



nVIDIA

WHITE PAPER

AI-POWERED TELECOM: SECURING THE FUTURE AND PROFITABILITY OF THE SECTOR

Leveraging Artificial Intelligence to Drive Innovation, Redefine Business, and Boost Revenue in the Telecommunications Industry.

Table of Contents

Introduction.....	1
Key Use Cases for AI in Telco	4
AI in Telco: Key Use Cases for Operational Efficiency.....	4
AI in Telco: Key Use Cases for Fueling Growth with New and Expanded Services.....	5
Navigating the Complexities: AI Adoption and Implementation Challenges	5
Supermicro with NVIDIA – Helping Telcos Embrace AI to Optimize Operations and Unlock Strategic Growth.....	6
Key Use Cases for Operational Efficiency.....	7
1. Reimagining Contact Centers to Enhance Customer Experiences	7
2. Streamline Network Operations.....	8
3. Predictive Analytics for Business Insights.....	8
AI in Telco: Key Use Cases for Fueling Growth with New and Expanded Services.....	9
1. AI Infrastructure for Sovereign AI Factories	9
2. 5G Monetization with Edge AI and 6G Research.....	10
Supermicro and NVIDIA: A Range of Solutions.....	11
Moving Forward	13
For More Information	13

Introduction

AI: Igniting the Next Evolution in Telecommunications

Organizations across the telecommunications industry (telcos) stand at a critical juncture. Telcos play a pivotal role in shaping the world's digital ecosystem, driving economic growth, and connecting people across the globe. As the sector with the potential to digitize entire countries and partner with various industries, telcos have the unique ability to empower other sectors and fuel widespread digital transformation. By leveraging cutting-edge AI technologies, telcos can enhance operational efficiency, deliver superior customer experiences, and stay competitive in an increasingly dynamic market.

“The telecommunications industry stands at the crossroads of AI, connectivity, security and sustainability in a digitized world. It drives the AI revolution and [shoulders the responsibility of securing critical infrastructure and safeguarding our world](#).¹”

¹ “[Why telecommunications is a lynchpin between cybersecurity and AI for good](#)”, World Economic Forum, 2024

Going forward, the new success mantra for telcos will be to extract value at each stage with new predictive and generative Artificial Intelligence (AI). According to BCG, “[The most innovative telcos have begun harnessing the power of GenAI.](#)”² By leveraging vast amounts of data generated by networks, customers, and operations, telcos can use advanced analytics and AI to extract valuable insights, enhance network performance, improve customer experiences, and optimize operational efficiency.

