PULSEMETER

Techstrong Group
a Futurum company

Leveraging the Edge When AI Has to Be Real-Time, Reliable, and Low Latency

by Guy Currier, Analyst, Futurum Group

SPONSORED BY
LatentAI

In AI's first great phase of lightning-quick adoption and development, the most common approach has been to use services based in the cloud. But many organizations have recognized the limitations of cloud-based AI: while ideal for rapid implementations and quick results, data management and security concerns have naturally led to investment in on-premises, data-center–based AI as well.

AI hosted at the infrastructure core may not perform as needed, though. Edge computing famously exists because of how critical real-time results, reliable operation, and low latency are for many applications. But these are important considerations for many AI applications as well—so when real-time, reliable, low-latency operation is needed, AI that has moved from the cloud to the data center is now moving further, to the edge.

Challenges in Edge AI Adoption

Despite edge AI's advantages of real-time results, reliability, and low latency, organizations face two critical barriers to adoption. First is a shortage of qualified talent to develop, and deploy, and operate edge AI systems. This talent gap is especially problematic because organizations frequently feel the need for highly customized edge AI solutions rather than off-the-shelf options.

Second is dissatisfaction with current tooling. In the view of implementers, edge AI platforms and development tools fall short, particularly in terms of customization capabilities. As most organizations seek fine-grained control over their edge AI parameters, well developed cloud-based AI tools hold more appeal—even for edge deployments.

Key Takeaways

- Edge AI is gaining traction due to its ability to provide real-time processing, reliability, and low latency.
- Talent shortages in edge AI development and operations are a major barrier for organizations.
- Organizations need highly customizable edge AI solutions, with very few finding out-of-the-box solutions acceptable.
- There is low satisfaction with current edge AI tools, signaling a strong demand for better platforms.
- Cloud-based tools for AI development and deployment are preferred, including for edge AI.
- Organizations need platforms that automate and simplify edge AI deployments, including proven, tested configurations.

Research Insights

In 2025, Techstrong Research polled our community of infrastructure, systems management, and AI readers and viewers to identify those involved in edge AI applications, gather their perspectives, and gain insight into how this promising area of AI is developing. The results tell this story of the evolution of AI deployments from the cloud to the data center and now to the edge, and how AI developers' needs both differ and align from one environment to the next.

Analyst View

Our poll showed how much cloud dominates AI deployment, with 42% of organizations choosing this path. And for good reason. The cloud offers ready-to-use infrastructure, seamless data integration, and established storage solutions, making it the natural starting point for most organizations.

In contrast, edge AI and on-premises data center deployments lag significantly, at 14% and 15% respectively. While these figures appear similar, they reflect different adoption curves: AI was deployed at the edge much earlier than the recent widespread AI "gold rush" that's centered so much in the cloud; while the data center figure reflects a recent surge in on-premises generative AI deployment mentioned earlier.

Edge AI adoption faces a clear primary barrier: the expertise gap. Organizations struggle equally with two aspects of this challenge. 34% report a lack of expertise to develop edge AI systems and 34% also report a lack of the knowledge to operate them. Cost and security concerns also figure, with 31% of organizations citing each.

New approaches are emerging, such as adaptive AI and automated ML pipeline optimization, to help mitigate the talent gap. By automating complex tasks like model selection or hyperparameter tuning, such tools reduce the need for specialized data science expertise. Pre-qualified model configurations or "recipes" are another new and promising way organizations can proceed with AI adoption without deep technical knowledge, using proven, ready-to-deploy solutions for common edge AI use cases.

The poll revealed a complex web of other operational hurdles that organizations must navigate: scaling, integration, hardware constraints, connectivity issues, and limited control. In fact, respondents expressed an overwhelming demand for control over their edge AI systems, with 55% requiring complete customization capabilities and another 40% needing at least the ability to adjust key parameters. This likely stems from the need to optimize performance for the widely variant and specific scenarios typical at the edge. Poll respondents were largely unimpressed with off-the-shelf edge AI, which is clearly anything but a one-size-fits-all solution.

