Forgepoint

2025 Q3 Report

/ Forgepoint Forward /

What's ahead in AI at the Edge



Table of Contents

2 Forgepoint Forward

3 The State of AI at the Edge

5 Challenges

6 Opportunity for Edge AI Startups

8 Trends

10 What Executives Want: Areas for Opportunity

13 Final Thoughts on the Future of Edge AI

14 Acknowledgements

16 Endnotes

> *My biggest problem with IoT and Edge- the assets are harder to manage. Getting firmware updates and maintaining security against attackers across these devices can get very difficult, because they're placed in remote locations. There are network protections on the assets, but when AI comes into play, it grows into a larger concern."*



Spencer Wilcox

Executive Director of Cybersecurity,
NextEra Energy

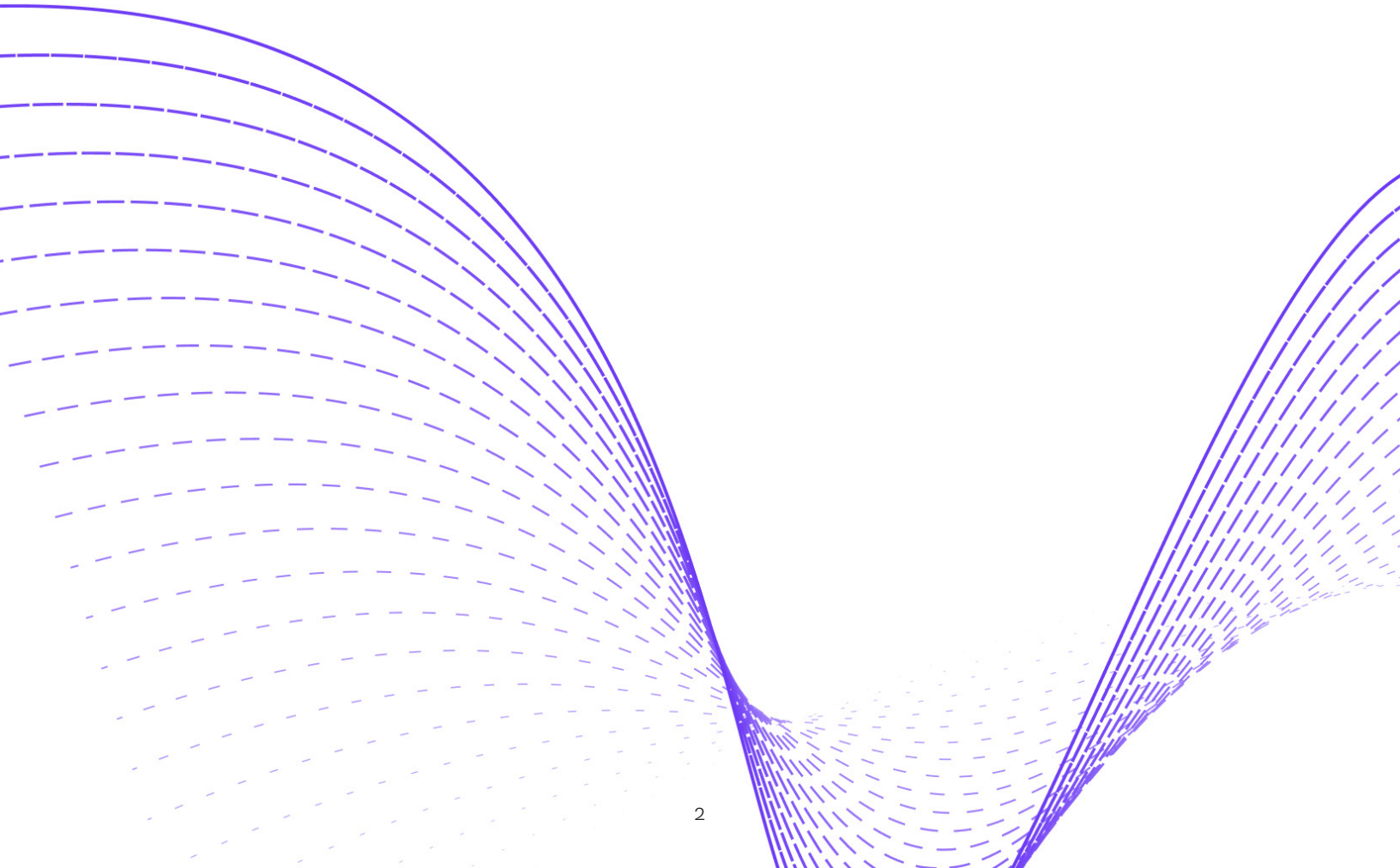
Forgepoint Forward

In today's rapidly evolving digital landscape, the role of cybersecurity has never been more critical. Modern companies navigate a complex threat environment, often relying upon big tech partners and legacy solutions, but still face critical security gaps. These gaps represent risks and opportunities – the linchpins of cybersecurity innovation, software development, and economic progress.