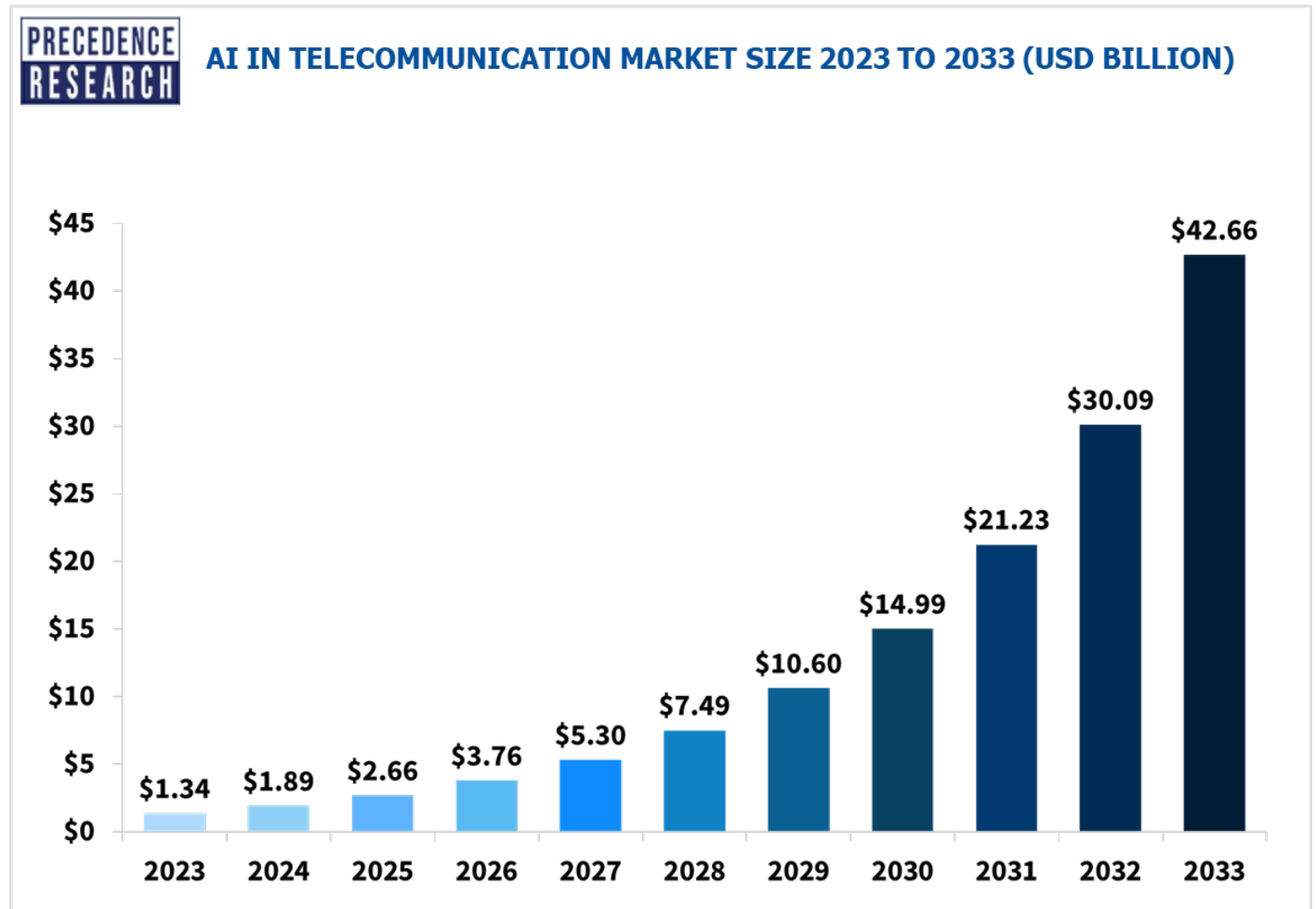


Figure 1- [AI in Telecommunication Market Size 2023 to 2033 \(USD Billion\)](#)³

Transforming the Digital Economy: How AI is Driving Innovation in Telco B2B2X Partnerships

Telcos that successfully build out their own AI infrastructure and effectively leverage AI in their own businesses are in an ideal position to offer AI services and infrastructure to the customers they already serve as a means of unlocking new revenue streams and becoming regional AI providers. They can provide critical infrastructure and data-driven insights to sectors such as healthcare, finance, manufacturing, and retail – fostering collaborative ecosystems. As B2B2X partners, telcos not only drive digital transformation but can also position themselves as indispensable enablers of cross-industry growth and development.

² [New Formula for Success | 2024 Telco Value Creators | BCG](#)

³ [Precedence Research, "AI in Telecommunication Market Size, Share, and Trends 2024 to 2033," 2023](#)

Some examples of these offerings include premium video conferencing services with innovations that can detect gaze, 3D audio call center chatbots featuring real-time translation, and services that improve network security or support infrastructure-as-a-service at the outside edge of the network.

Telco operators are perfectly positioned to expand their data centers with accelerated computing infrastructure to support nations by providing a Sovereign AI Platform that helps their national and economic development. This new class of data centers, known as AI Factories, enhances the AI capabilities of local governments, enterprises, and startups. Telco leaders need to continue investing in and innovating with their AI offerings to build their capabilities with the eventual goal of becoming an AI factory leader. Each new investment builds on prior ones and represents potentially profitable new offerings that allow telcos to participate in industry wide growth.

