WHITE PAPER

HOW AI IS PROPELLING INNOVATION IN FINANCIAL SERVICES

Supermicro Infrastructure with NVIDIA AI Enterprise Deliver Best-In-Class Outcomes for Financial Services Companies to Help the Sector Continue to Grow and Transform.

Table of Contents

Introduction

Over the past five years, Artificial Intelligence (AI) has radically reimagined how organizations engage with each other, their employees, and their customers. From 2017 to 2022, AI adoption has doubled from 20 percent to over 50 percent according to a McKinsey survey[1]—and that was before generative AI (GenAI) expanded its uses even further. With the release of ChatGPT in 2022, GenAI has changed not just how businesses operate, but how customers seek answers to their questions.

Current Landscape: Early Adopters Lead the Way

Over the last few years, the financial services industry has been working to integrate both predictive and generative AI into their business practices. Banking alone is expected to represent one of the largest opportunities for AI, with an annual potential of $200 billion to $340 billion in added value (equivalent to 9 to 15 percent of operating profits).[2] Use cases such as Machine Learning for algorithmic trading have been in use for years and have seen widespread successful adoption, and new use cases stand to elevate the significance of AI in the sector.

Early adopters of AI are already beginning to make topline contributions and stand out from their competitors. AI is enabling these organizations to offer new and innovative services that increase their efficiency, improve customer experiences, and reduce costs. As such, taking AI initiatives from the drawing board to production is now a top priority for executives looking to capitalize on these opportunities, whether their business is in banking, advanced trading, or insurance.

[1] McKinsey, "The State of AI in 2022—and a Half Decade in Review"
[2] Scaling GenAI in Banking: Choosing the Best Operating Model

In fact, Gartner reports that by the end of 2024, 75 percent of enterprises will shift from piloting to operationalizing AI.[3] To truly gain value from these projects, organizations need to make it easy for their employees to use AI and integrate it into their daily workflows. Employees must also be able to trust that the AI's results are accurate and beneficial to their work, especially beyond the pilot and testing stages in real-world situations. For customers, GenAI solutions for example chatbots also need to be simple to use and be trusted to give accurate advice.

Common Roadblocks for the Sector

Financial services organizations are tightly regulated and often have legacy systems that are difficult to update, requiring significant investment to integrate new technology offerings. To deploy AI effectively, organizations need AI-ready infrastructure that's easy to adopt and implement. Working with updated technology ensures that the sector can effectively analyze data and produce trustworthy solutions, allowing AI to deliver on its promise of transforming how financial services organizations interact with their customers. From large scale banking, trading, and call center services to in-person banking, insurance, and loan processing services at the edge where it's valuable to localized processing of data, AI is poised to transform the industry.

For example, AI can play a major role in protecting consumers by combating fraud. The consequences of fraud are costly, with credit card fraud causing merchants and card acquirers globally to lose more than 30 billion U.S. dollars in 2021 alone,[4] and AI offers a new avenue to protect customers and organizations' bottom line. Fraudsters use AI in their efforts to commit credit card fraud, circumvent Know Your Customer (KYC) verification, and scam customers. To stay ahead, financial services organizations must adopt AI for use cases such as advanced identity verification and credit card anomaly scanning.

Benefits and Common Use Cases

The benefits of AI adoption will continue to expand for organizations and their customers as new use cases and capabilities are developed. A common application today includes providing customers with immediate and accurate answers to questions. Tailored marketing is delivered both in-app and via email. AI also equips traders with algorithms that identify split-second opportunities. Additionally, insurance and loan providers can access unprecedented amounts of data to make more informed decisions. Because the opportunities AI presents across the financial services industry continue to evolve, this paper focuses on five common use cases making a major impact in the industry today.

Five Key AI Use Cases in Financial Services:

1. **Quant Finance:** Quantitative finance integrates AI and quantitative modeling with data science to help organizations make intelligent business decisions. Organizations can leverage mathematical models, AI and data analytics to discern trends, forecast asset valuations, manage risk, and optimize investment portfolios.
2. **Smarter Trading with Alternative Data:** Alternative data sets—including consumer transactions, logistics information, and employment trends—offer an informational edge that enhances trading strategies. AI models can analyze large volumes of financial data to generate predictive insights, optimize trading strategies, visualize anomalies and automate decision-making processes in trading operations.
3. **KYC, AML and Fraud Prevention Security**: AI helps streamline customer verification programs for Know Your Customer (KYC) processes and offers transaction monitoring by scanning global payment networks. It flags credit card transactions and collects evidence to protect customers and ensure compliance with regulatory practices.

