistent, high-quality experiences across all locations.

Making Edge AI solutions real with Supermicro and NVIDIA

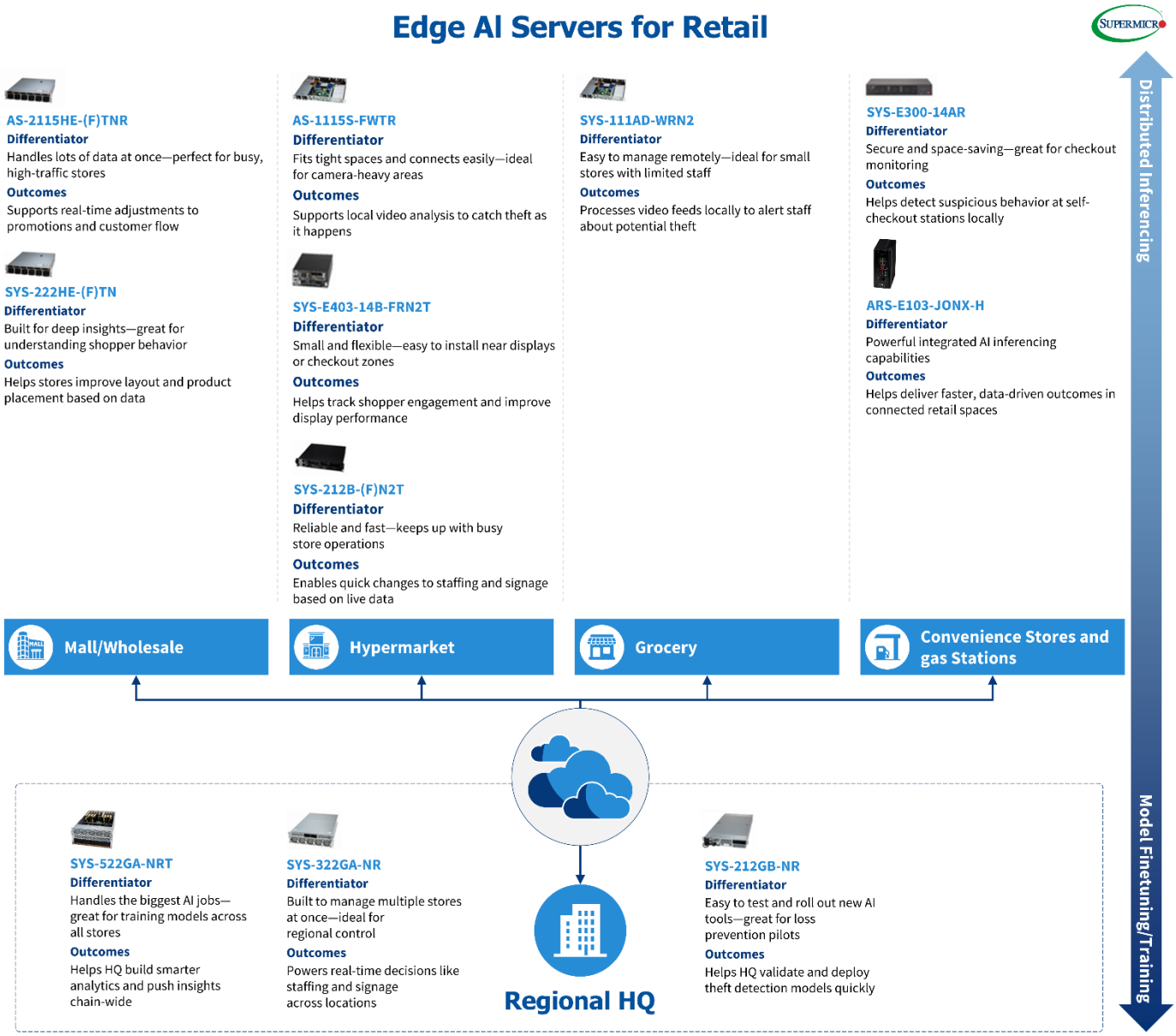
With these use cases—Loss Prevention, Store Analytics, In-Store Optimization, and Edge Management—clearly illustrating the transformative potential of Edge AI, the next question becomes: how can retailers deploy these capabilities effectively at scale? Delivering real-time insights, maintaining performance, and supporting operational continuity across hundreds or thousands of stores requires hardware that is both powerful and practical. This is where purpose-built Edge AI infrastructure comes into play, providing the compute, efficiency, and reliability needed to turn these use case opportunities into fully operational, high-impact solutions.

