AI in Telco – Unlock Growth Opportunities

Opportunity of AI in Telco

- The global AI in telecommunication market size is projected to hit around **USD 42.66 billion by 2033**[1], growing at a CAGR of 41.40% from 2024 to 2033.
- According to a recent NVIDIA survey, **90 percent of telco executives**[2] report that their organizations are currently engaged with AI from the experimentation stage to full scale deployment.

Common Pain Points of Adopting AI for The Telco Industry

- **Data Privacy and Compliance:** Telco data often contains sensitive information about customers, including personal identifiers, communication patterns, and location data which can jeopardize customer safety when not handled correctly.
- **Work Force Challenges:** Large tech organizations struggle to hire experienced AI developers since innovative AI solutions require a skilled workforce to generate value. The availability of skilled talent remains a challenge.

- **Scalability and Performance:** Telco organizations must handle massive volumes of data and support millions of users simultaneously through their extensive data centers and network infrastructure, mandating the need for highly scalable and performant AI solutions—especially during peak usage periods.
- **Legacy Systems and Infrastructures:** Legacy systems often lack compatibility with modern AI technologies, making integration complex and costly. This can slow down the deployment of AI solutions and limit their effectiveness.

Supermicro with NVIDIA Empowering Your AI Journey

Supermicro and NVIDIA® deliver best-in-class outcomes for Predictive and Generative AI implementations in the telco sector. This collaborative approach provides a comprehensive framework that integrates CPUs, GPUs, and optimized memory, all orchestrated within the resilient infrastructure of Supermicro's platforms. These AI solutions serve applications across multiple use cases in the telecommunication industry. Some of the common use cases you can adopt today are:

Key Use Cases of AI in Telco

Key Use Cases for Operational Efficiency

	Relevant Supermicro Systems
1. **Reimagining Contact Centers to Enhance Customer Experiences:** AI-enabled chatbots can transform contact centers by handling first-level inquiries and providing data-driven solutions to human representatives, reducing call times, improving customer satisfaction, and increasing productivity.	Supermicro SYS-221HE-FTNR(D) with NVIDIA H100 NVL GPU and Supermicro ARS-111GLNHR with NVIDIA GH200 Grace Hopper Superchip
2. **Streamlined Network Operations:** Use AI to streamline network maintenance, saving time, increasing efficiency, reducing troubleshooting time, and boosting performance. Create self-healing networks that fix issues automatically. Improve service by optimizing spectrum allocation and usage.	Supermicro SYS-221HE-FTNR(D) with NVIDIA's H100 NVL GPU, Supermicro ARS-111GL-NHR with NVIDIA GH200 Grace Hopper Superchip, Supermicro SYS-421GE-TNHR2-LCC with NVIDIA HGX H100/H200 GPU, and Supermicro SYS-821GE-TNHR with NVIDIA HGX H100/H200 GPU
3. **Predictive Analytics for Business Insights:** Use AI to gain insights into customer patterns and lifestyles to offer customized products, increasing engagement, satisfaction, and profit margins. Extend predictive analytics to B2B clients, providing insights into their customer behaviors and preferences.	Supermicro SYS-221HE-FTN (D) with NVIDIA's H100 NVL GPU, Supermicro ARS-111GL-NHR with NVIDIA GH200 Grace Hopper Superchip, Supermicro SYS-421GE-TNHR2-LCC with NVIDIA HGX H100/H200 GPU, and Supermicro SYS-821GE-TNHR with NVIDIA HGX H100/H200 GPU

