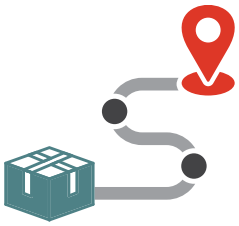
Free markets. Real solutions.

EXPLAINER

Securing the Future of AI at the Edge: An Overview of AI Compute Security

July 2024



Projections indicate that AI-enabled PC and GenAI smartphone shipments could reach **295 million** by the end of 2024—a tenfold increase from 29 million in 2023.

Introduction

In the global AI arms race, generative AI (GenAI) has significantly boosted analysis, speed, and capabilities across several sectors, including cybersecurity. Beyond these benefits and capabilities, AI continues to rapidly advance, with ongoing efforts focused on its integration into edge devices like smartphones and laptops. In fact, projections indicate that AI-enabled PC and GenAI smartphone shipments could reach 295 million by the end of 2024—a tenfold increase from 29 million in 2023. Understanding AI compute security's development and benefits, then, becomes important in protecting these devices from evolving security threats.

Background

Historically, AI—particularly specialized branches like GenAI—has relied on massive data centers for the storage, infrastructure, and computational power needed to train and deploy large models. This centralized approach presents challenges like high energy consumption, scalability issues, and latency that hinder real-time processing. Edge computing, which offers the necessary infrastructure for processing data near its source, is seen as a solution to these issues. Companies like Qualcomm, with its next-generation Snapdragon processor, and Google, with its Pixel 8 powered by Gemini Nano, are leading this transition with hardware and software advancements that bring AI capabilities closer to users. These advancements in edge AI enhance the reliability and speed of applications that require real-time response. However, this shift also has the potential to amplify existing cybersecurity vulnerabilities and introduce new risks.

AI compute security—the measures and practices incorporated to protect the infrastructure, data, and integrity of AI systems within edge devices—is crucial because it maintains the resilience of interconnected systems like edge devices, AI systems, and cloud networks. It currently encompasses a range of traditional cybersecurity measures and AI-specific security solutions including encryption, robust access controls, adversarial training, and more.

Benefits of AI Compute Security

When integrated with edge computing and edge AI, AI compute security represents a paradigm shift in AI technology and cybersecurity risk management. To harness the benefits of this paradigm shift, it is essential to understand the distinct roles and relationships between these components. Together, they establish a strong foundation for advanced AI technology, delivering three key benefits that illustrate the pivotal role of AI compute security in enabling secure and efficient AI operations at the edge.

- Strengthened Data Integrity and Privacy: Localized processing reduces the risk of unauthorized access and breaches. Techniques like secure boot processes, hardware-based security, and advanced encryption methods enhance data protection.
- Enhanced System Reliability and Resilience: Decentralizing AI computations allows continuous operation even during cyber threats, ensuring system functionality and service availability.
- Optimized Resource Allocation and Availability: Efficient security protocols minimize overhead, maintain performance, and reduce dependency on constant, high-speed internet connectivity, leading to energy savings.

Free markets. Real solutions.

To secure the future of AI at the edge, we must leverage AI compute security solutions to protect today's devices and networks.

EXPLAINER

Securing the Future of AI at the Edge: An Overview of AI Compute Security

July 2024

Security Considerations Across Layers of the Edge AI Infrastructure Stack

While decentralization can enhance defenses by reducing reliance on a single endpoint, it can also broaden the attack surface, increasing risks like physical tampering and side-channel attacks on locally processed data. Conventional cybersecurity threats are now compounded by the security risks associated with AI, cloud computing, the internet of things (IoT), and edge computing. This creates a multifaceted challenge that requires a layered approach to manage the interconnected nature of the security challenges of AI-enabled edge devices across infrastructure layers.

Edge Device Layer: Includes physical devices like IoT devices and embedded AI chips. Key security risks include tampering, data breaches, and cyber hijacking. AI compute security solutions involve role-based access control, isolation techniques like hypervisors and virtual machines, and ensuring an attested runtime.

Network Layer: Connects edge devices with local servers and cloud data centers. Security threats include Man-in-the-Middle attacks, data interception, and Distributed Denial of Service attacks. AI compute security measures include encryption, behavioral analytics, anomaly detection, and network segmentation.

AI Compute Layer: Includes edge servers and nodes with AI capabilities. Security considerations include unauthorized access, model poisoning, and adversarial attacks. AI compute security techniques include rigorous data validation and sanitation, adversarial training, and embedding context within machine learning models.

Conclusion

To secure the future of AI at the edge, we must leverage AI compute security solutions to protect today's devices and networks. This effort will become even more critical as the number of AI-enabled edge devices grows. Policymakers can contribute to this effort by supporting secure-by-design and secure-by-default principles to encourage the development of resilient AI ecosystems and by investing in collaborative research initiatives between government, industry, and academia to develop comprehensive AI compute security frameworks. By deepening our understanding of the dual-natured impact of edge AI on cybersecurity, we will be better prepared to harness the benefits of the next AI frontier.

Learn More

Exploring the Next AI Frontier: Understanding AI Compute Security in Edge Devices

Preparing for the Next AI Frontier: Leveraging AI Compute Security to Counter Threats in Edge Devices

Contact us

For more information, please contact:

Haiman Wong
Resident Research Fellow
Cybersecurity and Emerging Threats
hwong@rstreet.org