WHITE PAPER

ACCELERATING ADOPTION OF EDGE AI

Supermicro Servers with the NVIDIA AI Platform Deliver Best-In-Class Outcomes for Predictive and Generative AI Implementations at the Edge

TABLE OF CONTENTS

Introduction

Artificial Intelligence (AI) has become a global conversation and a business imperative across industries—both for its implications on how people interact in digital and physical spaces as well as the impact it stands to make on how people live and work in the future. **This is warranted, as Goldman Sachs estimates AI will add as much as $4.4 trillion in annual value to the global economy[1].**

Many industries were already heavily adopting the technology when ChatGPT was released in 2022, ushering in a new era of generative AI that has unlocked new revenue streams and experiences. Medical providers are using AI to improve diagnosis accuracy, improve outcomes for patients and save lives. Sports and event venues are debuting fully dynamic fan experiences, allowing visitors to interact in new ways with their favorite players and teams.

Ultimately, AI and its applications—both generative and predictive—across industries are among the most rapidly evolving technologies impacting our lives today. **According to a 2023 MIT Technology Review Insights survey, 78% of executives agree that scaling AI/ML use cases to create business value is a top priority[2].**

Alongside this trend, 'the edge' has emerged as a critical concept, particularly in industries where real-time data processing and privacy are paramount and where customers commonly interface with organizations at locations such as retail stores, restaurants, hospitals, or stadiums. In these locations, businesses are installing unified hardware & software solutions that enable new services to increase revenue, add differentiation, and make headlines.

The edge refers to computational processes occurring close to the data sources at these locations – sensors, devices, or end-users – rather than relying on centralized systems common in traditional AI processing. Edge processing unlocks hyper-personalized experiences like dynamic menu recommendations in quick-service restaurants, while also contributing to significant savings in data transfer bandwidth.

This shift to the edge is changing how industries ranging from retail and healthcare to manufacturing and smart city management engage with their customers. The immediacy of data processing allows for quicker responsiveness, enabling organizations across these industries to offer new and dynamic, AI-enabled customer experiences that would not otherwise be possible.

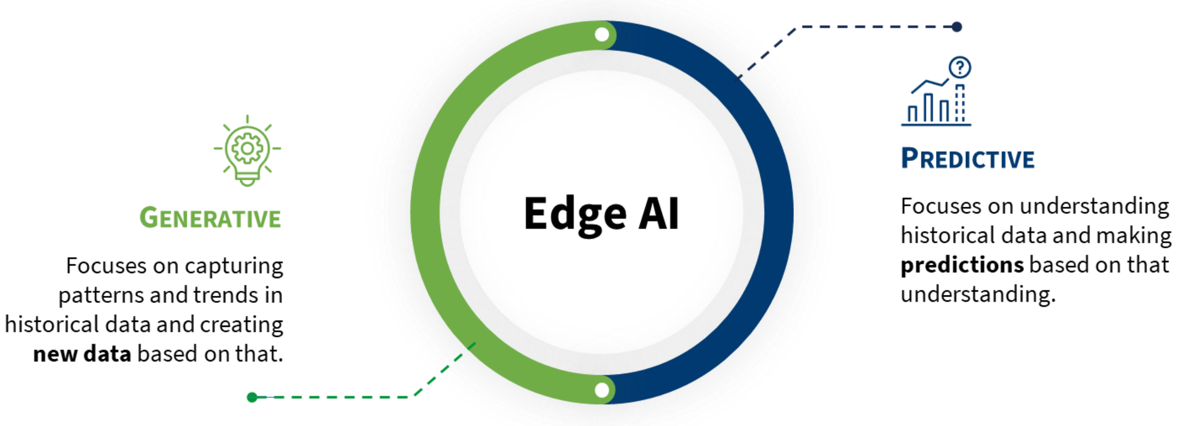
[1,2] MIT Technology Review Insights. (2023). *The Great Acceleration: CIO Perspectives on Generative AI*. Cambridge, MA: Massachusetts Institute of Technology.

The Two Halves of Edge AI

The deployment of AI at the edge can be for two distinct yet complementary purposes: predictive and generative AI. Predictive AI is the analytical powerhouse, leveraging historical data to forecast future events, spot unexpected anomalies, detect patterns in images and videos, and more.

As an example, in retail settings, Edge AI can be used to predict and manage inventory levels and facilitate automated loss prevention. Data can be combined from in-store video feeds and RFID-enabled shelves and processed in real-time for autonomous shopping experiences, providing better customer experiences while simultaneously protecting assets. **Shoplifting – referred to as inventory shrinkage – amounted to a record $112 billion in 2022, up nearly 20% from the year prior[3].** Predictive AI at the edge plays a critical role in addressing this problem by enabling informed decision-making at scale.