Supermicro and NVIDIA AI Infrastructure Portfolio

The Supermicro and NVIDIA AI Infrastructure portfolio represents a significant leap forward in the deployment of intelligent retail solutions. By combining Supermicro's robust, modular hardware with NVIDIA's cutting-edge GPUs, such as the NVIDIA RTX PRO™ 6000 Blackwell Server Edition, this portfolio delivers unparalleled performance and scalability for edge AI applications. Just as importantly, it is designed with the practical realities of retail environments in mind—where space is often limited, power efficiency is critical, and cooling needs must be carefully managed. These considerations make the portfolio uniquely capable of running advanced AI workloads directly at the edge without overwhelming store infrastructure.

This section will delve into the specific components and capabilities of the Supermicro and NVIDIA Edge AI Portfolio, highlighting how it not only meets the performance demands of AI-driven technologies but also addresses the operational constraints that retailers face every day.

The diagram below maps out the range of Supermicro systems, illustrating how their differentiated designs align with the unique requirements of various retail environments—from compact convenience stores to large-format retailers.



GPU	Why It Fits
NVIDIA RTX PRO™ 6000 Blackwell Server Edition	High-performance universal GPU, capable of multi-instance AI support for multi-camera analytics and concurrent workloads. Passive-cooled, dual-slot server GPU, built for 24/7 reliability in dense retail deployments.
NVIDIA RTX PRO™ 6000 Blackwell Max-Q Workstation Edition	Powerful workstation GPU for in-store servers, kiosks, or interactive experiences that need real-time inference. Active-cooled, full-length double-slot workstation card, delivering maximum edge performance.

NVIDIA RTX PRO™ 4000 Blackwell SFF Edition	Compact, efficient solution for smart signage, smaller analytics setups, or localized service monitoring. Active-cooled, single-slot workstation card, ideal for space- and power-limited environments.
NVIDIA L40S	Data center–class GPU for centralized retail hubs or private cloud, powering large-scale AI inference across many stores. Passive-cooled, dual-slot server GPU, optimized for reliability in dense deployments.
NVIDIA L4 Tensor Core GPU	Compact, power-efficient GPU for dense video inference, multi-camera monitoring, and fraud detection. Passive-cooled, low-profile single-slot server GPU, perfect for constrained retail edge servers.
NVIDIA Jetson Orin™	Embedded AI platform that drives advanced computer vision and generative AI in connected retail environments.

Below are some sample retail AI scenarios, each highlighting a different environment. We'll explore a system from each environment and dive into how it supports the outcomes outlined in the diagram above, illustrating the tangible benefits each Supermicro system delivers for these use cases.

Loss Prevention for Grocery Stores (SYS-111AD-WRN2)

Scenario: A grocery store installs the SYS-111AD-WRN2 in its front-end area to support AI-powered loss prevention at self-checkout stations and high-value product zones. The store aims to detect suspicious behavior in real-time and reduce shrinkage without relying on cloud processing.

How it helps: The SYS-111AD-WRN2 is a 1U short-depth edge server powered by 12th-14th Gen Intel Core processors, supporting up to 128 GB DDR5 memory. It includes dual 2.5GbE LAN ports and supports either one double-width or one single-width NVIDIA GPU (e.g. NVIDIA RTX PRO™ 4000 Blackwell, NVIDIA L4 Tensor Core GPU) via PCIe slots for AI acceleration. It connects to in-store cameras and runs AI models locally to detect behavior like item concealment, barcode switching, or loitering near high-theft areas. In addition to grocery stores, its remote management via IPMI, compact design, and front-I/O make it well-suited for gas station s and other space -constrained retail environments.

Outcome: Local AI alerts staff instantly, reducing shrinkage while protecting sensitive data.