[3] Gartner Top 10 Trends in Data and Analytics for 2020
[4] Statista, "Total value of losses due to card fraud, either credit card fraud or debit card fraud, worldwide from 2014 to 2022," 2023

4. **Intelligent Document Automation*:*** AI can streamline financial processes by automating tasks related to document handling, data extraction, and content validation. This automation not only increases efficiency but also reduces the risk of human error and accelerates AI workflows. GenAI helps ensure compliance in areas such as loan processing, contract management, and regulatory reporting. Additionally, insurance companies can speed claims intake, review photos and videos submitted by customers, and make decisions faster with high level analysis and decision-making.
5. **Customer Experience with Chatbots:** GenAI-enabled chatbots and virtual assistants improve customer interactions. Organizations can be more efficient with their support staff and deliver more personalized, customer-aware engagements for example customized financial recommendations for customers.

By establishing robust AI infrastructure to unlock these core use cases, financial services organizations also create a foundation for exploring other advanced applications such as high-frequency trading which require similarly performant infrastructure. Ultimately, embracing AI now sets the stage for organizations to capture new opportunities for innovation and solve new use cases that are all dependent on the right infrastructure being in place.

Navigating the Complexities: AI Implementation Challenges

While implementing AI in the financial services industry has the potential to be transformative, restrictive regulations and requirements for data privacy means that businesses must overcome several hurdles. Organizations such as banks often have legacy systems that must be upgraded before they leverage AI systems—meaning that investments in data capture and data accuracy, workforce knowledge, and system upgrades are necessary to realize high-level returns on investments from AI. However, despite these challenges, it is crucial to address and navigate these hurdles effectively to fully harness AI's benefits. The following section will delve into these challenges in detail.

Data Privacy, Sovereignty, and Disparate Locations

- Securing Personal Data: AI systems require huge quantities of often sensitive customer financial data, for instance bank account details and transaction histories, which must be handled carefully to maintain compliance. Building your own on-premises AI infrastructure is challenging and requires significant investment, but it allows you to secure your data within your own facility, offering direct control over security measures. Conversely, cloud computing provides scalable storage and processing power but requires stringent security protocols to ensure data protection and regulatory compliance.
- Changing Regulatory Landscape: As AI evolves rapidly, regulatory requirements need to keep pace. In the US, the 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence' was issued in October 2023. Similarly, the EU's AI Act was enacted in March 2024. These regulations vary across jurisdictions and are updated frequently to reflect the rapid pace of AI innovation.
- Elevated Risks Caused by a Breach: A breach from a poorly configured and under-secured system could be costly and damage an organization's reputation. Financial services executives are particularly concerned with customer privacy, with 69 percent identifying it as a high concern compared to 56 percent in all industries surveyed by KPMG in 2023.[5]
- Data Sovereignty Concerns: Maintaining data sovereignty creates additional concerns for multi-national organizations that must adhere to the often-disparate regulations of different countries.

Modern AI infrastructure solutions can help navigate these challenges by providing robust data management capabilities through scalable cloud platforms. They also ensure compliance with evolving regulations via automated regulatory compliance tools, advanced security measures such as end-to-end encryption, and AI-driven threat detection. Modernized AI infrastructure also facilitates adherence to international data sovereignty for the safe, secure, and compliant use of AI in financial services.

[5] KPMG, The Generative AI Advantage in Financial Services, 2023

Workforce Challenges

- Recruiting Shortfalls: In a 2023 NVIDIA survey, 32 percent of respondents identified recruiting[6] and retaining AI experts and data scientists as the biggest challenge to achieving their organization's AI goals. Without an experienced workforce, organizations cannot operationalize AI, and they cannot offer ongoing support to employees to build trust and see returns.
- Technology Adoption Challenges: Often employees resist adopting newer technologies that require significant upskilling and agile adaptation to rapidly changing systems and processes. To mitigate these challenges, organizations need an AI platform that is easy to deploy, integrates well into existing workflows, is simple to operate and manage, and includes tools that are familiar to teams in financial services today.