Generative AI, conversely, is a creative catalyst that uses underlying data patterns to synthesize new information, driving innovation and personalized experience. For example, generative AI can be harnessed by automated kiosks within hospitals to provide customized route wayfinding for hospital visitors. Unlike a simplistic, traditional information kiosk – or a human receptionist who cannot leave their post – dynamically generated wayfinding can make use of opt-in facial recognition or contact information to guide visitors through hallways to their destination or provide turn-by-turn directions sent directly to a mobile device. **With hospital operating costs up 11% since pre-pandemic times, generative AI at the edge can help lessen the burden on overworked staff while improving experiences for individuals[4].**



Focuses on capturing patterns and trends in historical data and creating **new data** based on that.

Focuses on understanding historical data and making **predictions** based on that understanding.

Figure 1 – Generative vs Predictive Edge AI

The Impact on Industry Functions and Customer Engagement

The integration of AI at the edge is revolutionizing how industries operate and interact with customers. By harnessing AI at the point of data generation, businesses are unlocking new capabilities. In manufacturing, for example, companies are using AI on the factory floor to conduct automated quality inspections alongside human engineers. Automated quality inspection leads to superior defect detection, better resilience from workforce attrition among inspectors, increased inspection rates, and increased yields at greater throughput. **Despite only 11% of surveyed manufacturers saying AI is now a critical part of the function, that number is projected to grow to 38% by 2025[5].**

[3] Reuter, Dominick. "Retailers Lost $112 Billion to Inventory Shrink in 2022." Business Insider, September 26, 2023. https://www.businessinsider.com/retailers-lost-112-billion-to-inventory-shrink-in-2022.

[4] American Hospital Association. "2022 Hospital Expenses Increase Report." April 2022. 2022-Hospital-Expenses-Increase-Report-Final-Final.pdf (aha.org).

[5] MIT Technology Review Insights. (2023). *The Great Acceleration: CIO Perspectives on Generative AI*. Cambridge, MA: Massachusetts Institute of Technology.

From a customer-facing lens, retail stores and quick-service restaurants can customize window signage and advertising on the fly based on a customer's attention outside their storefront. Highlighted deals or items can be perfectly tailored to the shopper, increasing conversion before entering the store.

The approach of implementing AI at the edge not only enhances operational efficiency but also opens avenues for innovative services and products tailored to the dynamic needs of end-users.

Why Edge AI?

A critical factor in Edge AI adoption is the ability to reduce latency by avoiding transmission of data across large distances, which is essential in scenarios where even a fraction of a second's delay can have significant consequences. Additionally, processing data at the edge reduces the load on central data networks, mitigating bandwidth constraints. This architecture is particularly crucial in industries where vast amounts of data are generated continuously, such as in video surveillance or sensor networks – **the amount of data generated annually by IoT devices is expected to grow over 400% by 2025 to 73 ZB (zettabytes)[6]**. By handling AI functions in the locations where those systems are serving customers, the adoption of Edge AI can lessen the bandwidth required to carry out this work while also decreasing the time it takes to process requests.

In modern industrial applications of Edge AI, the interplay of predictive and generative AI is crafting a new narrative. While predictive AI continues to play a crucial role in forecasting and decision-making, generative AI is opening new doors for innovation and creativity. **Across use cases like employee training, automated quality inspection, and predictive maintenance, manufacturing companies surveyed by MIT expect AI implementations to be either widespread or business critical within 68% of business operations by 2025[7].** A balanced approach between predictive and generative solutions is vital for businesses looking to leverage Edge AI for both the optimization of current operations and the exploration of new opportunities. Supermicro's role in this domain is critical, providing tailored solutions that support the diverse requirements of both AI types, thereby enabling businesses to stay ahead in the competitive landscape.

Applications of Edge AI are Everywhere

Edge AI is rich with diverse applications, each tailored to harness the distinct capabilities of predictive and generative AI. For example, in retail, predictive AI plays a pivotal role in loss prevention, using data analytics to anticipate and prevent potential losses. On the other hand, quick-service restaurants are innovating with generative AI, creating more natural and intuitive customer service experiences through advanced speech recognition and natural language processing, enabling them to move beyond the rigid command/response systems of the past. These examples and those provided below are just a sample of a broad and deep set of possible use cases for Edge AI across retail, food service, manufacturing, healthcare, and smart spaces. Although each implementation varies in the usage of predictive, generative, or both types of AI, the shared thread is that all of them have proven ROI in production. Businesses already find that Edge AI can reduce costs, increase revenue, and improve workforce or customer safety.