Store Analytics for Gas Stations (SYS-E300-14AR)

Scenario: A gas station installs the SYS-E300-14AR system to analyze customer behavior and interactions with in-store displays, checkout areas, and fuel lanes. The goal is to understand traffic flow, dwell time at product zones, and engagement with promotions to optimize layout and merchandising.

How it helps: The SYS-E300-14AR is a compact, fan-based embedded system powered by an Intel® Core™ Ultra Series 2 processor with up to 96GB DDR5 memory. It features multiple LAN ports of up to 10 GbE and a single PCIe 5.0 slot, capable of supporting a GPU such as the NVIDIA L4 Tensor Core GPU for AI acceleration This enables the system to connect to in-store cameras and run AI models that analyze customer movement, dwell time, and interactions with products and displays. Its small form factor and front I/O make it easy to deploy in tight retail spaces, and its embedded design ensures quiet, reliable operation.

Outcome: Edge AI tracks customer engagement, guiding layout and merchandising for better results.

In-Store Optimization for Hypermarts (SYS-212B-FN2T)

Scenario: A Hypermart installs the SYS-212B-FN2T in its back office to support real-time decisions on customer flow, staffing, and promotional engagement. The store experiences variable traffic patterns and wants to dynamically adjust operations based on live conditions.

How it helps: The SYS-212B-FN2T is a 2U compact front I/O edge AI system equipped with a single Intel® Xeon® 6700/6500 series processor, up to 2 double-width GPUs such as the NVIDIA L40S or NVIDIA RTX PRO™ 6000 Blackwell Max-Q Workstation Edition, and up to 2TB DDR5 ECC memory. It connects to in-store cameras and sensors and runs AI models that monitor queue lengths, foot traffic, and shopper behavior. PCIe 5.0 slots and redundant AC/DC power ensure continuous, high-performance inferencing in demanding retail environments.

Outcome: Edge AI informs real-time adjustments, optimizing flow, staffing, and promotions.

These examples highlight how Edge AI can transform retail operations—reducing shrinkage, optimizing traffic, and providing real-time shopper insights. Next, we'll explore how to bring these possibilities to life through a structured implementation process that creates smarter, more responsive stores.

Implementing Edge AI in Retail: Your Path to Smarter Stores

Bringing Edge AI into retail isn't just about adding new technology—it's about shaping stores that are more responsive and efficient. While every retailer's journey will look a little different, following a structured process like the one below sets the stage for lasting impact.

1	2	3	4	5	6
Identify problems and quantify opportunity	**Identify possible solutions**	**Define KPIs and methods to track rollout success and growth**	**Validate through pilot in real-world settings**	**Phased production rollout**	**Roll out to production in phases**
Scenario: *Are customers leaving empty-handed because products are out of stock? Could checkout lines be shortened through smarter traffic flow insights?*	**Scenario:** *What if computer vision could detect when shelves need restocking before employees or customers notice? Could AI-driven recommendations help create more personalized in-store promotions?*	**Scenario:** *If AI predicts product demand with great accuracy in a single store, how much waste or lost revenue could that prevent chain-wide?*	**Scenario:** *What happens to customer satisfaction scores when AI helps staff focus on high-value interactions instead of manual tasks?*	**Scenario:** *If self-correcting inventory systems are running across every store, how much staff time could be redirected toward customer service?*	**Scenario:** *As customer behaviors shift, how can your AI models adapt to detect new shopping patterns or seasonal changes automatically?*

1 Start by exploring what's possible with Edge AI for your retail locations. Identify the challenges you want to solve—like optimizing inventory, enhancing in-store experiences, or improving operational efficiency—and define the business outcomes you want to achieve. At this stage, consider which ISVs and services can best support your objectives. While enterprise retailers may have internal capabilities to manage implementation, SMBs often benefit from external expertise to accelerate adoption and reduce risk.

2 Work with the Supermicro and NVIDIA team and any necessary technology partners, ISVs, and service providers to brainstorm AI-driven solutions tailored to your needs. Evaluate different AI models, algorithms, and technologies that could help solve your challenges. Market research, retail-specific insights, and partner expertise help ensure approaches are viable and scalable, whether for SMBs with a few locations or enterprise chains with hundreds.