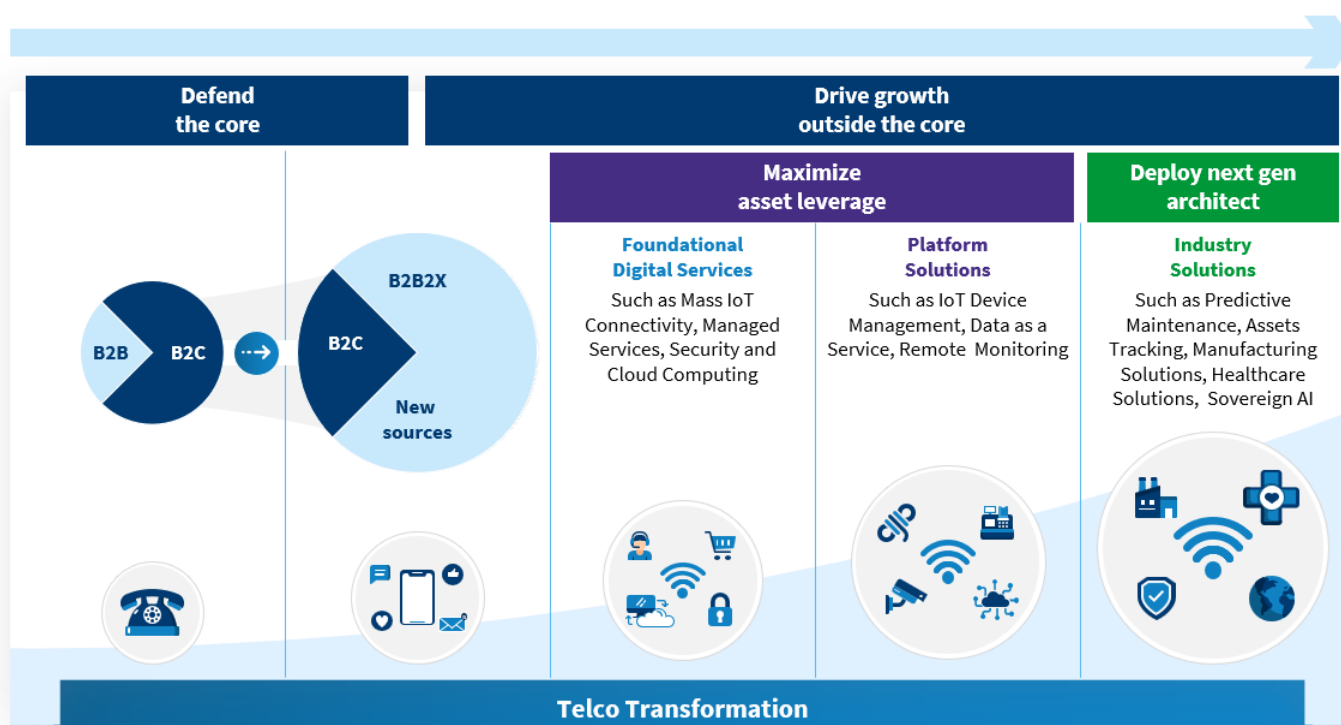


Figure 2- Telco Transformation with Continued Investment

Most Common AI Trends in the Industry

1. Telcos are Swiftly Embracing AI

According to a recent NVIDIA survey, [90 percent of telco executives](#)⁴ report that their organizations are currently engaged with AI from the experimentation stage to full scale deployment. AI leaders in the sector are already seeing impressive increases in customer satisfaction and decreases in costs. The top quintile of companies surveyed by McKinsey experienced a five-year revenue CAGR that is 2.1 times higher than that of peers and a [total return to shareholders that is 2.5 times larger](#).⁵

⁴ [State of AI in Telecommunications. NVIDIA, 2024.](#)

⁵ [The AI-native telco: Radical transformation to thrive in turbulent times. McKinsey, 2023.](#)

Since 2020, there has been an explosion of organizations using telecommunications networks to connect employees and improve access to healthcare, communications, and essential services, all bolstered by high-speed connectivity. AI-fueled opportunities are providing the much-needed momentum to stimulate the industry.

2. Customer Experience Remains a Key Driver of AI Investment

Telcos are leveraging predictive and generative AI to meet various business needs, with the greatest AI opportunity lying in enhancing customer experiences. Telcos are using AI to improve customer engagement through virtual assistance, enriched experiences, and recommendations, all while improving the service experience using predictive data analytics. Additionally, AI-driven customer service tools handle routine inquiries efficiently, reducing wait times and ensuring prompt resolutions. For B2B customers, telcos can become essential partners for building their AI infrastructure through platforms that go beyond traditional connectivity.

3. Telcos are Boosting Network Planning and Operations with AI

AI is revolutionizing how telcos manage their networks. AI enables telcos to interact with their networks in innovative and powerful ways, such as responding to critical issues, identifying incidents within specified timeframes, and recommending effective solutions. This can improve the service experience for customers by providing swift resolutions. For example, predictive AI solutions can detect and remediate (self-heal) network issues before they impact customers, effectively managing service queries in advance and reducing customer churn.

These are just some of the many strategic areas where AI has started showing massive potential. The usage of this technology can best be described by some common use cases which have seen the most significant adoption by a number of major telco organizations. In this whitepaper, we will focus on five common use cases with the greatest potential to significantly impact the telco sector.

Key Use Cases for AI in Telco

AI has the potential to transform organizations' internal operations and create an entirely new suite of product offerings, with key use cases spanning across both areas.

AI in Telco: Key Use Cases for Operational Efficiency

1. Reimagining Contact Centers to Enhance Customer Experiences

AI-enabled chatbots can transform the way contact centers operate to enhance customer experiences. Chatbots handle first-level inquiries and provide data-driven solutions to human representatives, reducing call times, improving customer satisfaction, and increasing productivity.

2. Streamlined Network Operations

Telcos can use AI to streamline costly network maintenance. Using AI saves operations teams time and increases efficiency while reducing time-to-resolution for troubleshooting and boosting performance. Telcos can create self-healing networks that can identify and remediate issues before they impact customers without human intervention and improve service by using AI to optimize spectrum allocation and usage for existing customers.

3. Predictive Analytics for Business Insights

Get next level insights into your customers' use patterns and lifestyles with AI to offer them customized premium products on top of core offerings. Increase customer engagement and satisfaction while also improving profit margins with services beyond connectivity. Extend predictive analytics capabilities to B2B clients by providing other businesses with valuable insights into their own customer behaviors and preferences.

AI in Telco: Key Use Cases for Fueling Growth with New and Expanded Services

1. AI Factories for Sovereign AI Infrastructure

AI Factories are transforming sovereign AI infrastructure development and deployment. AI factories enable telcos to create tailored AI solutions that comply with data sovereignty and privacy regulations. This approach ensures full control over data and AI models and helps foster innovation while meeting national interests and regulatory compliance. Implementing AI Factories for sovereign AI infrastructure marks a significant advancement towards a smarter, more secure, and efficient telecom future.

2. 5G Monetization with Edge AI and 6G Research

Telcos can leverage the AI ecosystem to deliver innovative 5G applications, including intelligent video analytics for autonomous shopping, smart traffic monitoring to reduce urban congestion, and immersive 5G extended reality (XR) experiences. In addition, edge computing can play a significant role in 6G networks, creating significant growth opportunities for the sector such as real-time data analysis and smarter decision-making when deploying AI at the network edge. Telcos can use AI to accelerate 6G research by optimizing network design, enhancing signal processing, and enabling real-time data analysis. Additionally, AI is enhancing the development of innovative applications such as intelligent edge computing and immersive extended reality, pushing the boundaries of next-generation connectivity. AI enables telcos to go beyond research by creating revenue opportunities today with 5G monetization at the edge.

Navigating the Complexities: AI Adoption and Implementation Challenges

As in many industries, fully embracing AI and AI-driven strategies continues to be challenging in the telco sector. Some of the common roadblocks telco organizations face when trying to implement predictive and generative AI solutions include:

Data Privacy and Compliance: Ensuring data privacy and regulatory compliance is crucial and sometimes challenging for telcos that handle sensitive customer information and operate across diverse jurisdictions.