Use cases for Edge AI cater to a variety of personas, each with unique requirements and objectives. Business leaders seek AI solutions that drive growth and operational efficiency. Operators focus on reliability and the seamless integration of AI into existing processes. Developers, meanwhile, look for robust platforms and tools that enable the creation of innovative AI applications. Understanding these diverse needs is crucial in tailoring Edge AI solutions to meet these challenges.

While there are myriad ways to leverage AI at the edge, certain applications stand out for their proven impact. Let's delve into specific industry verticals and some of the most effective Edge AI implementations being done today:

[6] Jovanovic, Bojan. "Internet of Things Statistics for 2023 – Taking Things Apart." DataProt, May 5, 2023. https://dataprot.net/statistics/iot-statistics/.

[7] MIT Technology Review Insights. (2023). *The Great Acceleration: CIO Perspectives on Generative AI*. Cambridge, MA: Massachusetts Institute of Technology.

Retail: Asset Protection

In the retail sector, predictive AI is revolutionizing asset protection**. At a cost of over $100 billion per year globally, inventory shrinkage is one of the most significant problems today for physical retailers[8].** By analyzing patterns and identifying potential threats, AI systems provide retailers with preemptive solutions to reduce losses and enhance store security. This is done by preventing label switching, mis-scans, and even shoplifting. In-store security teams can be alerted to monitor high risk areas before a crime occurs so that loss prevention can be ready to respond. This predictive approach is crucial in a sector where more and more stores are altering layouts, closing off entrances, or even shuttering stores entirely. These calculations must happen at the edge to facilitate the real-time responses that are required to respond effectively. **As shrinkage is projected to grow at nearly 20% annually on average, responses like predictive Edge AI will become table stakes in staying competitive[9].**

Quick Service Restaurants: Speech-Enabled Food Kiosks

Quick service restaurants are embracing generative AI to transform their customer experience. Speech-enabled food kiosks, powered by generative AI, offer personalized interactions, understanding, and responding to customer preferences in a more natural and engaging manner. Imagine solutions that go far beyond the typical automated kiosks that many of us have experienced, where a simple menu is shown, and the customer is guided through a stepwise checkout. Instead, generative Edge AI can offer upsells and complimentary suggestions on the fly or offer human-like customer service right within the kiosk. When combined with predictive AI that learns from a customer's loyalty card history, each subsequent visit can provide deeper customizations and more delightful interactions. This blend of generative and predictive AI, made possible by processing at the edge, marks a significant evolution from traditional customer service models. It can maximize revenue while creating an improved customer experience. **It is no surprise that Business Insider found that 60% of restaurants that implemented AI technology say it's critical to improving their processes[10].**

Manufacturing: Workforce Health and Safety

The manufacturing industry is leveraging predictive AI to address one of its most important and expensive challenges: workplace injuries. **In 2022, associated costs from these injuries – many of which are avoidable – reached $167 billion in the U.S. alone[11].** AI can help create a safer work environment by predicting potential hazards and ensuring compliance with safety protocols. Companies are using Edge AI hardware to connect data feeds from CCTV, machinery, and operational software to provide real-time predictive analysis of hazards as they emerge. As mentioned previously, real-time processing of such large amounts of data is only practical using Edge AI. Given that most manufacturing facilities already employ video monitoring systems, this boost in safety is a matter of utilizing that data most effectively.

Manufacturing: Automated Quality Inspection

Another significant application in manufacturing is automated quality inspection, a predictive AI-driven process. By continuously monitoring a production line using Edge AI, predictive systems can provide real-time insights for quality control and significantly reduce the risk of defects. For example, in a traditional inspection system, 1 in 20 parts might be inspected for defects. However, by using computer vision and Edge AI, manufacturers can inspect every part coming off the production line. **It is, therefore, no surprise that leveraging AI in manufacturing can reduce machine downtime by 30% to 50% and that**

[8] Reuter, Dominick. "Retailers Lost $112 Billion to Inventory Shrink in 2022." Business Insider, September 26, 2023. https://www.businessinsider.com/retailers-lost-112-billion-to-inventory-shrink-in-2022.

[9] Reuter, Dominick. "Retailers Lost $112 Billion to Inventory Shrink in 2022." Business Insider, September 26, 2023. https://www.businessinsider.com/retailers-lost-112-billion-to-inventory-shrink-in-2022.

[10] Graves, Allen. "AI for Restaurants: Accelerate Marketing and Improve Guest Experience." Bloom Intelligence, February 23, 2023. https://bloomintelligence.com/blog/artificial-intelligence-restaurant-industry/.