What is driving AI to the edge? It's not solely about "field applications"; remember, everything not in the data center is the edge, including offices, stores, laptops, and cell phones. In our poll, performance stands out as the dominant priority for AI systems in general, with a majority (51%) of organizations ranking it as their most critical requirement. Cost considerations follow closely behind: 40% of organizations prioritize

infrastructure costs, while 37% focus on operating costs. Both of these areas are addressed at the edge in many cases—performance by proximity to data collection and use, cost by better resource utilization, lower power usage, and everyone's favorite, less need for network bandwidth.

The poll confirmed the degree to which the dominant motivation for edge AI is similar to the performance issues that drive edge applications generally: 43% of respondents prioritize edge AI's ability to process data at its source and enable real-time analysis. This capability addresses two

critical operational needs, each cited by 39% of respondents: improved reliability and reduced latency. By eliminating the need to transmit data to distant servers, edge AI both accelerates analysis and creates more dependable systems.

Notably absent from general AI priorities is energy consumption. Only 23% of organizations consider energy use to be an important factor—surprisingly low given AI's substantial power requirements. This disconnect between operating costs and energy awareness shows that, in the rush to adopt, many organizations may be overlooking a significant component of their AI systems' total cost impact, one that could in fact be addressed by the edge's classic low-power capabilities.

There is a clear preference for cloud-based AI development tools. 56% favor cloud platforms even for edge AI development. In contrast, only 27% specifically seek edge-focused tools and platforms. Poll respondents could select multiple options, so these preferences may overlap somewhat. Nonetheless, the desire to leverage the cloud within the edge AI application lifecycle is clear and strong.

A striking majority of organizations are frustrated with the current state of edge AI development tools, with 53% reporting dissatisfaction with available platforms and solutions. The depth of this problem becomes even clearer when looking at the other end of the spectrum: only 17% of organizations report being very satisfied with their tools. This widespread discontent suggests a serious mismatch between what organizations need for edge AI development and what they believe current tools deliver.

But modern optimization frameworks are beginning to address these issues. Integrated development environments are combining automated model optimization, hardware-aware deployment tools, and unified debugging capabilities to streamline the entire edge AI workflow from initial sandboxing to testing, deployment, and operation. Automation of complex tasks like model quantization or hardware-specific optimization allows more organizations to perform key elements of edge AI that previously required extensive manual intervention, time, and expertise.

Organizations are increasingly drawn to edge AI for its core benefits: increased performance, minimal latency, and enhanced reliability. However, the lack of qualified talent for both development and deployment of edge AI systems and dissatisfaction with the current tooling ecosystem present significant challenges. The future of edge AI looks bright, however, thanks in part to new platforms that bridge crucial gaps—enabling deep customization while simultaneously reducing the expertise required for implementation, providing cloud-based development while supporting edge-based deployment. As the focus on performance continues to grow in AI, the platforms and tools that successfully balance these contrasting needs will likely drive and shape the next wave of edge AI adoption.

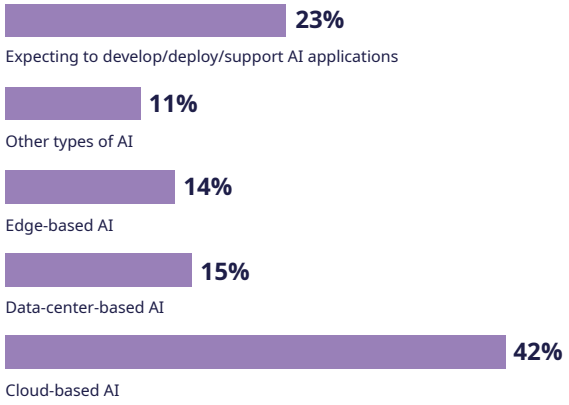
Recommendations

Edge AI must equally be accessible to a wide range of organizational capabilities and customizable to the diverse technical requirements and challenges inherent to edge deployments. Architects, developers, and operators should:

- Take advantage of new development platforms that feature intuitive interfaces and automated tooling to democratize edge AI development, enabling model optimization and edge device adaptation without requiring deep expertise.
- Prioritize platforms that automate adaptation of AI models to edge hardware constraints while maintaining performance targets. This automation should be able to handle quantization, pruning, and architecture optimization without manual intervention.
- Also prioritize platforms that offer flexible customization options by edge use case so teams can fine-tune implementations based on operating conditions.
- Ensure seamless cloud-based development and management integration with the edge systems, providing centralized control while enabling rapid updates, consistent performance across distributed systems, and operational agility.

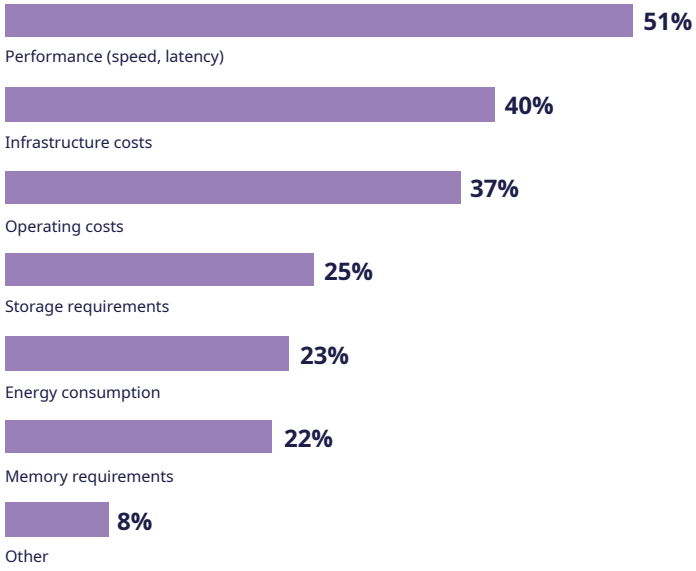
Research Results

Which of the following types of AI application have you ever developed, deployed, or supported?



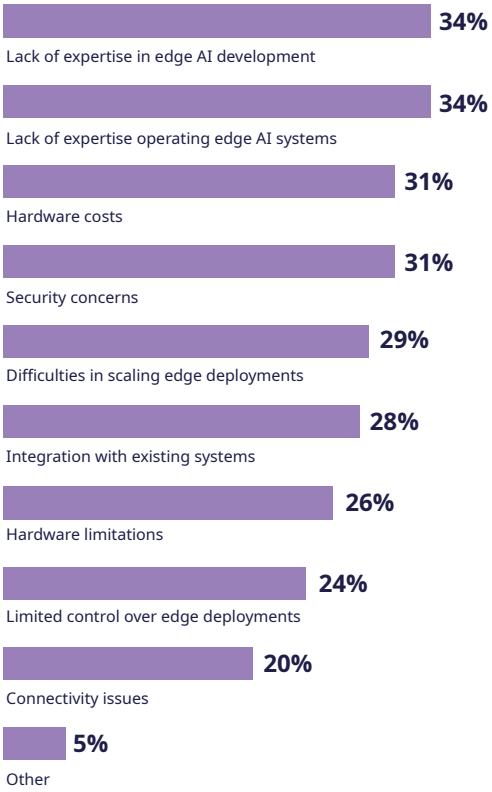
Cloud-based AI has been the most prevalent, with 42% of organizations reporting development, deployment, or support in this area. This is followed by data-center-based AI at 15% and edge-based AI at 14%, with 23% of organizations expecting to develop or deploy AI applications in the future. The fact that edge-based AI is nearly as common as data-center AI suggests that, despite cloud-based AI's dominance, there is significant recent movement towards edge deployment, especially when considering the attention paid early on in the current boom to AI implementations in the data center.

When you think about AI systems generally, what's most important to you?