Key Use Cases for Fueling Growth with New and Expanded Services

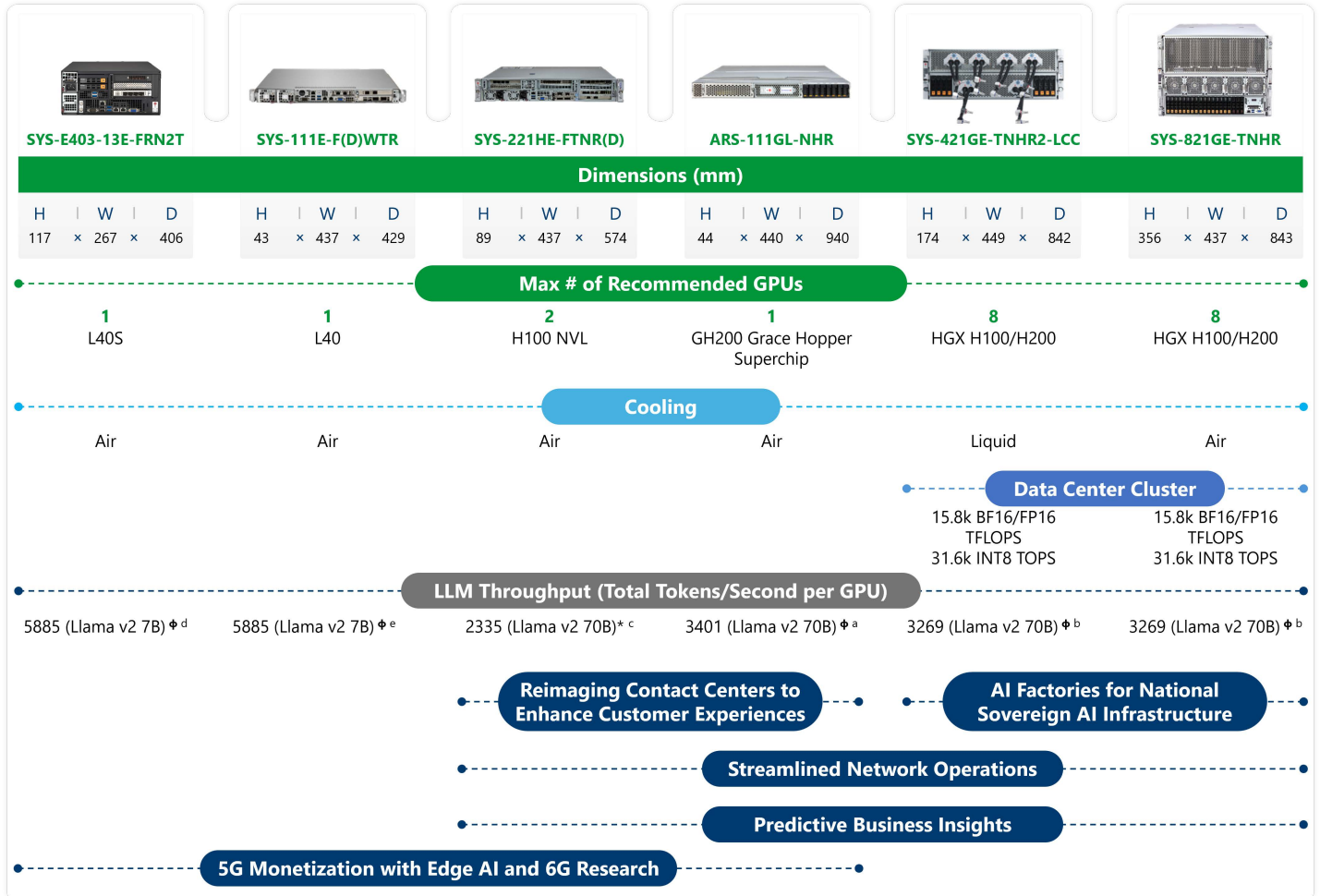|	
1. **AI Factories for National Sovereign AI infrastructure:** Build secure AI infrastructure for customers to innovate on through improvements to existing data centers. Telco organizations are perfectly positioned to expand their data centers with accelerated computing infrastructure to support sovereign AI models. This new class of data centers, known as AI factories, enhance the AI capabilities of local governments, enterprises, and startups. It allows nations to utilize telcos to host their sovereign AI infrastructure for enhanced security and performance.	SYS-421GE-TNHR2-LCC with NVIDIA HGX H100/H200 GPU and Supermicro SYS-821GE-TNHR with NVIDIA HGX H100/H200 GPU
2. **5G Monetization with Edge AI and 6G Research:** Accelerate 6G research by optimizing network design, enhancing signal processing, and enabling real-time data analysis powered by AI. AI is enhancing the development of innovative applications such as intelligent edge computing and immersive extended reality, pushing the boundaries of next-generation connectivity.	Supermicro SYS-E403-13E-FRN2T with NVIDIA L40s GPU, Supermicro SYS-111E-F(D)WTR with NVIDIA L40 GPU, Supermicro SYS-221HE-FTNR(D) with NVIDIA H100 NVL GPU, and Supermicro ARS-111GL-NHR with NVIDIA GH200 Grace Hopper Superchip

[1]AI in Telecommunication Market Size, Growth, Report 2033 (precedenceresearch.com) | [2]https://resources.nvidia.com/en-us-ai-in-telco/state-of-ai-in-telco-2024-report?ncid=no-ncid

August 2024

Supermicro Solutions for Powering Telco AI

	SYS-E403-13E-FRN2T	SYS-111E-F(D)WTR	SYS-221HE-FTNR(D)	ARS-111GL-NHR	SYS-421GE-TNHR2-LCC	SYS-821GE-TNHR

Dimensions (mm)

H	W	D	H	W	D	H	W	D	H	W	D	H	W	D	H	W	D
117	× 267	× 406	43	× 437	× 429	89	× 437	× 574	44	× 440	× 940	174	× 449	× 842	356	× 437	× 843

Max # of Recommended GPUs

1	1	2	1	8	8
L40S	L40	H100 NVL	GH200 Grace Hopper Superchip	HGX H100/H200	HGX H100/H200

Cooling

Air	Air	Air	Air	Liquid	Air

Data Center Cluster

				15.8k BF16/FP16 TFLOPS 31.6k INT8 TOPS	15.8k BF16/FP16 TFLOPS 31.6k INT8 TOPS

LLM Throughput (Total Tokens/Second per GPU)

5885 (Llama v2 7B) ϕ d	5885 (Llama v2 7B) ϕ e	2335 (Llama v2 70B)* c	3401 (Llama v2 70B) ϕ a	3269 (Llama v2 70B) ϕ b	3269 (Llama v2 70B) ϕ b

Reimaging Contact Centers to Enhance Customer Experiences

AI Factories for National Sovereign AI Infrastructure

Streamlined Network Operations

Predictive Business Insights

5G Monetization with Edge AI and 6G Research

*: 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

Supermicro Systems Accelerating Your AI Journey

Supermicro's cutting-edge AI-ready infrastructure solutions enable **large-scale training to intelligent edge** inferencing allowing telcos to streamline and accelerate AI deployment. Their AI-infrastructure empower workloads with **optimal performance and scalability** while **optimizing costs and minimizing environmental impact.**

Supermicro's flexible range of solutions ensures that telcos implementing AI solutions can scale up their implementation as much as needed. Whatever the requirements, solutions are available to expand memory, processing power, and storage to meet any situation.

Selecting the Optimal Supermicro Systems for your AI Applications

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 video analytics of a smart store or processing LLMs in building chatbots.

For more information, visit:
https://www.supermicro.com/en/solutions/ai-deep-learning

SUPERMICRO