3 Test a specific use case in a controlled environment to demonstrate impact. This could be in a store or lab setting, providing tangible evidence of value before broader rollout. Integration with existing in-store systems is key, supported by ISVs and service partners for implementation, data integration, and analytics—especially valuable for SMBs—while larger enterprises may leverage internal teams for validation at scale.

4. Deploy the solution in a small number of stores to evaluate real-world performance. Adjustments can be made based on actual operations, and the pilot helps confirm alignment with your business needs. Guidance from ISVs and service providers ensures SMBs have the expertise needed to scale effectively, while enterprise teams may focus on cross-store consistency and broader operational integration.

5. Once validated, implement the Edge AI solution across your retail locations. Collaborate with your team and partner ecosystem to train staff, define KPIs, and help ensure smooth adoption. ISVs and services continue to provide integration support, optimization, and troubleshooting, enabling SMBs to deploy confidently, while enterprises may rely on internal resources for large-scale orchestration.

6. With the infrastructure fully in place, Edge AI solutions can continue evolving as your business grows. Regular updates, hardware upgrades, and new store deployments maintain peak performance, while ISVs and services help implement improvements, monitor results, and leverage the latest AI capabilities—supporting SMBs and enterprises alike in sustaining long-term value and agility.

Supermicro and NVIDIA Services and Software Support

Edge AI is transforming retail by enabling real-time insights, smarter operations, and responsive customer experiences across the retail space. Supermicro offers flexible services and support to keep edge hardware running efficiently, while NVIDIA AI Enterprise is tailored for larger retail organizations, providing cloud-native tools and frameworks that simplify AI model deployment, optimize performance, and maintain security at scale. These capabilities help retailers manage Edge AI effectively, maximizing operational efficiency and accelerating AI-driven outcomes.

Supermicro Global Services and Support provides flexible and customizable Service Level Agreements for logistics, remote service desk, and onsite support to cover the Supermicro hardware solutions. The Onsite Service Programs offer a 4-hour Response time option for mission-critical environments or any tailored solution that will meet your specific business requirements.

For more info visit Supermicro's Onsite Global Services site.

NVIDIA AI Enterprise is a cloud-native suite of software tools, libraries, and frameworks designed to deliver optimized performance, robust security, and stability for production AI deployments. Easy-to-use microservices enhance model performance with enterprise-grade security, support, and stability, ensuring a smooth transition from prototype to production for enterprises that run their businesses on AI. With broad adoption across industries and an extensive partner ecosystem, NVIDIA AI Enterprise enables organizations to build and deploy agentic AI systems anywhere—across clouds, data centers, or at the edge. By streamlining development and optimizing hardware utilization, NVIDIA AI Enterprise helps businesses accelerate time to market and reduce infrastructure costs while ensuring reliable, secure, and scalable AI operations.

For more info visit the NVIDIA AI Enterprise site.

Conclusion: Building the Store of the Future

Edge AI is redefining the role of the physical store—transforming it into an intelligent, responsive environment that combines operational efficiency with personalized customer experiences. From fresh food counters to convenience stores and hypermarkets, retailers now have the opportunity to align people, processes, and technology in ways that were not possible before.

Supermicro and NVIDIA are at the center of this transformation. Together, they provide the edge AI infrastructure and software capabilities that help retailers overcome the complexity of adoption while delivering measurable outcomes: reduced shrink and operational costs, empowered employees, and better service for customers. By making AI both scalable and accessible, this partnership equips retailers to unlock new levels of productivity and build future-ready operations.

The path to intelligent retail is clear. Now is the time to reimagine your store experience and accelerate AI adoption at the edge. Engage with Supermicro to start building the store of the future today.

Visit Supermicro's AI Retail solutions site to learn more.

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. See www.supermicro.com.

NVIDIA

NVIDIA accelerated computing platforms power the new era of computing, performing exponentially more work in less time with much lower energy consumption than traditional CPU-based computing. Accelerated computing revolutionizes energy efficiency across industries by harnessing NVIDIA GPUs, CPUs, and networking, all optimized through NVIDIA enterprise software solutions.
More information at https://nvidianews.nvidia.com.