[11] Muscad, Ossian. "The True Cost of Workplace Safety: A Complete Guide." DataMyte, October 23, 2023. https://datamyte.com/blog/cost-of-safety/.

quality-related costs can be reduced by 10 – 20%[12]. Another benefit is mitigating the impact of skilled employee attrition, as it requires less manual intervention and training. The combination of Edge AI with skilled labor in manufacturing complements human oversight, ensuring higher accuracy and efficiency in quality control, thereby reducing costs.

Healthcare & Medical Devices: Diagnostic Imaging

In healthcare, predictive AI is increasingly being trialed with diagnostic imaging, attempting to enhance the accuracy and efficiency of medical diagnoses to augment trained experts. **Of the more than 3.6 billion imaging procedures conducted annually, approximately 97% of that data goes unused[13]**. Herein lies the potential of leveraging AI to process this vast amount of imaging data, aiding clinicians in identifying patterns that might be missed by the human eye. Edge AI would allow individual providers to make use of predictive technology directly in care settings to support their life-saving work. Although this application involves heavy regulatory scrutiny due to its nature, **49% of healthcare executives surveyed this year said they thought it was at least somewhat likely that AI would be utilized this way by 2028[14].**

Healthcare: Wayfinding

In healthcare, particularly in hospital settings, generative AI is being employed for wayfinding - an application that enhances the patient and visitor experience. Digital kiosks and concierge services, powered by AI, guide individuals through complex hospital layouts, offering directions and information on wall-mounted screens that follow the visitor's progress or updates delivered to a mobile device. The timing is crucial because healthcare staffing costs are increasing, leading to heavier burdens on hospitals: **a variety of studies have shown that administrative staffing expenses total nearly $1 trillion per year. Therefore, finding ways to improve patient and visitor experiences – like wayfinding – without adding to staffing overhead costs is an important win-win[15].** This application is not only limited to healthcare but is also relevant in large spaces with high traffic—like airports or sports stadiums—where efficient guidance systems significantly improve the visitor experience. Not only can routes be planned to a specific location in a building, but generative AI can suggest stops along the way, such as the best coffee shop or shortest line to grab lunch. These real-time guide and recommendation engines rely on the low latency of AI servers installed at the edge. AI-driven wayfinding solutions exemplify the potential of Edge AI in creating more navigable and user-friendly environments while also driving unique revenue opportunities.

Smart Spaces: Crowd & Parking Management

Smart spaces, including venues and stadiums, are adopting predictive AI for crowd and parking management. By analyzing real-time data, these systems effectively manage foot traffic and parking logistics, enhancing the overall customer experience and improving safety. **One case study from a music festival showed that entry times were reduced by 30% following the implementation of AI crowd management systems[16].** Edge AI in these settings underscores the ability to maintain privacy while providing valuable insights and generating value from real-time processing.

Each of these verticals showcases the power of Edge AI, offering a glimpse into the future of industry-specific applications. By harnessing the right AI technology - predictive or generative -businesses can unlock new levels of efficiency, safety, and customer engagement. While some industries and use cases will develop slower than others, the meteoric rise in adoption and interest across industries is undeniable. Business leaders and technical teams in 2024 will be tasked with identifying their use cases and beginning to solve challenges on the road to implementation.

[12] Khanna, Ayesha. "How AI Is Reshaping Five Manufacturing Industries." Forbes, January 17, 2024. https://www.forbes.com/sites/forbestechcouncil/2024/01/17/how-ai-is-reshaping-five-manufacturing-industries/.

[13, 14] American Hospital Association. "How AI Is Improving Diagnostics, Decision-Making and Care." AHA Center for Health Innovation Market Scan, May 9, 2023. https://www.aha.org/aha-center-health-innovation-market-scan/2023-05-09-how-ai-improving-diagnostics-decision-making-and-care.

[15] Chernew, Michael, and Harrison Mintz. "Administrative Expenses in the US Health Care System: Why So High?" JAMA 326, no. 17 (2021): 1679–1680. https://doi.org/10.1001/jama.2021.17318.

[16] Crockett, Shomari. "AI's Watchful Eye: Orchestrating & Analyzing Crowds Like Never Before!" Secure AI Insights, October 23, 2023. https://secureaiinsights.com/2023/10/23/ais-watchful-eye-orchestrating-analyzing-crowds-like-never-before/.