As longtime investors, operators, and company-builders with decades of experience in cybersecurity, the Forgepoint Capital team has the privilege of collaborating with a deep network of CISOs, CIOs, CEOs, industry experts, and national security leaders. These individuals are at the forefront of defending against cyber threats, implementing robust security postures, and fostering organizational and industry resilience.

What follows is the fourth and final edition of Forgepoint Forward, a series of quarterly reports on the most critical emerging spaces in cybersecurity, artificial intelligence, and infrastructure software. Forgepoint Forward presents our findings from extensive research and interviews with experts across our network-including our Global Advisory Council. Our goal is to highlight investment trends and market projections, as well as startups involved in promising areas, to identify key opportunities for entrepreneurs and technology leaders across the cybersecurity community.

We would like to extend our thanks to everyone in our community who shared their insights and contributed to this report. A full list of contributors can be found at the end of this report under Acknowledgements.



The State of AI at the Edge

The "edge" represents a monumental shift in how data is processed, managed, and distributed. As enterprises increasingly utilize applications that rely on real-time data processes, such as connected vehicles or telehealth services, a need for reduced latency emerges. In response, they must rapidly scale to efficiently handle soaring amounts of generated data.

Computing at the edge differs from traditional cloud computing in that it processes data faster and closer to the source or device generating the data, rather than in centralized data centers. Applications can then avoid delays associated with sending data to and receiving data from centralized cloud locations. Edge computing also has lower network bandwidth requirements and enhances privacy by eliminating the need for data to be transmitted to external servers.

In the past, computing power, storage, and connectivity requirements constrained applications of artificial intelligence. Recent advancements in CPU density and declining power requirements now allow more computing power to be pushed to endpoints. Now, AI can effectively operate on devices that were previously too resource-constrained for complex tasks, making edge computing more promising than ever; according to STL Partners, edge AI addressable revenue is expected to reach $157 billion by 2030.[1] Edge AI has already begun revolutionizing various industries:[2]

- Smart sensors can perform on-device data analysis to predict equipment failures and minimize downtime
- Energy distribution grids leverage edge computing to improve efficiency and reliability
- Manufacturing quality control uses edge AI to inspect products on the production line and quickly detect defects
- Automated vehicles analyze data from cameras and sensors at the edge to make safer real-time driving decisions
- Wearable health devices use edge AI to deliver personalized user experiences

As organizations seek to capitalize on the benefits of edge computing and edge AI, executives must determine whether near edge or far edge best suits their applications, taking into consideration:

1) Whether centralized computing or edge computing is best suited for specific workloads

2) An understanding of the required level of flexibility and scalability

3) The need for robust, secure communication across decentralized environments – identifying encryption and authorization protocols that safeguard against physical device tampering and cyberattacks

This report explores key trends driving AI adoption at the edge, the challenges companies face when migrating across the near and far edge, emerging edge AI startups, and the capabilities executives want to see in new tools.



Near Edge vs. Far Edge

"Near edge" infrastructure refers to cases in which enterprises integrate edge capabilities with traditional cloud computing. For example, lightweight distributions designed for the edge enable hybrid setups wherein heavy AI model training happens in centralized cloud clusters, but inference (real-time prediction) happens at edge nodes managed under the same Kubernetes fabric.[3] Kubernetes can scale up or scale down AI inference services locally at the edge based on real-time demand, albeit within the limit of the available hardware at the given location. This allows rolling updates across the edge fleet without downtime and the deployment of different versions of AI models to different nodes—a particularly useful capability for A/B testing.

While Kafka is typically used for high throughput data streaming, it can also be deployed at the near edge to allow quick data movement between edge devices and analytic engines. Though it is a powerful tool familiar to many developers, Kafka is often too heavyweight for the far edge and managing it at the near edge using Kubernetes can become complicated. MQTT operates similarly but is more lightweight and better for small, resource-constrained devices that can't handle big network loads or heavy software. NATS is becoming increasingly popular as a high-performance, lightweight, open-source, and low-latency messaging system with inbuilt security features. Lastly, custom protocols can fill gaps where standard protocols don't fit; for example, in locations without reliable connectivity. Custom protocols can be great for low-power, real-time AI needs while being highly secure and efficient.

Full-service cloud vendor solutions such as AWS SQS or Azure IoT Hub are also valuable tools for enterprises that want to take advantage of near-edge computing. These solutions offer easily integrated services that can be quickly scaled to millions of devices, enabling enterprises to take advantage of existing data centers around the world. While these solutions can be very convenient, they come with certain tradeoffs including vendor lock-in, limited customization, and high costs at scale.