- Telco data often contains sensitive information about customers, including personal identifiers, communication patterns, and location data which can jeopardize customer privacy when not handled correctly.
- Telcos that operate across many regions struggle to create a robust and auditable AI system that meets the data privacy regulations of every jurisdiction their products are used in.
- Telco providers face the challenge of building robust security measures and compliance frameworks when leveraging AI for data analysis and insights.

Telcos must create AI infrastructure for their customers that is reliable and secure to ensure data sovereignty and compliance. Varying regulations across jurisdictions require a secure platform for building AI solutions to provide a trustworthy foundation for telcos.

Workforce Challenges: In a 2023 NVIDIA survey, telco executives identified finding the appropriate [skilled labor with the necessary skills as a top challenge for AI \(34%\) in general and for generative AI \(55%\) more specifically](#).⁶

- Without skilled workers well-versed in AI, organizations struggle to successfully deploy AI solutions and see value. Organizations ranging from large tech companies to small operations compete to hire experienced AI developers. Professionals must be familiar with their existing infrastructure to successfully integrate AI solutions into their operations.
- Often, employees resist adopting newer technologies that require significant upskilling and agile adaptation to rapidly changing systems and processes.

⁶[State of AI in Telco 2024 Report. NVIDIA, 2024.](#)

Organizations need AI-ready infrastructure that is easy to deploy, integrates well into existing workflows, and is easy to operate and manage with tools that are familiar to IT teams.

Scalability and Performance: As telcos strive to manage massive data volumes and meet high customer expectations without disruptions, scaling with continued infrastructure performance can pose a considerable challenge.

- As customer demand grows, telco organizations struggle to handle massive volumes of data and support millions of users simultaneously through extensive data centers and network infrastructure.
- Customers expect high performance and support, so any dips or outages can negatively impact their experience, leading to dissatisfaction, potential churn, and a damaged reputation for the service provider.
- High latency and slow responsiveness can delay AI responses, leading to missed sales opportunities, customer frustration, and operational inefficiencies.

AI solutions in telco environments must be highly scalable and performant, especially during peak usage periods. To innovate rapidly and meet customer expectations, organizations need powerful GPU-enabled infrastructure with high-performance networking purposely built to scale and deploy AI solutions quickly and securely.

Legacy Systems and Infrastructures: Outdated technology and fragmented data hinder effective AI deployment and sustainable growth.

- Legacy systems are often incompatible with modern AI technologies, making integration complex and costly. Compatibility issues can slow down the deployment of AI solutions and limit their effectiveness.
- Legacy infrastructures can create fragmented data environments, leading to data silos. This makes it difficult to consolidate and analyze data comprehensively, hindering the AI's ability to generate accurate insights and predictions.
- Older systems may not have the capacity to handle the increased data processing and storage demands of AI applications, limiting the scalability of AI initiatives and preventing telcos from fully leveraging AI capabilities.

Telcos need modern infrastructure that can streamline the integration of siloed data systems to achieve sustainable growth.

Supermicro with NVIDIA – Helping Telcos Embrace AI to Optimize Operations and Unlock Strategic Growth

Supermicro infrastructure with the NVIDIA AI Enterprise solution empowers the industry with telco optimized AI solutions that can help organizations improve customer experiences, streamline complex network operations, extract business insights through data science – all while unlocking new ways to positively impact their bottom line.

Here's a look at some of the key use cases of AI for the telco sector that Supermicro with NVIDIA helps power for your organization and some of the recent customers that have adopted and implemented AI solutions for their businesses.

Key Use Cases for Operational Efficiency

1. Reimagining Contact Centers to Enhance Customer Experiences

Generative AI chatbots are the first line of interaction for customer service, answering questions about service outages and bill changes and allowing human representatives to solve more pressing and complex inquiries. Chatbots streamline the customer experience as reported by TIDIO, a customer service software company, which found that [62 percent of customers surveyed would rather use a customer service bot than wait for a human representative](#).⁷

Once a request is elevated to a contact center representative, AI-powered solutions review previous customer calls, complete sentiment analysis, and create a list of solutions for the representative to consider offering the customer. This can assist the contact center representative with insights into the likelihood of customer churn and successful solutions to previous calls such as temporary bill decreases. Contact center representatives are empowered with additional context from previous customer touchpoints with AI-powered omnichannel strategies and are no longer playing catch up on customer complaints.

The Intelligent Virtual Assistant AI Workflow

The intelligent virtual assistant AI workflow accelerates building and deploying an end-to-end intelligent virtual assistant solution. The workflow contains ASR and TTS training with NVIDIA NeMo™ and inference with NVIDIA Riva built on NVIDIA Triton™ Inference Server.

Powering Other Industries

Telcos can also offer AI-driven contact center solutions as a product or service to other industries, helping businesses enhance their customer service operations. By providing integrated customer service experiences that span multiple touchpoints, including brick-and-mortar, web, and mobile devices, telcos enable companies to adopt omnichannel strategies that bridge the gap between online and offline interactions. This not only improves customer satisfaction but also increases operational efficiency, making AI-powered contact center solutions an attractive offering for B2B clients.

Supermicro + NVIDIA

Supermicro's GPU-optimized systems, combined with NVIDIA's AI Enterprise solutions, help telco organizations create intelligent chatbots, copilots, and virtual assistants. These tools can access and use important data to improve efficiency and provide a competitive edge. This is done through a method called retrieval-augmented generation (RAG), which links large language models (LLMs) to a company's knowledge base, unlocking many new possibilities.