Navigating the Complexities: Edge AI Implementation Challenges

While there are clear applications of Edge AI across multiple market verticals, bringing solutions to production is more challenging in practice than in theory. **Across industries, the U.S. Census Bureau's Business Trends and Outlook Survey recently found roughly two companies planning to use AI for every company currently using AI[17].** One of the hurdles in bridging this gap between goal and action is the number of technical components involved. The implementation of Edge AI presents a complex landscape of challenges that are significantly different from those encountered in more controlled data center environments. Addressing these challenges is critical for the effective deployment and operation of Edge AI solutions.

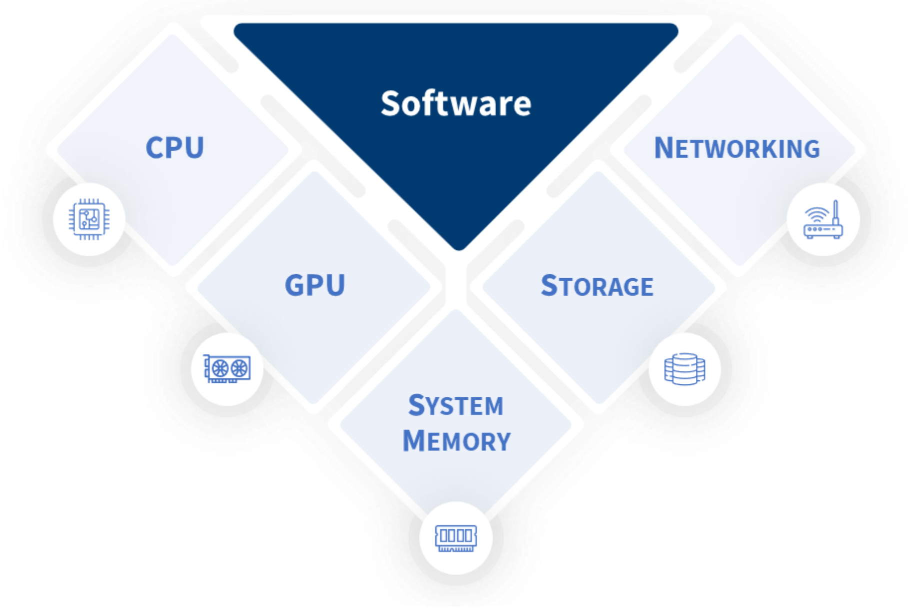


Figure 2 - Components of Edge AI

Size, Weight, and Power Constraints (SWaP):

One of the foremost challenges is adhering to SWaP constraints. Edge environments often lack the spaciousness and power availability of traditional data centers. If a system needs to be installed, for example, inside a hospital's existing network room, there is likely not significant free space. Alternatively, one might picture a speech-enabled food kiosk in an airport where all the hardware must fit inside a cramped and poorly ventilated chamber underneath the kiosk. In this scenario, access to higher voltage outlets might also be restricted. Supermicro's solution to these issues lies in its compact and efficient systems, designed to operate in confined spaces while maintaining high performance even if power supplies are limited. These systems are tailored to fit unconventional installation sites, ranging from wall-mounted setups to compact shelf placements, offering versatility without sacrificing power or efficiency.

Environmental Factors

Edge devices frequently operate in challenging environments, including exposure to dust, moisture, or extreme temperatures. Think of hardware that must be installed in a restaurant kitchen environment or outdoors in a networking cabinet at a sports stadium. Installations in a kitchen might be within an area that needs to be sprayed down frequently for cleaning, meaning containment is of the utmost importance. Another environmental consideration is the noise generated by Edge AI servers. Not only are Supermicro's high-reliability systems built to withstand harsh conditions, but they are also designed to dampen sound.

[17] U.S. Census Bureau. "Business Trends and Outlook Survey." January 18, 2024. https://www.census.gov/hfp/btos/about.

Together, these traits provide reliability, durability, and convenience. The robust build quality means they can function consistently and effectively, even in demanding operational scenarios.

Data Management Challenges

The data management challenges at the edge that need to be addressed include the volume and frequency of data generated and the numerous different types and formats of the data. Given these factors, it may be too costly or impossible to transmit this data to a central location for analysis, storage, and visualization. This means that efficient processing, storing, and securing data at the edge is crucial. Supermicro's systems leverage NVIDIA's software stack to address these challenges, providing powerful security, optimizing data processing, and ensuring that the data's value is maximized.

Hacking and data breaches at edge locations also need to be considered – as the physical security of the system comes into play. Encryption ensures that any stolen equipment does not represent a data security risk, while intrusion risk is reduced through physical locking systems.