> *"We were one of Google's larger Apigee customers and spending tens of millions annually on the API gateway. It was a necessary evil—variable costs made it harder to deliver high-quality products. We paused innovation for two quarters to build our own solution. Time to market, cost, and the need to diversify node placement drove that decision—we couldn't rely on a single cloud provider...We later shifted to a startup built on Cloudflare, where serverless architecture made deployment far easier. Unlike Kubernetes, you don't have to worry about who runs and monitors it. Our thesis was that inferencing closer to the customer would drive revenue and improve experience."*
>
> 
>
> Chris Sharp
>
> Managing Director of Product, JPMorgan Chase

"Far-edge" refers to cases in which computing is pushed further out to edge devices, often with limited or intermittent connectivity. When AI is deployed at this level, it typically takes the form of tiny, localized AI models making real-time decisions without external help, requiring ultra-low latency (milliseconds) and ultra-low power.

Challenges

1) The edge introduces a limit on processing capacity

Due to edge devices' smaller size and energy efficiency, enterprises face a tradeoff between greater processing power and reduced latency. Sensors, smartphones, and IoT devices have much less processing power than data centers. Edge devices that are battery-powered and operate in environments with intermittent connectivity must be particularly optimized for low energy consumption.

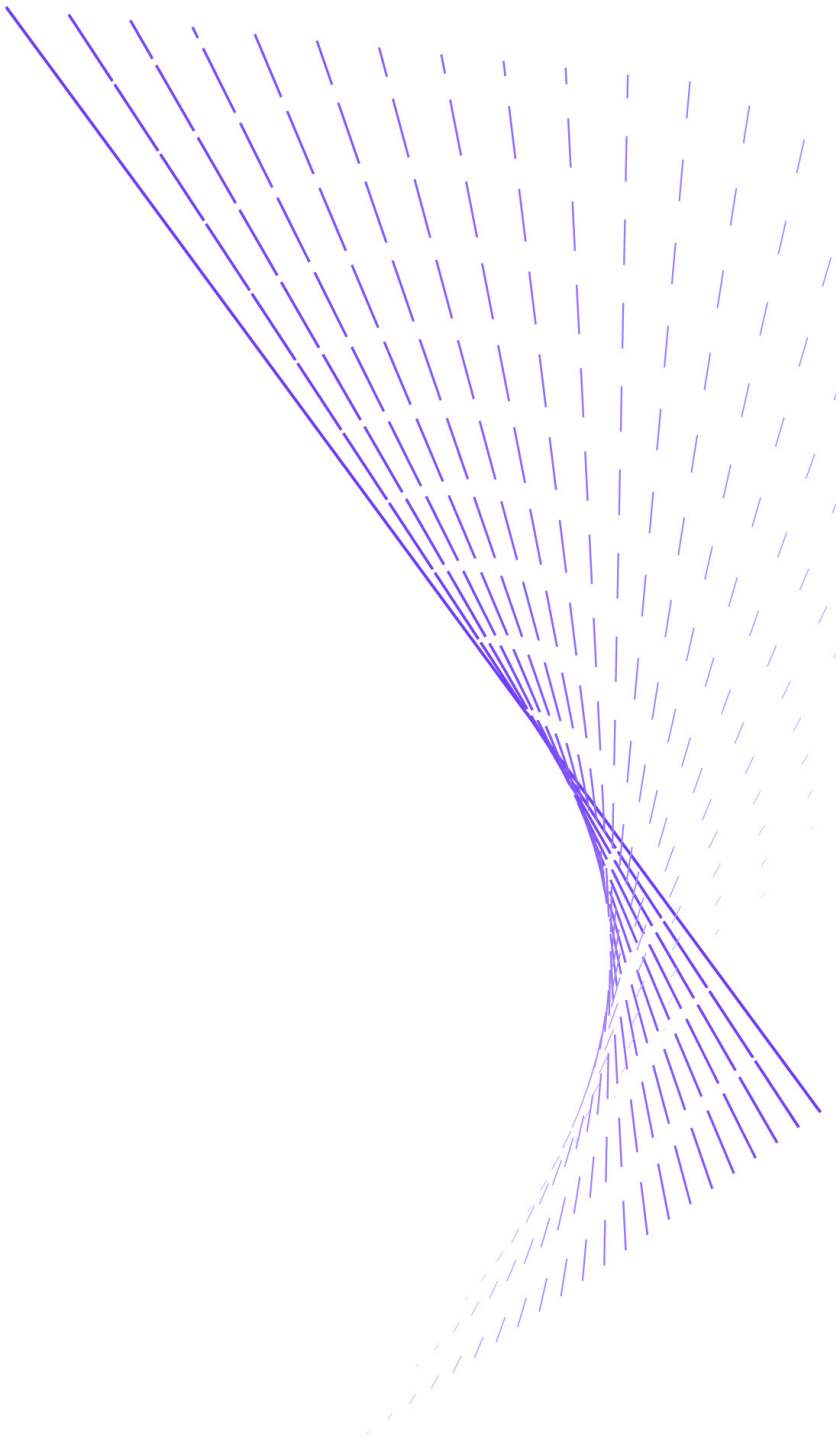
CXOs must evaluate AI use cases to determine whether to prioritize lower latency and real-time decision making at the edge or more advanced AI models, data storage, and processing power with data centers.