- Achieve 2335 LLM throughput (total tokens/second per GPU) on Supermicro's [SYS-221HE-FTNR\(D\)](#) with NVIDIA H100 NVL GPU or for even greater throughput utilize [Supermicro's ARS-111GLNHR](#) with the NVIDIA GH200.
- NVIDIA NeMo™ and NVIDIA NIM™ help speed the development and deployment of enterprise-ready GenAI models.
- NVIDIA NeMo™ is an end-to-end, cloud-native framework for curating data, training and customizing foundation models, and running inference at scale. It supports text-to-text, text-to-image, text-to-3D models, and image-to-image generation and speeds the development and deployment of enterprise-ready GenAI models.
- NVIDIA NIM™ facilitates scalable and efficient deployment of GenAI models by serving models as microservices, allowing applications to request inference via API calls.
- NVIDIA Merlin™, an open-source recommendation system powered by machine learning and deep learning algorithms, accelerates the entire pipeline, from ingesting and training to deploying GPU-accelerated recommender systems.

⁷ ["The Future of Chatbots..." TIDIO, 2024.](#)

- For state-of-the-art conversational AI, NVIDIA® Riva delivers accelerated data pipelines, tools for developing high quality virtual assistant apps, real-time transcription, and chatbots.

2. Streamline Network Operations

Managing networks can be one of the most time and labor-intensive undertakings for the operations arm of a telco organization. Just the field and service operations aspect of [network maintenance accounts for 60 to 70 percent of most telco organizations' operating budgets](#).⁸ Not only do maintenance teams need to constantly update and repair aging infrastructure and install new upgrades, they also must make urgent repairs when networks are damaged by unforeseen circumstances such as weather events.

AI, the Internet of Things (IoT), and accelerated computing at the edge are turbocharging network performance and efficiency. AI is proving to be a powerful ally for telcos, responding to operational roadblocks by helping them identify potential network failures, recommending solutions, and helping optimize field service while also continuously monitoring for anomalies and automating the response process for minimal downtime. With AI-powered systems, organizations can enable networks to self-configure, self-optimize, and self-heal without human intervention while optimizing resource management.

Over the past several years, telcos have worked to build up their 5G infrastructure and expand coverage. 5G operations are crucial to maintaining high quality and consistent services and AI can assist in several key areas. AI can boost dynamic network management and help create self-optimizing networks. These networks adjust to real-time conditions and optimize performance, ensuring high-quality service delivery even during peak usage. Predictive analytics performed by AI can forecast network resource demands, enabling better planning and allocation of resources to prevent congestion and ensure optimal performance. With intelligent routing, AI can prioritize routing specific applications and services to reduce congestion and boost overall service.

Supermicro + NVIDIA

Supermicro's AI-ready telco infrastructure with NVIDIA's AI Enterprise Solution can help telcos create AI-enabled solutions, build software-defined and accelerated infrastructure for 5G, and bring connected intelligence to smart devices at the edge. Supermicro's SYS-221HE-FTNR(D) IOT SuperServer combines with NVIDIA's H100 NVL to support telco microdata centers and network function virtualization while Supermicro's GPU SuperServers ARS-111GL-NHR, [SYS-421GE-TNHR2-LCC](#), and [SYS-821GE-TNHR](#) support high performance computing, generative AI, and AI training and inference.

[Supermicro's high performance AI Development Workstations](#) can supercharge your GenAI development process with powerful, dedicated, and secure platform to create and train models. Build a comprehensive generative AI system equipped with LLMs, multimodal, vision, and speech AI with NVIDIA NeMo™ with inflight batching through NVIDIA TensorRT™ to help deliver low-latency and high-performance inference with generative AI solutions that are simpler to deploy.

3. Predictive Analytics for Business Insights

In an industry as competitive as telecommunications, any additional advantage to attract and retain customers can boost profit margins and give organizations an important competitive edge. Using advanced analytics and machine learning, telcos can extract valuable insights to improve network performance, customer experiences, and operational efficiency. With advanced analytics alone, telcos can reduce cloud costs for use cases such as customer churn prediction, predictive maintenance of network equipment, advanced security, fraud detection, and much more. Forward-thinking telco leaders are using this as an opportunity to deliver new services. In a B2C context, this can translate to next-level insights into telco customers to identify upselling opportunities, build brand loyalty, and provide elevated service.

⁸ [How AI is helping revolutionize telco service operations | McKinsey](#)

Powering Other Industries

In a B2B setting, telcos can leverage data to help other businesses optimize their services. For instance, they can predict surges or downtimes, giving operators the time and resources to resolve issues promptly and effectively. Telcos can offer predictive analytics as a service, enabling other industries to harness the power of AI for their own operational improvements. By extending these sophisticated tools and insights to B2B clients, telcos not only diversify their revenue streams but also reinforce their position as innovative leaders capable of driving business success across multiple sectors.

Supermicro + NVIDIA

Supermicro's high performing infrastructure for machine learning/AI training with NVIDIA's generative AI software suite enables quicker time to result for AI and machine learning initiatives while improving cost-effectiveness.

- Choose between Supermicro's GPU SuperServers ARS-111GL-NHR, SYS-421GE-TNHR2-LCC, and SYS-821GE-TNHR to support high performance computing, generative AI, and AI training and inference. Use Supermicro's SYS-221HE-FTNR(D) IOT SuperServer combined with NVIDIA's H100 NVL for 5G edge and core computing.
- The NVIDIA RAPIDS™ Suite of software libraries help accelerate ML processing.