Data Challenges for Model Training and Accuracy

- Lack of Data Volume to Train an AI Model: Without sufficient data, an AI model may not capture the full variability of a problem, resulting in overfitting, where the model becomes too tailored to the training data and fails to generalize, or underfitting, where the model is too simplistic and misses important patterns and complexities inherent in the data, leading to poor performance on new, unseen data.
- Need for Data Consistency: Ensuring data consistency through robust data governance is crucial for accurate AI insights, as inconsistencies in data formats and unaddressed biases can lead to misleading results and flawed decision-making.
- Need for Accuracy of Data: Ensuring the accuracy of data is crucial to the effectiveness of AI applications. For instance, if an AI model is trained on data from one industry (for example trading) but applied to another (for example banking) it may produce inaccurate insights and recommendations.
- Improper Data Labeling: Improperly labeled data can significantly impact AI insights. For instance, if sentiment analysis training data for customer reviews lacks proper labeling to accurately distinguish between positive and negative sentiments, the resulting AI model may misclassify sentiments, leading to unreliable analyses and recommendations.

Addressing these complex challenges requires organizations to optimize data management, label data precisely, and ensure accuracy when training AI models. By leveraging advanced technologies tailored to these specific needs, organizations can enhance data integrity, refine AI model accuracy, and unlock unparalleled insights that drive smarter decisions and create a competitive advantage in today's dynamic digital landscape.

Legacy Systems and Infrastructure

- Outdated Infrastructure: Many banks utilize aging core banking systems (CBS), some dating back as far as the 1970s. Without efforts to modernize legacy infrastructure, implementing new technologies especially around AI can be almost impossible.
- High Costs for Innovation: Maintaining these systems can be highly costly and can prevent organizations from investing in innovative technology such as AI that is becoming increasingly necessary to compete. Out of date infrastructure often results in lagging productivity and slower performance for the entire organization.
- Challenges Integrating New Technology: Assembling AI infrastructure alongside existing legacy systems is challenging, costly, time-consuming, and requires a highly skilled technical workforce.

It is easier to overcome this model by investing in an existing AI-ready platform built into a hardware stack that is tailor-made to run AI workloads and can exist alongside your legacy systems.

[6] NVIDIA, State of AI in Financial Services:2024 Trends

Supermicro + NVIDIA: Enabling Key Use Cases for AI in Financial Services

Supermicro's versatile and robust infrastructure solutions seamlessly integrated with NVIDIA's cutting-edge AI platform, GPUs, and software can help organizations overcome these challenges. Moreover, they can scale up and down depending on the needs of the industry and more quickly realize the opportunity AI represents.

While there's a myriad of ways to leverage AI in the financial services industry, certain applications stand out for their proven impact. Here, we'll delve into real-world examples showcasing the adoption of AI-specific use cases by the financial services sector across banking, insurance, and trading. Although each implementation varies in the usage of predictive, generative, or both types of AI, and the scale of deployment, the shared thread is that all of them have proven ROI in production. Financial services organizations are already finding that AI can reduce costs, increase revenue, and improve workforce or customer safety.

Let's delve into some of the most effective AI implementations being done today:

1. Quant Finance

For traders using AI to enable quant trading, speed and computational power are key. Quant traders use advanced algorithms to sift through large quantities of data and create advanced risk assessments to execute smart trades and manage portfolios. AI allows for more accurate predictions, better risk management, and improved overall trading strategy performance. These algorithms are highly confidential and proprietary, to the extent that organizations usually refrain from running them in remote locations to safeguard their intellectual property and maintain strict control over their execution environments. Traders need a system that can guarantee low latency and make enhanced predictions by using on-premises scalable data centers and make enhanced predictions, enabling them to complete smarter trades quickly. To illustrate how AI specifically supports these needs in quantitative finance, the following points outline some key ways in which AI contributes to improved trading strategies and decision-making.