2) Edge devices may have difficulty handling sophisticated AI models due to hardware diversity and software fragmentation

Some AI applications require an extensive volume and diversity of data. Large, computationally intensive deep learning models may not be able to be deployed at the edge due to hardware requirements. Models that are successfully deployed at the edge must be carefully pruned and optimized to be lightweight enough to run on edge hardware without compromising accuracy.

Edge AI must be uniformly deployed across a wide range of devices, each with different Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), CPUs and memory configurations. Edge devices capable of running AI workloads also require more powerful and expensive chips, which can be a barrier in cost-sensitive industries.

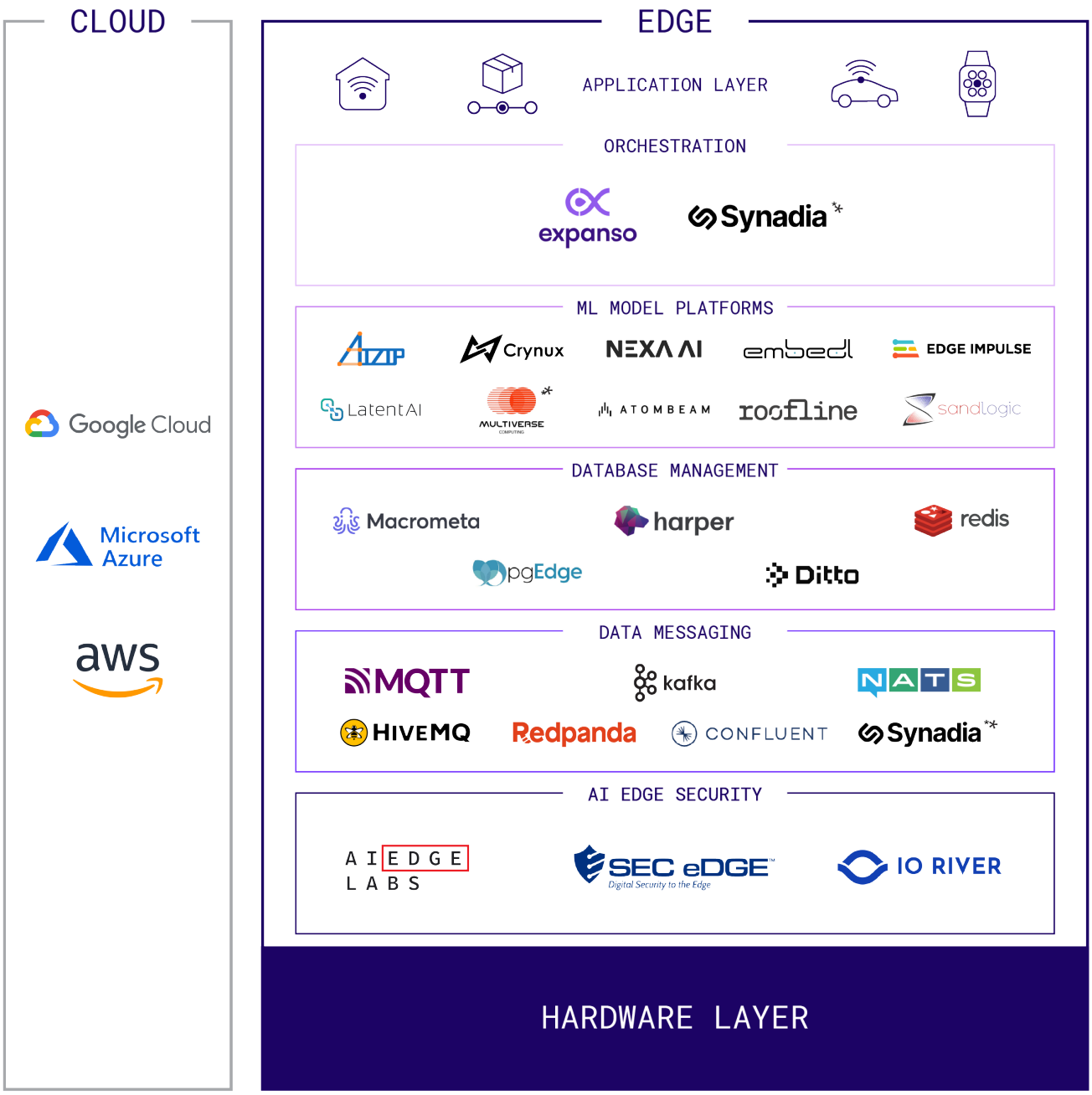
Edge platform software may also be unstandardized, with different OSs and frameworks. The heterogeneity of edge device hardware and software complicates AI model deployment and maintenance.