Energy Efficiency and Network Connectivity

Given the often-limited power sources in edge environments, energy efficiency becomes paramount. At the same time, the consumption of energy generates heat. Consequently, Edge AI servers need to be designed with energy efficiency *and* efficient cooling in mind. Furthermore, ensuring reliable network connectivity is vital, especially in remote or mobile edge settings where network stability can be a concern. Many edge locations would not have originally been built with AI computing requirements in mind, which puts added pressure on finding robust hardware & software to work within these limitations.

For network connectivity, I/O formats may differ at the edge rather than the universal network connections found in a data center. Servers at the edge may connect to cameras or IoT devices that require USB or other formats. Therefore, Edge AI servers must account for these connectivity variations while maintaining energy efficiency, stable connections, and previously discussed requirements around SWaP.

Remote Management and Maintenance

Edge deployments necessitate robust remote management and maintenance capabilities. Owners must be able to install new software and manage these systems from a distance. For many organizations, this challenge may be a new one. Ensuring the health of the system, software installations, and reliable operations remotely is a challenge that Supermicro's solutions are equipped to handle. Although one upside to Edge AI is the reduction in data transfer costs to centralized storage locations, the downside that comes with it is the likely lack of technical staff being co-located with the hardware. Thankfully, Supermicro provides the necessary infrastructure for effective remote management, ensuring the longevity and effectiveness of Edge AI deployments.

Software Integration and Orchestration Challenges

The transition from Proof of Concept (PoC) to scaled deployment also involves software integration and orchestration. AI systems might ultimately interface with dozens of software systems already in use at that business or location. For example, a manufacturing setting might have a machinery execution system and a safety system. Despite executives and business leaders clamoring to implement AI for their businesses, many projects languish in PoCs due to technical challenges. The number of overlapping variables and components in a launch increases the risk of failing to achieve production.

Harmonious hardware and software designed for Edge AI help mitigate such risks. While Supermicro provides a robust hardware foundation, NVIDIA AI Enterprise – an enterprise-grade software platform that includes optimized AI frameworks, libraries, pre-trained models, and tools such as NVIDIA Triton™ Inference Server and NVIDIA® TensorRT™ – addresses the efficiency of utilizing the hardware, ensuring businesses achieve the highest inference density possible. Even though integrators deal with the majority of connections between the diverse systems interfacing with Edge AI systems, the combination of Supermicro's hardware and NVIDIA’s software nonetheless delivers proven results to reduce risk in this process while also facilitating excellent performance.

Challenges Unique to Predictive vs Generative AI Use Cases

The challenges in Edge AI are further nuanced when distinguishing between predictive and generative AI implementations. Predictive AI, primarily focused on analysis and forecasting, requires systems capable of processing large volumes of data efficiently to provide timely insights. Supermicro’s solutions cater to these needs with high-performance computing capabilities optimized for data-intensive tasks.

Generative AI introduces additional layers of complexity. It requires not just high computational power but also significant memory capacity, as it involves creating new, synthetic data. This requirement is where NVIDIA’s advanced GPU technologies are crucial, offering the memory and processing power needed for generative AI applications. Supermicro’s integration of these technologies ensures that businesses can leverage generative AI at the edge effectively, increasing speed from PoC to production and hedging risks inherent to complex tech stacks.

Both types of AI implementations have their distinct challenges, but through the combined expertise and innovative solutions of Supermicro and NVIDIA, these challenges are effectively addressed. This partnership ensures that businesses can harness the full potential of AI at the edge, whether it’s for predictive analytics or generative modeling.

Supermicro + NVIDIA: A Range of Solutions

At the heart of any Edge AI implementation is a robust and cohesive solution architecture that underpins the system. Supermicro and NVIDIA's collaborative approach provides just that—a comprehensive full-stack solution that integrates CPUs, GPUs, optimized memory, and the NVIDIA AI Enterprise software platform, all orchestrated within the resilient infrastructure of Supermicro's platforms. This architecture is designed to support a wide array of applications, from domain-specific tasks handled by specialized SDKs like NVIDIA RAPIDS™ for data science and NVIDIA Aerial™ for 5G to I/O-intensive operations facilitated by NVIDIA Magnum IO™.

The stack is engineered to ensure maximum efficiency and performance, featuring NVIDIA's industry-leading technologies, such as TensorRT for optimizing high-performance deep learning inference and Triton Inference Server for deploying AI models at scale. Middleware solutions like Base Command Manager Essentials for Edge AI and NVIDIA Fleet Command™ for deployment management ensure that the systems are not only powerful but also intelligently coordinated and easily integrated into existing workflows.

Security, storage, and networking are foundational elements of this architecture, guaranteeing that data integrity and transmission are never compromised. This robust backend is encapsulated within Supermicro's hardware, known for its reliability, and designed to meet the demands of a variety of edge environments. The result is a scalable solution architecture that empowers end-users to unlock the full potential of Edge AI. This architecture is shown in Figure 3 below.