- **Big Data Processing:** Analyze vast amounts of structured and unstructured data, including financial reports, news articles, social media sentiment, and economic indicators with AI.
- **Algorithm Optimization:** Optimize trading strategies by continuously scanning market data and adjusting parameters to maximize returns and minimize risk.
- **Market Sentiment Indicators:** AI-driven insights enable traders to anticipate market movements driven by public sentiment and develop sentiment-based trading strategies.
- **Portfolio Management:** AI optimizes portfolio allocation by considering risk-return trade-offs. Machine learning models adjust portfolios dynamically based on changing market conditions.

Supermicro + NVIDIA

Supermicro and NVIDIA offer the speed and power to take advantage of split-second opportunities. With the low latency NVIDIA AI Platform and the NVIDIA Triton inference system powered by NVIDIA GPUs running in Supermicro servers designed for high-performance computing, traders can digest and respond to new data quickly.

Supermicro offers a range of high-performance, powerful, and scalable solutions designed to meet the demands of traders. The SYS-521GE-TNRT and AS -4125GS-TNRT systems provide robust, scalable on-premises infrastructure that helps ensure low latency and secure execution of proprietary algorithms. For even greater computational power, the ARS-111GL-NHR and ARS-111GL-NHR-LLC systems, along with the ARS-221GL-NHIR (featuring dual GH200 GPUs with NVLink), deliver the speed and efficiency required for big data processing and real-time trading strategy optimization. These solutions are engineered to handle complex AI workloads, enabling quant traders to make smarter, faster decisions—helping maximize returns and minimizing risk in competitive financial markets.

2. Smarter Trading with Alternative Data

Trading organizations can use AI to move beyond traditional metrics, i.e., earnings reports and price trends by analyzing alternative data to make smarter, more informed trades. Organizations using this technology can instantly view trades in the market and respond to them, using contextual knowledge gleaned from a variety of non-traditional sources. These sources can include anything from trends in the news and social media, to annual reports and sentiment analysis, to satellite images of retail environments. Ultimately, AI enables organizations to make the right trade at the right moment by considering all relevant factors. Alternative data helps organizations make quicker and more informed decisions than humans could alone to create and capitalize on opportunities that might otherwise have been missed.

- **Gain a Competitive Advantage:** Financial services organizations that leverage alternative data with AI can gain a competitive edge by uncovering unique insights and investment opportunities.
- **Enhanced Predictive Power:** Alternative data can provide early signals of market trends and company performance that traditional data sources might miss.
- **Risk Management:** AI-driven analysis of alternative data can help in identifying potential risks and market anomalies, improving overall risk management strategies.

Supermicro + NVIDIA

Supermicro's fully integrated large-scale AI-training infrastructure along with their inference systems utilizing NVIDIA's GPUs allows traders to take advantage of powerful AI models that review conventional and alternative data and use it to make strategic trades. Users can collect alternative data from the edge with Supermicro and NVIDIA's advanced Edge AI capabilities and use it to make more informed trades. The Supermicro AI platform powered by NVIDIA GPUs delivers a 1,000x speedup over the previously set benchmark for back testing in algorithmic trading. With NVIDIA GPUs running in powerful Supermicro servers, traders can respond to surges of data, reducing latency and expediting the trading process.

To effectively leverage alternative data in smarter trading strategies, Supermicro's cutting-edge systems provide the computational power and scalability necessary for real-time AI-driven analysis of large datasets. The SYS-521GE-TNRT and AS -4125GS-TNRT systems deliver high-performance processing capabilities, ensuring that trading organizations can quickly analyze and act on a diverse range of data sources, from news trends to satellite imagery. For enhanced performance, the ARS-111GL-NHR and ARS-111GL-NHR-LLC systems, along with the dual-GPU ARS-221GL-NHIR model featuring NVLink, offer advanced data processing and AI capabilities that support rapid decision-making and risk management. Additionally, the SYS-821GE-TNHR and AS -8125GS-TNHR systems are designed to handle the complexities of alternative data, enabling organizations to gain a competitive edge through unique insights and more informed trading decisions. These solutions are ideal for financial services firms seeking to capitalize on the predictive power of AI and maintain an edge in today's fast-paced markets.