Opportunity for Edge AI Startups

Innovative startups have emerged to address challenges around edge computing and deploying AI applications at the edge. The solutions they develop cover a range of edge AI capabilities from managing and processing data across distributed environments, to tailoring AI applications to devices, to enhancing security at the edge. There is a significant opportunity for this emerging market to make the edge more accessible for enterprises across industries.

/AI At the Edge/ Market



Forgepoint

This market map is meant to be representative of the AI at the Edge market and is not exhaustive.

** = Forgepoint portfolio company*

We have categorized edge AI startups based on how they enable AI at the edge, with a particular focus on universal rather than vertical solutions. Notably, many vendors in the space offer point solutions focused on layers between hardware and applications.

Here are the growing edge AI market segments that we are particularly interested in:

1) AI Edge Security companies secure edge AI deployments by protecting AI models and data on distributed and vulnerable endpoint devices. They often help secure model updates, implement access controls, and conduct anomaly detection at the edge. For example, AI EdgeLabs blends security with edge intelligence to offer threat detection and response capabilities.

2) Data Messaging systems include widely adopted protocols such as MQTT, Kafka, and NATS, and determine how data is transferred between edge devices, sensors, and applications. These capabilities are critical for systems that require lightweight, fast communication, common for IoT communications. There are various Kafka and MQTT brokers including HiveMQ and Redpanda, while NATS is managed and maintained by Synadia.

3) Database Management companies offer local databases designed to handle storage at the edge for quick data retrieval close to the device. Some of these startups offer edge-native global databases that replicate, distribute and synchronize data at the edge (Macrometa, pgEdge, HarperDB). Ditto, on the other hand, focuses on peer-to-peer local synchronization between edge devices without using a central cloud, while Redis is an in-memory database adapted for the edge.

4) Machine Learning (ML) Model Platforms deliver toolkits to help companies design, train, compress, and deploy Small Language Models (SLMs)-lightweight models trained on focused datasets and designed to excel at specific tasks with lower latency.[4]

Some startups offer a full stack platform to build and deploy models on edge devices (EdgeImpulse, Aizip) while others focus on model compression and optimization functionalities (Multiverse Computing, LatentAI, Roofline).

5) Orchestration providers manage AI workload deployment, scaling, and coordination across edge devices. Synadia, creators and maintainers of NATS, also offer comprehensive management of NATS clusters, a streaming engine, and inbuilt authentication features for security.

With enterprise spend on edge computing expected to grow to $378 billion by 2028, according to the IDC, we see potential for these startups to capture a significant share of an immense market.[5]



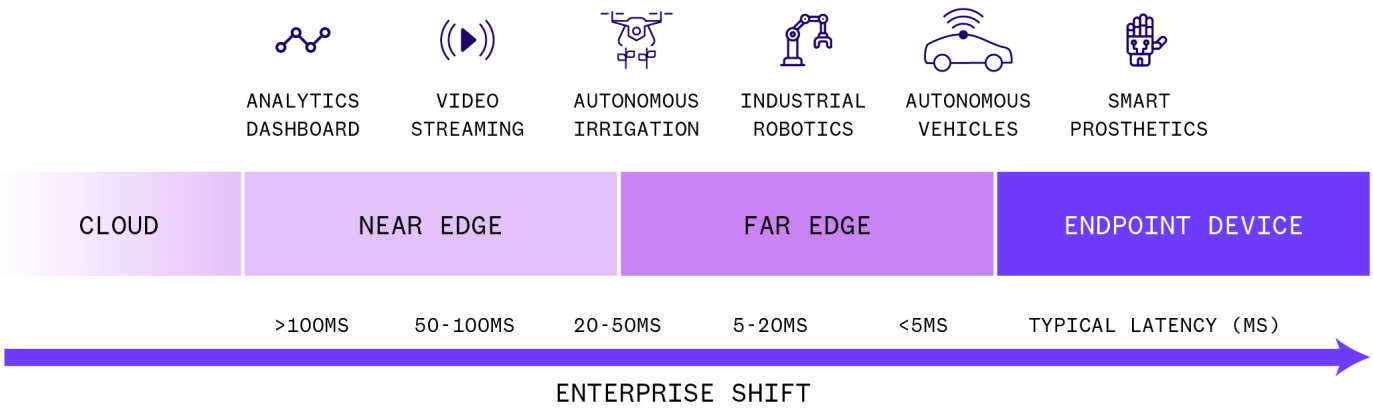
Trends

1) Demand for lower latency and autonomy is driving near edge applications to the far edge