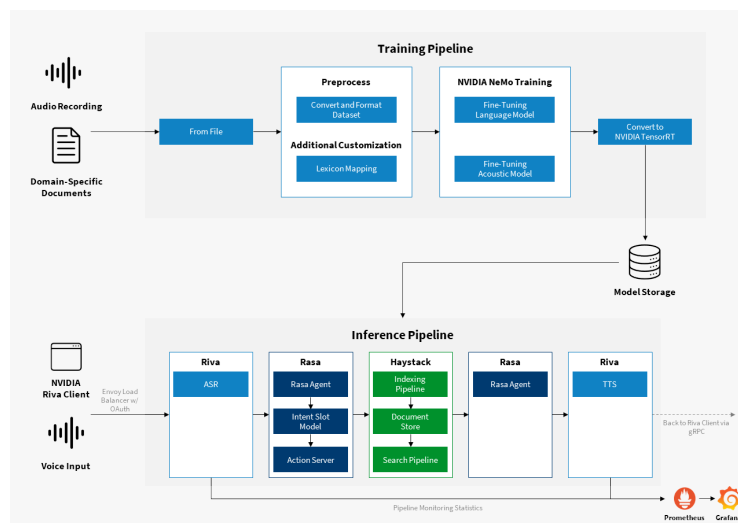


Figure 4- AI Training and Inference pipeline

AI in Telco: Key Use Cases for Fueling Growth with New and Expanded Services

1. AI Infrastructure for Sovereign AI Factories

Telco organizations are already trusted partners for governments, enterprises, and startups, which all depend on telco data centers for secure, high-speed connectivity. By expanding data centers with accelerated computing infrastructure to support AI models, telcos can transform their existing data centers into AI Factories. AI Factories are specifically designed for AI workflows, particularly LLM foundational pre-training, fine-tuning, RAG, and inference, which are the most computationally intense tasks ever created. As AI adoption grows and most apps begin to incorporate GenAI as their new user interface, AI infrastructure will become increasingly necessary across regions. Nations will quickly realize the importance of AI for their national development, economic growth, job creation, and innovation. Telco service providers can leverage their expertise in managing capital, distributed infrastructure, and access to data center space and power. This enables them to provide foundational AI Factory capabilities to the national ecosystem, including government, academia, enterprises, and startups. Supermicro and NVIDIA can help organizations design and create their AI factories by [reading this whitepaper](#).

Revenue Generating Opportunities

Telco companies don't need to limit themselves to providing AI infrastructure as a service. They can also develop and deploy their own tailored AI solutions to address key use cases and challenges for customers across industries. This approach unlocks new revenue streams and facilitates new customer experiences within their geography.

By offering AI as a service to other industries, telco organizations can tap into new markets and diversify their revenue streams, providing AI-driven insights and automation to sectors such as healthcare, finance, and manufacturing. Some of the key AI Factory enabled use cases across these sectors are predictive maintenance for manufacturing, personalized healthcare recommendations, real-time fraud detection in finance, and advanced customer service chatbots for retail. This cross-industry application not only maximizes the return on investment in AI infrastructure but also positions telco operators as leaders in the broader AI ecosystem, driving innovation and growth beyond their traditional market boundaries.

Supermicro + NVIDIA

Telco operators can use their distributed regional data centers (RDCs) as accelerated computing infrastructure to build, fine-tune, and deploy sovereign AI models. Supermicro with NVIDIA delivers one of the broadest selections of NVIDIA-Certified systems providing the best performance and efficiency for organizations to build their own AI capabilities and ultimately deliver these services to others. The solutions ensure the fastest deployment, supportability, and performance leveraging NVIDIA References Architectures on Supermicro AI optimized systems.

- For optimal data center performance, leverage Supermicro SuperServer SYS-421GE-TNHR2-LCC or SYS-821GE-TNHR with NVIDIA HGX H100/H200 delivering up to 15.8k BF16/FP16 TFLOPS and 31.6k INT8 TOPS in a data center cluster.
- NVIDIA NeMo™ and NVIDIA NIM™ help speed the development and deployment of enterprise-ready GenAI models.
- NVIDIA NeMo™ is an end-to-end, cloud-native framework for curating data, training and customizing foundation models, and running inference at scale. It supports text-to-text, text-to-image, and text-to-3D models, and image-to-image generation and speeds the development and deployment of enterprise-ready GenAI models.

2. 5G Monetization with Edge AI and 6G Research

While 5G technology has already been deployed worldwide to meet an ever-expanding need for mobile data services, 6G networks represent the next step in wireless services, integrating AI and ML to create new applications and use cases previously not possible with 5G networks. By partnering with telcos to deploy 5G at the edge, organizations across many industries such as transportation, healthcare, logistics, manufacturing, robotics, smart cities, and retail can visualize and optimize their physical environment and expand their computational power to new places, becoming more connected and efficient. Telco service providers are monetizing 5G at the edge by providing services including enterprise and B2B services, managed services, consumer services, and more.

AI is also critical to the research and development of 6G, as AI-driven simulations and modeling accelerate the testing of new technologies and use cases. Organizations can now accelerate the development of 6G technologies that will one day connect trillions of devices, laying the foundation for a hyper-intelligent world supported by autonomous vehicles, smart spaces, and a wide range of extended reality and immersive education experiences and collaborative robots.