3. KYC, AML, and Fraud Prevention Security

Just as financial services institutions add AI to their tool belts, so do fraudulent actors. Financial services organizations must deploy AI algorithms to learn from previous attacks and scan for increasingly advanced threat tactics. Malicious activity perpetrated against customers such as fraudulent customer identity verification and money laundering can also be a serious compliance and reputational risk for financial services institutions. AI can help mitigate these challenges by streamlining the Know Your Customer (KYC) and Anti-Money Laundering (AML) processes by analyzing customer data, transaction histories, and risk profiles to detect suspicious activities, for example money laundering or terrorist financing, and ensure that customers meet regulatory requirements. AI can also analyze a broad picture of a customer's financial history to assign risk scores to customer accounts that may engage in fraudulent activities.

One challenge to identifying patterns of fraud in customer transactions such as credit card payments is the vast amount of data. AI can speed up the process of identifying abnormal transactions by scanning transactions in real-time, allowing customers and organizations to quickly freeze accounts and limit losses. Because vendors and credit card issuers often take on the cost of fraudulent charges, limiting the number of fraudulent charges is an urgent priority. Once fraud has occurred, AI can assist fraud investigators by compiling evidence, analyzing transaction histories, and creating reports that can be shared with law enforcement.

- **Improved Compliance:** AI can help organizations review more data when completing KYC and AML processes to ensure compliance is maintained.
- **Reduce False Positives:** By refining models and continuously learning from new data, AI can reduce the number of legitimate transactions that are incorrectly flagged as fraudulent.
- **Cost Savings:** Automating fraud detection and response processes can reduce the operational costs associated with manual fraud investigation.

Supermicro + NVIDIA

Supermicro solutions using NVIDIA GPUs allow you to create advanced fraud detection algorithms. By developing Machine Learning (ML) models optimized with NVIDIA TensorRT and running on NVIDIA Triton Inference Server, financial services organizations can scan transactions for anomalies and view suspicious activity in context.

The SYS-221H-TNR, AS -2025HS-TNR, and AS-2015HS-TNR, SYS-221GE-NR, and ARS-221GL-NR are advanced Supermicro systems, that provide the computational power necessary for real-time AI-driven fraud detection as part of, KYC, and AML processes. With systems ARS-111GL-NHR-LLC and ARS-111GL-DNHR-LCC (2-node), and the SYS-521GE-TNRT and AS -4125GS-TNRT, organizations can efficiently analyze vast amounts of data to identify suspicious activities. The ARS-221GL-NHIR with dual GH200 GPUs and NVLink ensures rapid processing to immediately respond to threats and suspicious activity. These systems support financial institutions in maintaining compliance, reducing false positives, and automating fraud prevention, ultimately safeguarding both the organization and its customers.

NVIDIA RAPIDS™, an open-source software speeds transactions with GPU-accelerated data processing. Financial services organizations can run advanced identity verification and anomaly detection in large and small data centers in conjunction with edge AI sensors and computing centers.

4. Intelligent Document Automation

Many financial services processes require considerable documentation to make informed decisions in situations such as loan approvals and insurance claims decisions. AI-powered Optical Character Recognition (OCR) can convert scanned documents, PDFs, and images into machine-readable text while Natural Language Processing (NLP) algorithms can extract critical information to streamline their document workflows, reduce operational costs, and improve overall service quality.

For insurance claims management, AI can speed the intake process by integrating with external sources leveraging Retrieval-Augmented Generation (RAG) for enhanced data retrieval, and utilizing smart image and video review, thereby reducing the need for in-person inspections. The insurance claims process can often be a frustrating experience for many customers, so giving them easy to use tools for instance AI-powered applications to submit photos and have their status explained to them can improve experience and ultimately save organizations labor and money.

- **Credit Scoring and Loan Underwriting**: AI analyzes a wide range of data sources to assess the creditworthiness of individuals and businesses, leading to more accurate and fair lending decisions.
- **Increased Efficiency:** AI significantly speeds up document processing times, reducing the need for manual intervention and allowing staff to focus on more strategic tasks.
- **Scalability:** AI systems can handle large volumes of documents, making it easier to scale operations without a corresponding increase in manual workload.
- **Improved Accuracy:** AI reduces human errors associated with manual data entry and document handling, leading to higher data integrity and reliability.

Supermicro + NVIDIA

With NVIDIA's advanced GPUs on Supermicro's AI optimized systems, financial services organizations can build high-powered AI models that can reduce the length of the insurance claims process from days to seconds. Data can be processed in advanced data centers and stored at the edge for more rapid access.