Most if not all applications can be improved with more speed. A core driver for edge technology is a need for ultra-low latency. Lower response times (sub-100ms at the near edge and as low as sub-5ms at the far edge) are essential for complex real-time control systems (self-driving cars) and real-time handling and user (re)display of device inputs (speech to text, camera filters/image correction). In certain specialized cases like industrial robotics, jitter- the consistency of latency- is also critical because predictable and consistent response times are needed to ensure safe and accurate machine actions. Jitter tends to be lower at the far edge than the near edge.

As applications built at the edge diversify and multiply, they become more complex, requiring lower latency and flexibility, interoperability, and inbuilt security. To meet these heightened requirements, enterprises with distributed architectures are beginning to move to the far edge rather than simply collecting and preprocessing data at the near edge. More computational power and storage capabilities are placed where the data is being generated or consumed, enabling decision making to happen in milliseconds. This capability is fueling unique edge applications for contexts including robotic surgery and patient monitoring systems.

On the other hand, there are inherent difficulties in lifting and shifting applications further on the edge. As such, for applications where latency isn't as crucial or exclusive autonomy is not needed, companies may stick with Kubernetes-enabled near-edge solutions.



2) The adoption of continual learning is fueling an emerging market segment

As AI applications are deployed at the edge, enterprises must ensure they are maintaining performance, accuracy and reliability. Models benefit from being able to adapt according to user interactions and novel real-world conditions. There is a persistent need for data orchestration that enables models to be continuously monitored and updated based on changing real-world conditions. This has driven the emergence of orchestration tools or frameworks that manage, validate, and monitor model updates across thousands or millions of edge nodes. Version control rollback is an important capability in case of failure or poor performance. Periodic or event-driven synchronization of AI models between cloud and edge devices is equally essential to enhance prediction accuracy and mitigate model drift.

3) AI model inference is growing at the edge while model training remains cloud-based

According to S&P Market Intelligence, by 2027, 62% of enterprise data is expected to be processed at the edge.[6] Edge inference is emerging as a dominant AI use case as enterprises seek to utilize models with lower latency. While AI model training still happens primarily in centralized cloud data centers due to immense computational requirements, edge inference allows trained models to run directly on endpoint devices. This results in faster response times, autonomy from network connectivity, and cost savings on cloud egress.

This is a logical trend: inferencing requires lower computational power than training but demands high accuracy and ultra-low latency, making edge deployments ideal.[7]

Meanwhile, edge training, where autonomous systems store data, autotune, and update models locally, is slowly gaining some traction. However, edge training is limited by infrastructure requirements and high compute costs. We expect many enterprises to continue training models in the cloud while shifting the most latency-sensitive and cost insensitive classes of inference to the edge.

> *"There are two problems at the edge with respect to AI: training on the edge and inference at the edge. The latter is more interesting. We expect customers' inference layers to explode in growth, but many prompt augmentation or context management approaches are naïve. Solutions must be able to scale at the same pace."*



Derek Collison

Founder and CEO, Synadia

What Executives Want: Areas for Opportunity

1) Lower latency and energy requirements

Enterprises are placing a premium on solutions that deliver ultra-low latency and energy efficiency without sacrificing performance- a shift driven by the rise of mission-critical real-time applications such as autonomous vehicles, industrial robotics, telemedicine, and augmented reality. Moving to the far edge is an attractive investment for these applications, which are relatively cost insensitive and require near-instantaneous data processing.

Certain optimization strategies can make running models both cheaper and faster by enabling less complex computation. Model quantization reduces the size and complexity of machine learning models and enables them to run on GPUs or FPGAs. By using certain data types that reduce the model's memory footprint, quantization enables faster execution, which extends battery life. There is a lower amount of bandwidth required to initially push a smaller model out to the edge, an appealing proposition for industries using devices in resource-constrained environments. Edge AI startups that deliver optimization and determine the sweet spot between efficiency and accuracy will be in high demand.

However, improvements in latency and energy efficiency come with trade-offs. Enterprise leaders must evaluate the balance between gains and increased costs and complexity associated with building and maintaining far-edge infrastructure. As enterprises increasingly scrutinize the total cost of ownership (TCO) for edge deployments, startups must offer scalable architecture and deliver consistent low-latency performance.