AI-fueled 6G research is unlocking exciting new revenue streams for telcos by enabling cutting-edge services including ultra-reliable low-latency communications and immersive AR/VR experiences. With AI-driven network optimization and predictive maintenance, operational costs are slashed, paving the way for investment in innovative technologies and profitable new services.

Supermicro + NVIDIA

Supermicro's servers for next-gen networks from edge to core to cloud with NVIDIA 6G Research Cloud platform are helping telco organizations rapidly accelerate research. Industry-leading researchers can use all elements of the 6G development research cloud platform to advance their work.

- [Supermicro SYS-E403-13E-FRN2T](#) with NVIDIA L40s, [Supermicro SYS-111E-F\(D\)WTR](#) with NVIDIA L40, Supermicro SYS-221HE-FTNR(D) with NVIDIA H100 NVL, and Supermicro ARS-111GL-NHR with NVIDIA GH200 are all effective options to facilitate 5G optimization and forward-looking 6G research.
- NVIDIA Aerial Omniverse Digital Twin for 6G: A reference application and developer sample that enables physically accurate simulations of complete 6G systems, from a single tower to city scale. It incorporates software-defined RAN and user-equipment simulators, along with realistic terrain and object properties. Using the Omniverse Aerial Digital Twin, researchers will be able to simulate and build base-station algorithms based on site-specific data and to train models in real time to improve transmission efficiency.
- NVIDIA Aerial CUDA-Accelerated RAN: A software-defined, full-RAN stack that offers significant flexibility for researchers to customize, program and test 6G networks in real time.
- NVIDIA Sionna Neural Radio Framework: A framework that provides seamless integration with popular frameworks such as PyTorch and TensorFlow, leveraging NVIDIA GPUs for generating and capturing data and training AI and machine learning models at scale. This also includes NVIDIA Sionna, the leading link-level research tool for AI/ML-based wireless simulations.

Supermicro and 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, networking, and the NVIDIA AI Enterprise software platform, all orchestrated within the resilient infrastructure of Supermicro's platforms.

From large-scale training to intelligent edge inferencing, Supermicro's AI-optimized solutions streamline and accelerate AI deployment. They help the telecommunication sector empower AI workloads with optimal performance and scalability while optimizing costs and minimizing environmental impact.

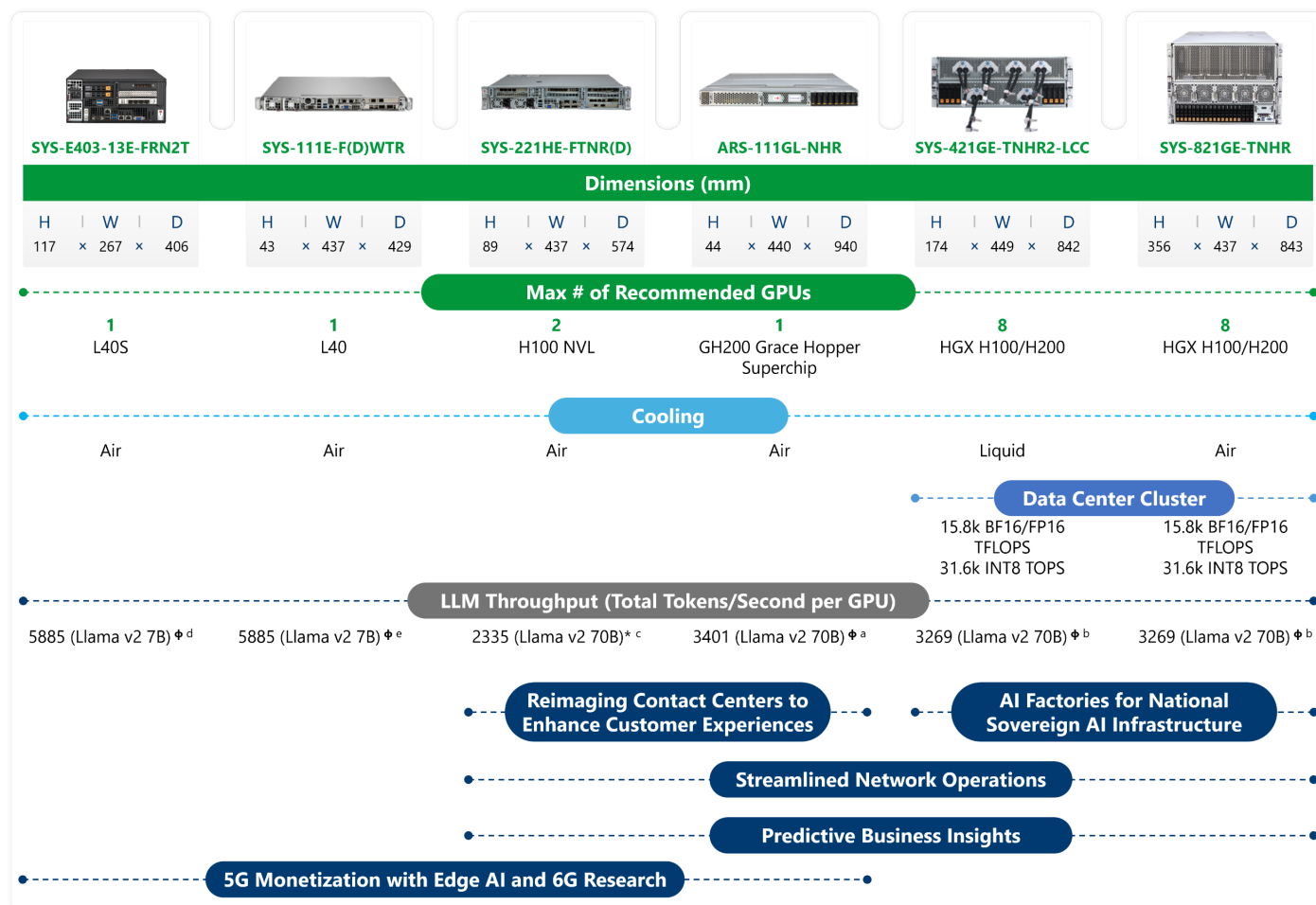
Supermicro and NVIDIA can help telco customers with a diverse range of AI workload-optimized solutions while empowering them to be a powerful ally to other industries that rely on telcos for their operations and connectivity needs. While there are a variety of systems that organizations can choose from based on their unique environments and needs, we provided a brief set of recommendations in the product table on the following page.