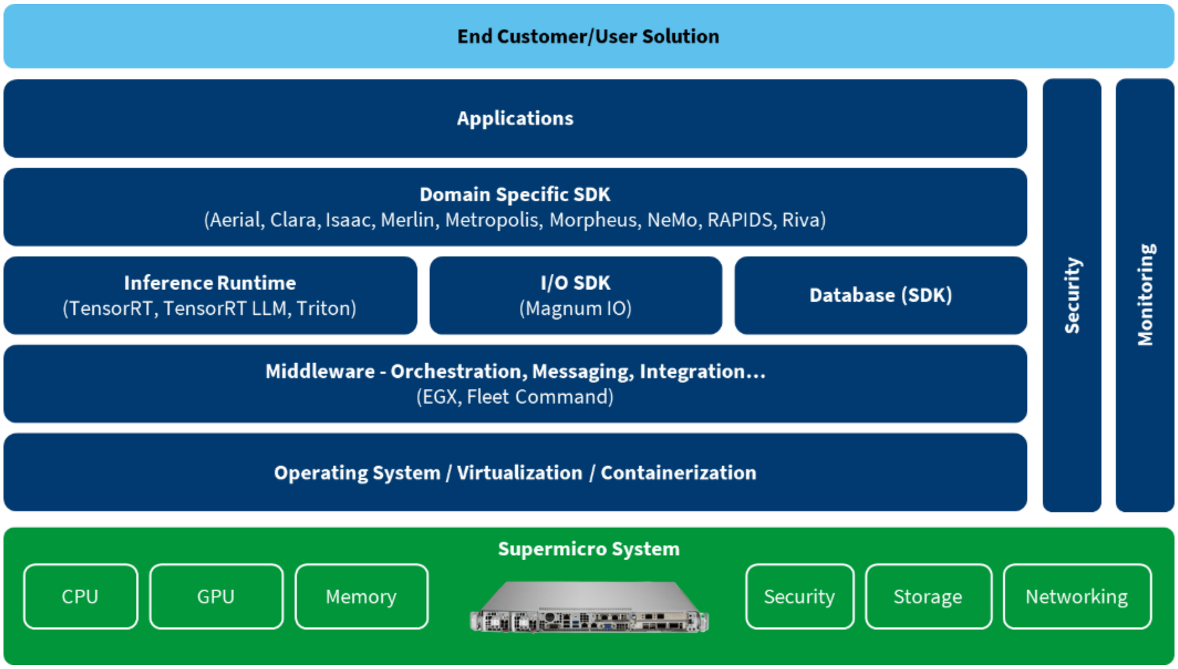

End Customer/User Solution
Applications
Domain Specific SDK
(Aerial, Clara, Isaac, Merlin, Metropolis, Morpheus, NeMo, RAPIDS, Riva)
Inference Runtime
(TensorRT, TensorRT LLM, Triton)
I/O SDK
(Magnum IO)
Database (SDK)
Middleware - Orchestration, Messaging, Integration...
(EGX, Fleet Command)
Operating System / Virtualization / Containerization
Security
Monitoring
Supermicro System
CPU
GPU
Memory
Security
Storage
Networking

Figure 3 - The Comprehensive Landscape

Supermicro and NVIDIA provide a diverse range of solutions tailored to numerous Edge AI applications, ensuring that organizations can select the right combination of hardware and software for their specific needs. For instance, Supermicro's SYS-221HE is an excellent choice for large, multi-GPU inferencing and training workloads, offering robust performance for demanding AI applications. For more compact scenarios, the SYS-E300 and SYS-E403-12P offer powerful multi-GPU inferencing in a more space-efficient package. For medium workloads where single GPU-based inference is sufficient, the SYS-110D provides an optimal balance between performance and space utilization.

Supermicro's Resilient Solutions for Extreme Edge Environments

Supermicro's focus on high-reliability systems is evident in its product offerings, which are designed to withstand challenging operational environments. These environmental challenges are often made more complex by overlapping constraints on physical space. Thankfully, the range of offerings allows for reliable systems in various form factors. For example, the Supermicro SYS-E403-12P offers compact multi-GPU inferencing capabilities, ideal for edge deployments in industrial, retail, or smart city environments where space and environmental conditions are a concern. These systems ensure reliable operation even in adverse conditions, making them an asset for industries that require robust and dependable computing at the edge.