> *"At the edge, latency is a primary factor. It depends on the target industry, the data they want and how often. It depends on the target industry, the data they want and how often [they want it]. The closer you get to the operation, the more latency becomes a factor."*



Miles Hill

Co-founder and former CEO, Rivitt

2) Security at the edge

According to IDC, 75% of organizations say the workloads they run at the edge are highly critical or critical.[8] Given growing enterprise concerns over data privacy and security, edge computing can help ensure that sensitive data remains protected by staying in local environments, circumventing the need to transfer large datasets to the cloud. This also simplifies compliance with regulations like the EU's General Data Protection Regulation (GDPR).

However, unlike in the cloud, AI-based applications running on edge environments currently lack traditional security technologies and standard privacy protection mechanisms. Security solutions like EDR are too heavyweight to operate in far edge devices and require continuous connectivity. Securing AI applications at the edge requires a different approach.

Edge environments are also at an increased risk of certain cyber threats due to their distributed nature. Edge environments support a large number of users, introducing risks like intellectual property theft and model poisoning that targets continual learning-malicious attempts to poison the local model before it is aggregated to the centralized server. Such poisoning can reduce the accuracy of the global model, leading to widespread hallucinations and false outputs.[9]

Given these risks, user access policies are an important consideration, and executives will seek advanced authentication and authorization systems to protect their applications at the edge. Concerns regarding data confidentiality will also affect how companies build their edge environments and select solutions. High-risk AI applications, particularly in industries with strict privacy requirements such as finance, healthcare, and government, may push executives to look for edge solutions with embedded data governance, security, and user privacy capabilities.

> *"The more power given to the edge, [the more] there is a tradeoff between innovation and security. We must look for ways to power next generation use cases while still being secure, but we're still in the Wild West. To deploy, enterprises must define clearly security and service policies and ensure the exposure is worth the risk."*
>
> 
>
> Tim Booher
>
> SVP, Leidos; former VP of Autonomy, Boeing

> *"As we start to look at the capabilities that are going out to these IoT and operational technology assets in the field, there's a huge appetite to [integrate] artificial intelligence.*
>
> *The integrity of that asset is critical since the data that it's presenting back is what I operate on. Key considerations around the adoption of these capabilities and technologies include human-in-the-loop requirements and mechanisms to be able to validate when the machine is wrong."*
>
> 
>
> Spencer Wilcox
>
> Executive Director of Cybersecurity, NextEra Energy

3) Flexibility and interoperability

The overall efficiency of edge computing and edge AI depends on standardized data formats and seamless communications between systems, given the diversity of edge data sources and applications. Executives want to connect a large volume of devices and monitor them at the same time. Since the location of edge nodes is flexible, managing those nodes must be equally flexible, adapting to new requirements with efficient updates across diverse environments including remote sites, factories, and retail sites. The need to scale at the edge may push executives toward solutions

that offer centralized end-to-end management and orchestration. Investing in these capabilities will shorten time to market and may help companies gain a competitive advantage.

> *"Latency requirements are driving development from the near edge to the far edge, but the rules differ around security and autonomy. The edge frontier is competing against cloud services, but when exclusive autonomy is required due to the difficulty of moving workloads off the cloud, companies need dedicated technologies that provide the platform to succeed in a short amount of time. What worked in the cloud doesn't work at the far edge."*



Derek Collison

Founder and CEO, Synadia

Final Thoughts on the Future of Edge AI

Edge computing is poised for a dramatic acceleration driven by the global adoption of 5G across industries such as logistics, agriculture, and industrial automation. This unlocks new opportunities and the potential for innovative applications, in addition to new operational needs.

AI models deployed at the edge must continuously adapt with on-device model tuning or federated learning to maintain performance levels. Security vulnerabilities will need to be quickly patched across large, distributed environments while performance must be autonomously monitored in real time to minimize downtime.

Our take on the market: Bullish

There is significant potential in edge AI startups that enable lower latency AI inferencing, proactive edge security, and lower cloud egress costs and energy consumption. As the demand for edge-native infrastructure grows, we expect continued innovation in lightweight model architecture and distributed AI monitoring.

Here's what could change our minds:

1) If model accuracy degrades significantly as organizations compress and fine-tune models for deployment on edge devices, there is a risk of unacceptable performance tradeoffs, especially in highly regulated or safety-critical environments. If AI models fail to perform with sufficient accuracy, enterprises may hesitate to utilize AI inference at the edge.

2) If cloud incumbents can develop lighter-weight, integrated solutions that address cost and performance challenges at the edge, emerging startups may struggle to compete based on platform depth and pricing. Legacy providers that prioritize ease of use and support for global infrastructure could convince enterprise customers to delay investments in new edge tools.

As we look ahead, we believe that edge AI is a critical frontier for global enterprises across industries. We look forward to seeing how this exciting market evolves.

Acknowledgements

Special thanks to the following individuals for their insights and contributions.