The SYS-221H-TNR, AS -2025HS-TNR, and SYS-221GE-NR systems are part of Supermicro's powerful hardware solutions, that are designed to support AI-driven intelligent document automation in financial services. These systems enable rapid Optical Character Recognition (OCR) and Natural Language Processing (NLP) for efficient document handling, reducing manual intervention and improving service quality. For more complex tasks, like insurance claims processing and loan underwriting, systems like the ARS-221GL-NR and ARS-111GL-NHR-LLC ensure seamless integration with AI-powered tools for smart data retrieval and image analysis. With the SYS-521GE-TNRT and ARS-221GL-NHIR (featuring dual GH200 GPUs) systems, organizations can handle large volumes of documents with high accuracy, scalability, and speed, leading to better decision-making, reduced costs, and superior customer experiences.

5. Customer Experience with Chatbots

AI has the potential to transform how customers interact with the financial services industry. Rather than having to spend time on the phone to reach a person in a call center, often with limited hours, customers can get quick personalized service based on a comprehensive view of their account. Chatbots can also streamline and simplify the process of setting up and using a bank account and reach new customers that might lack technology or language skills. Once customers have established accounts, organizations can use generative AI to create personalized upselling recommendations which is, often a time-consuming task for marketing teams.

- **Personalization:** AI enables highly personalized interactions and recommendations, making customers feel valued and understood.
- **Accessibility:** AI-driven tools provide 24/7 support and access to financial services, making it easier for customers to manage their finances anytime, anywhere.

- **Efficiency:** Automated processes reduce wait times and improve the speed of service delivery, enhancing overall customer satisfaction.

Supermicro + NVIDIA

Supermicro's AI-optimized systems, combined with NVIDIA GPUs, help organizations create intelligent chatbots, copilots, and virtual assistants. These tools can access data to improve efficiency and provide a competitive edge. Chatbots are developed with a method called RAG (RAG or Retrieval-Augmented Generation is a framework that combines the capabilities of large language models (LLMs) with information retrieval systems to improve the accuracy of AI-generated text), which links large language models to a company's knowledge base, unlocking many new possibilities. RAG enables companies to utilize document repositories without needing to retrain their LLM.

For organizations looking to implement chatbots, Supermicro's solutions deliver exceptional performance in personalization, accessibility, and efficiency. The SYS-221H-TNR, AS -2025HS-TNR, and AS -2015HS-TNR system are designed to handle the high-performance computing needs of sophisticated chatbots and generative AI applications. The SYS-221GE-NR and ARS-221GL-NR offer robust infrastructure for scalable, 24/7 customer support platforms, ensuring reliable and efficient service delivery. The SYS-821GE-TNHR and ARS-111GL-NHR-LLC systems, provide the necessary computational power and thermal management for real-time personalization and upselling recommendations.

NVIDIA® Riva expands access and improves service by enhancing chatbots with advanced speech recognition and speech synthesis, language understanding, and vision learning models.

Ultimately, this helps the sector elevate the customer experience and empowers employees to focus on strategic and critical tasks while routing routine engagements to AI-powered tools.

Supermicro + NVIDIA: A Range of Solutions

At the heart of any 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.

Selecting the Optimal Supermicro Systems for Your AI Applications

Supermicro's cutting-edge AI-ready infrastructure solutions help with large-scale training to intelligent edge inferencing, enabling financial services organizations to streamline and accelerate AI deployment. Their AI infrastructure empowers workloads with optimal performance and scalability while optimizing costs and minimizing environmental impact.

Supermicro and NVIDIA excel in guiding organizations to select the right system for their specific AI applications. This support involves considering factors such as the size of AI models, system compatibility, and specific use case requirements. Whether it's handling large-scale data for alternative trading or processing LLMs in building chatbots, **Supermicro's portfolio of platforms offers unparalleled breadth, available across a wide range of form factors. This extensive range allows Supermicro to deliver custom-tailored solutions that precisely meet each customer's needs, budget, and power requirements, ensuring an optimal fit for any deployment scenario.** When combined with NVIDIA's powerful GPUs, it provides a range of solutions to meet these diverse needs effectively. The solution is designed to meet customers at any stage of their AI journey and can be tailored to their specific needs, including cost, power, and other related requirements.