Selecting the Optimal Supermicro System for Your Edge AI Application

Supermicro and NVIDIA excel in guiding organizations to select the right system for their specific Edge AI applications. This support involves considering factors like the size of AI models, system compatibility, and specific use case requirements. Whether it's handling large-scale video analytics in smart cities or processing complex AI algorithms in healthcare diagnostics, Supermicro's SYS-E300, SYS-110D, SYS-E403, and SYS-221HE combined with NVIDIA's powerful AI and computing platforms, provide a range of solutions to meet these diverse needs effectively.

The versatility of Supermicro and NVIDIA’s solutions is key to their widespread applicability across various industries. Their systems are not limited to specific sectors but are adaptable to various computational demands. From enhancing diagnostic imaging in healthcare with the SYS-221HE to optimizing traffic management in smart cities with the SYS-110D, these solutions demonstrate the ability to address a broad spectrum of Edge AI applications.

Small	Medium	Large	Extra Large
SYS-E300-13AD	SYS-110D-20C-FRAN8TP*	SYS-E403-13E-FRN2T	SYS-221HE-TNR(D)
Dimensions (mm)			
H W D 43.0 x 264.8 x 225.8	H W D 43.0 x 437.0 x 399.0	H W D 117.3 x 266.7 x 406.4	H W D 88.9 x 436.9 x 574.0
Single NVIDIA A2, L4 or RTX A1000	**Single NVIDIA L4 or RTX 4000 Ada**	**Single NVIDIA L40S, L40, or RTX 6000 Ada**	**Multiple H100PCIe, L40S, or RTX 6000 Ada**
Key Features for Predictive AI			
AI computer vision for up to 8 streams[1]	AI computer vision for up to 16 streams[1]	AI computer vision for up to 32 streams[1]	AI computer vision for up to 48 streams per GPU[1]
Up to ~2,000 automatic speech recognition (ASR) samples per second[2]	Up to ~6,000 automatic speech recognition (ASR) samples per second[2]	Up to ~22,000 automatic speech recognition (ASR) samples per second[2]	Up to ~32,000 automatic speech recognition (ASR) samples per second per GPU[2]
Key Features for Generative AI			
LLM up to 8 billion parameters	LLM up to 24 billion parameters	LLM up to 48 billion parameters	LLM up to 80 billion parameters
Image and video generation with Stable Diffusion at ~1 image every 4-5 seconds.	Image and video generation with Stable Diffusion at ~1 image every 2 seconds.	Image and video generation with Stable Diffusion at ~1-2 images per second.	Image and video generation with Stable Diffusion at ~3 images per second.

*Additional models include SYS-110D-4C-FRAN8TP, SYS-110D-8C-FRAN8TP, SYS-110D-14C-FRAN8TP, and SYS-110D-16C-FRAN8TP

[1] Based on using an image classification model similar to EfficientNet-B4, dependent on video stream compression and other workloads on the system

[2] Based on using an ASR model like QuartzNet

Figure 4 - Most Commonly Used Supermicro Solutions for Edge AI Applications

The Path Forward

The exploration and adoption of Edge AI, encompassing both predictive and generative AI, underscores a significant shift in how industries operate and innovate. **Since 2022, interest among executives has risen swiftly to the point that surveyed leaders predict AI to be a critical function in 20% of their operations by 2025**[18].

Predictive AI, with its ability to forecast and analyze, continues to be a cornerstone in decision-making processes across various sectors, from retail to manufacturing. Generative AI, on the other hand, has rapidly emerged as a transformative force, offering new possibilities in customer engagement and operational efficiency. Supermicro and NVIDIA innovation and partnership have led to a range of solutions that have successfully demonstrated how these technologies can be effectively harnessed to drive growth, enhance efficiency, and foster innovation. The Edge AI landscape is vibrant and diverse, with each industry finding unique and powerful applications for these technologies.

Integration of AI at the edge is not just a technological advancement; it represents a total shift in business operations and customer interactions. By bringing computation and AI closer to the point of data generation, businesses are achieving faster, more responsive, and more personalized outcomes. This evolution is pivotal in an era where real-time insights and actions are increasingly critical for success.

Businesses seeking to leverage Edge AI have a clear pathway forward with Supermicro and NVIDIA. Their combined expertise and range of solutions offer a solid foundation for any Edge AI initiative. This synergy reduces risk and increases the speed of arrival in production. In turn, successful implementations can contribute to more delightful customer experiences, increased revenue, and improved safety across various industry applications.

For More Information

To learn more, visit our Edge AI solution page at www.supermicro.com/edge-ai.

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

Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com.

[18] MIT Technology Review Insights. (2023). *The Great Acceleration: CIO Perspectives on Generative AI*. Cambridge, MA: Massachusetts Institute of Technology.