Tim Booher
SVP, Leidos; former VP of Autonomy, Boeing

Derek Collison
Founder and CEO, Synadia

Miles Hill
Co-founder and former CEO, Rivitt

Chris Sharp
Managing Director of Product, JPMorgan Chase

Spencer Wilcox
Executive Director of Cybersecurity, NextEra Energy



Preisha Agarwal
Analyst



Rey Kirton
Principal



Kathryn Shih
Venture Partner

With Conor Higgins

About Forgepoint

Forgepoint is a leading venture capital firm that partners with transformative cybersecurity, artificial intelligence, and infrastructure software companies protecting the digital future. With the largest sector-focused investment team, over $1 billion in AUM, and an active portfolio of over 40 companies, the firm brings over 100 years of collective company-building expertise and its Global Advisory Council of more than 100 industry leaders to support exceptional entrepreneurs. Founded in 2015 and headquartered in the San Francisco Bay Area and London with a presence in Madrid and Paris, Forgepoint is proud to cultivate a diverse, global community dedicated to backing builders of the digital future.

Learn more at https://www.forgepointcap.com and https://www.linkedin.com/company/forgepoint-capital.

Company Highlights: AI at the Edge

MULTIVERSE COMPUTING

The AI model compressor to make AI systems faster, cheaper and energy efficient

multiversecomputing.com

Synadia

Secure, real-time messaging powered by NATS for distributed systems

synadia.com

Portfolio

Forgepoint is proud to partner with these companies securing the way people live and work. For more information, see https://forgepointcap.com/companies/.

1KOSMOS, aka identity, ANITIAN, AppMap, BISHOPFOX, cinchy, Constella INTELLIGENCE, Converge, CyberCube, DATABAHN, FORESCOUT, GetReal, HUNTRESS, hyperproof, Instabug, interos, loginradius, LUMU, Lynx, maisa, MULTIVERSE COMPUTING, NowSecure, nudge, Qevlar AI, RAD SECURITY, RAFAY, RAPIDFORT, REVERSINGLABS, SECURE CODE WARRIOR, SolCyber, SPHERE, STRATA, SUREFIRE CYBER, SYMMETRY, Synadia, tadaweb, uptycs, VERITUITY, whistic

Exits

CLOUDFLARE AREA 1 SECURITY, Attivo NETWORKS A SentinelOne Company, BehavioSec, SECUREAUTH | CLOUDENTITY, CYSIV A FORESCOUT COMPANY, ermetic a Tenable company, IDX A ZEROFOX COMPANY, Akamai | noname, Remediant NOW PART OF netwrix, snowflake | truera, concourse LABS

Notes

[1] Glanville, George. "Edge AI Market Forecast: The USD157 Billion Opportunity," *STL Partners*, Jan 29 2025. Edge AI market forecast: The USD157 billion opportunity - STL Partners

[2] Caspi, Ido. "Internet of Things and Devices Poised for Edge AI-Driven Upgrades," *GlobalX ETS*, Jul 25 2024. Internet of Things and Devices Poised for Edge AI-Driven Upgrades

[3] Mahler, Charles. "Kubernetes on the Edge: Getting Started With KubeEdge and Kubernetes for Edge Computing. *Cloud Native Computing Foundation*, Aug 18 2022. Kubernetes on the edge: getting started with KubeEdge and Kubernetes for edge computing | CNCF

[4] Whiting, Kate. "AI: What are SLMs and are they Good for Businesses?" *World Economic Forum*, Jan 23 2025. AI: What are SLMs and are they good for businesses? | World Economic Forum

[5] "IDC Estimates Global Spending on Edge Computing to Grow at 13.8% Reaching Nearly $380 Billion by 2028." *International Data Corporation*, Mar 17 2025. https://my.idc.com/getdoc.jsp?containerId=prUS53261225

[6] Sholom, Nati and Jeff White. "Inferencing at the Edge," *Dell Technologies*, Feb 28 2024. Inferencing at the Edge | Dell Technologies Info Hub

[7] Hsu, Jane and Peter Lin. "AI Chip Market Evolution: Cloud vs. Edge in Training and Inference." *Researcher and Research*, Feb 19 2025. AI chip market evolution: Cloud vs. edge in training and inference Part 2: Edge training, inference, and market trends | Researcher and Research

[8] "Accelerating the Future with AI and Edge Computing," *Dell Technologies*. Accelerating-the-future-with-ai-and-edge-computing-ebook.pdf

[9] Tolpegin et al. "Data Poisoning Attacks Against Federated Learning Systems." *Computer Security – ESORICS 2020*, Sept 12 2020. Data Poisoning Attacks Against Federated Learning Systems | SpringerLink

Forgepoint

What's ahead in AI at the Edge

Backing builders of the /digital future/

forgepointcap.com
650.289.4455
400 S. El Camino Real / Suite 1050
San Mateo, CA 94402

Website LinkedIn