The versatility of Supermicro and NVIDIA's solutions is key to their widespread applicability across different use cases in financial services and across other industries. Their systems are adaptable to various computational demands.

Figure 1– Most Commonly Used Supermicro Solutions for AI Applications

Security, storage, and networking are foundational elements of this architecture, guaranteeing that data integrity and transmission are never compromised. The robust backend is encapsulated within Supermicro's hardware, known for its reliability, and designed to meet the demands of a variety of environments. The result is a scalable solution architecture that empowers end-users to unlock the full potential of AI. This architecture is shown in Figure 2 below.

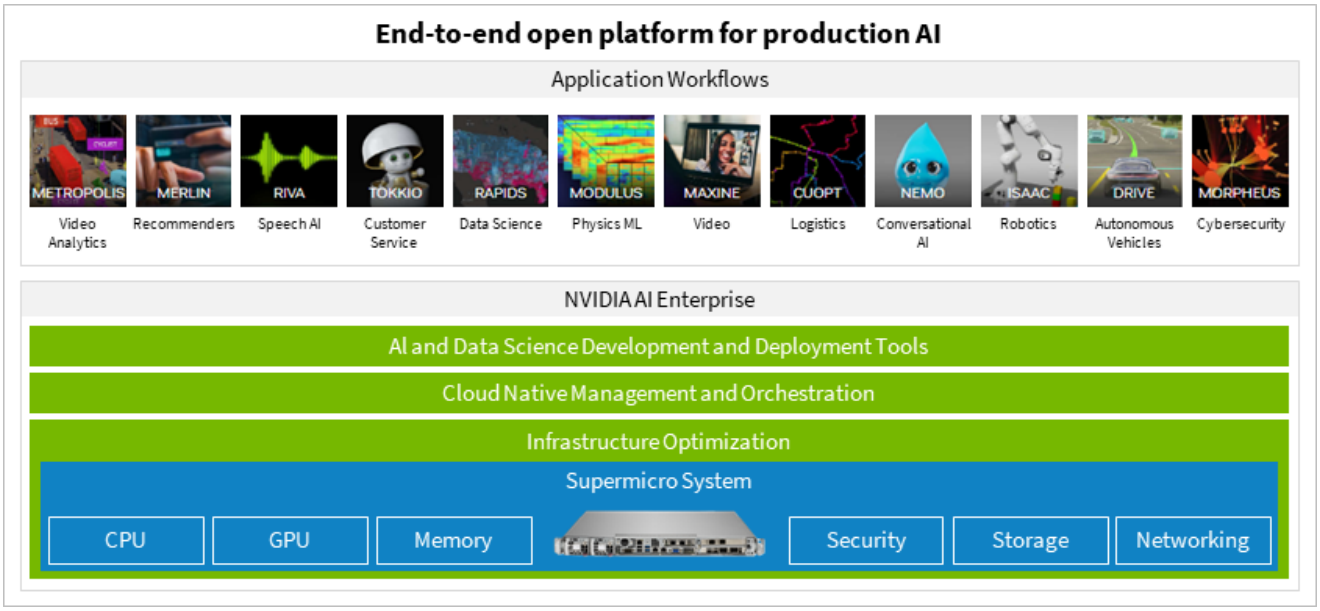

Figure 2 – The Comprehensive Solution

The Path Forward

The AI landscape is vibrant and diverse, with each industry finding unique and powerful applications for these technologies. The exploration and adoption of AI is already underway in the financial services industry, and organizations that most successfully implement it into their workflows are poised to generate the most value. Those that do not could be left behind. Over 80 percent of financial services executives surveyed in 2023 expect their organization's GenAI investment to increase by 50 percent or more, with 41 percent expecting it to increase by over 100 percent.[7]

Predictive AI, with its ability to forecast and analyze, continues to be a cornerstone in decision-making processes across various sectors, from financial services 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's partnership and innovation has 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.

Integration of AI 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 AI have a clear pathway forward with Supermicro and NVIDIA. Their combined expertise and range of solutions offer a solid foundation for any 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 operations across various financial services applications.

[7] The Generative AI Advantage in Financial Services

For More Information

To learn more, visit our Edge AI solution page https://www.supermicro.com/en/solutions/ai-deep-learning.

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.