Selecting the Optimal Supermicro Systems for Your AI Applications

Supermicro's flexible range of AI-ready infrastructure solutions ensures that telco operators implementing AI solutions can scale their implementation as much as needed. Whatever the requirements, Supermicro has solutions to expand memory, processing power, and storage to meet any situation.

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 the large-scale video analytics of a smart store or processing LLMs in building chatbots, Supermicro's portfolio of edge platforms is available across a wide range of form factors, including a compact box, 1U and 2U, wall mount, rack mount, and even fan-less models. When combined with NVIDIA's powerful AI and accelerated computing platforms, they 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 different use cases in telco and other industries. Their systems are adaptable to various computational demands.



*: Input Length = 128, Output Length = 128, Batch Size = 256
 Φ : Input Length = 128, Output Length = 128, Batch Size = max
a: TP = 1, with 1x GH100
b: TP = 2, with 2x H100
c: TP = 1, with 1x H100 NVL
d: TP = 1, with 1x L40S
e: TP = 1, with 1x L40
TP = Tensor Parallelism

Figure 5- 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. NVIDIA's high-performance networking platforms, alongside telco ecosystem partners, such as Arrcus, bring software-defined, hardware accelerated carrier grade routing, switching and telemetry to every Supermicro server as shown in Figure 6. This 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 telcos to unlock the full potential of AI. This architecture is shown in Figure 6 below.

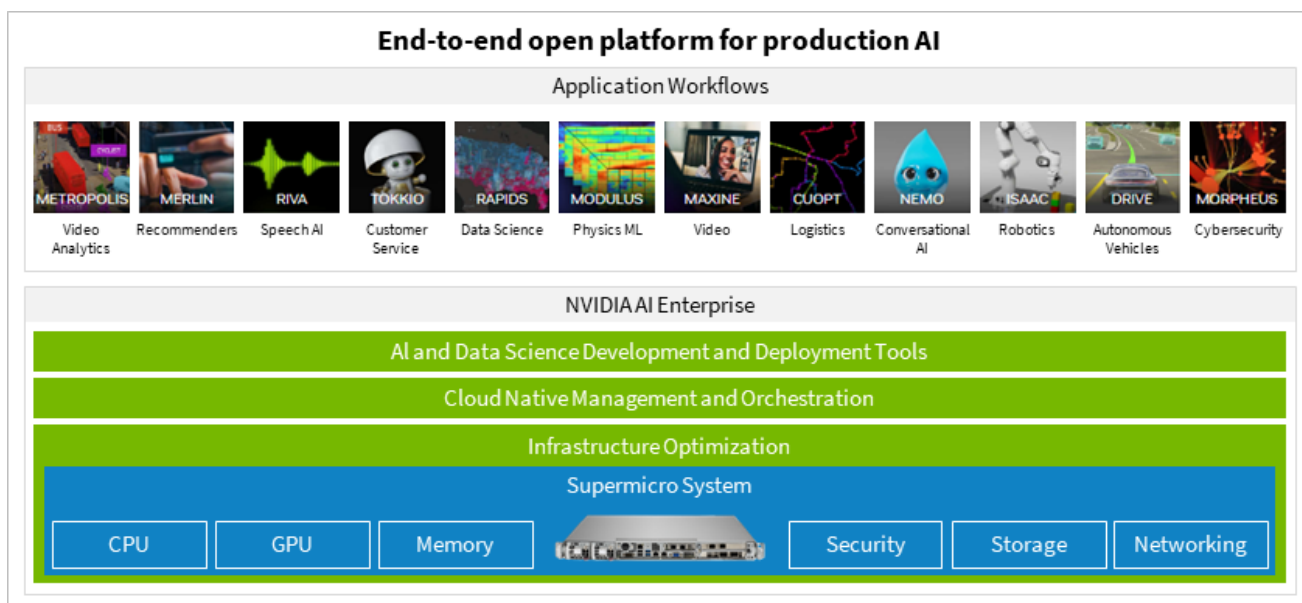


Figure 6 – The Comprehensive Solution

Moving Forward

The exploration and adoption of AI is already underway in the telco industry, and those who most successfully implement it into their workflows are poised to generate the most value. Organizations 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 diverse AI initiatives. 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 about our AI solutions, visit <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>.