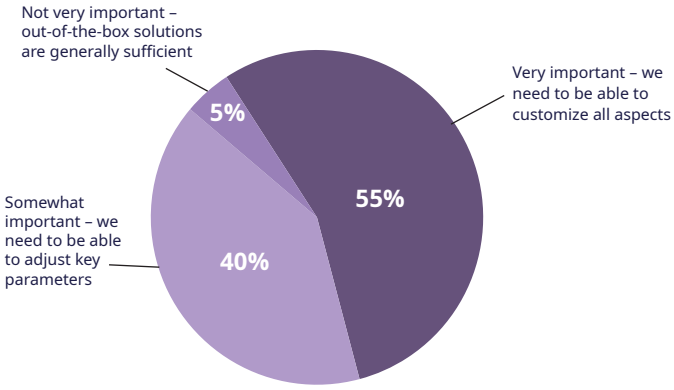
Performance (speed, latency) is the most frequent factor across all of AI considered important by organizations, at 51%. This is followed by infrastructure costs (40%), and operating costs (37%). This supports the concept that organizations are turning to edge AI to improve speed and latency of AI applications used in the field, in addition to the secondary factors such as cost. Other important factors cited for AI generally include storage and memory requirements.

If you are using or planning to use edge-based AI, what are the biggest challenges you have faced or anticipate facing?



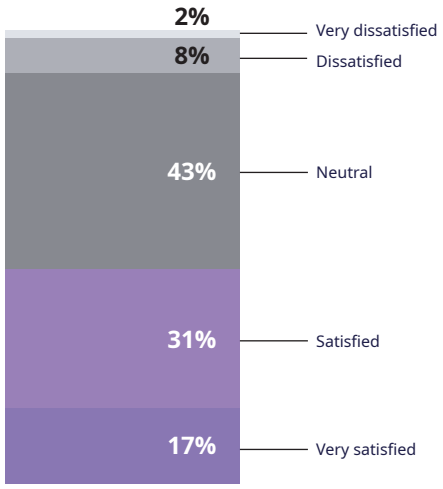
Organizations face a range of significant challenges when adopting edge-based AI. A lack of expertise in edge AI development and operations are the most prevalent, at 34% each. Hardware costs (31%) and security concerns (31%) are also significant hurdles. These results emphasize that human and knowledge resources are the primary obstacles in edge AI adoption, rather than the technology or hardware alone. Other challenges are primarily operational and include difficulties scaling edge deployments, integration issues with existing systems, and hardware limitations.

In your view, how important is it to be able to control the configuration of edge-based AI solutions?



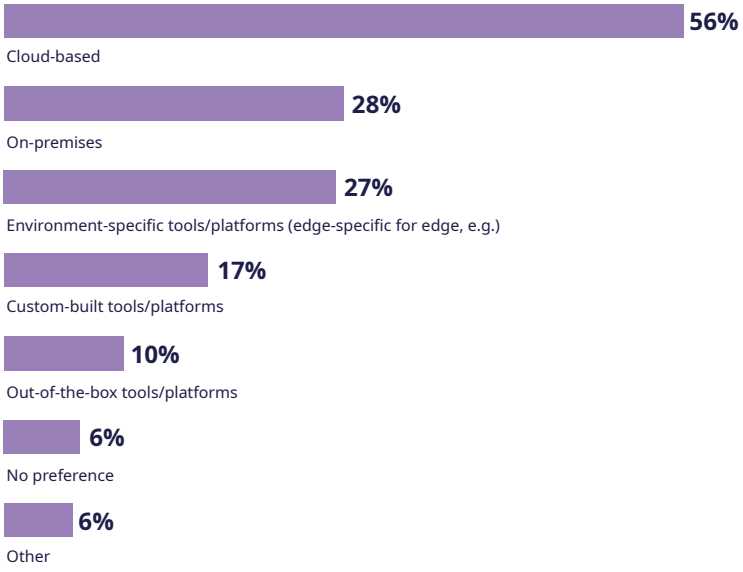
The ability to control the configuration of edge-based AI solutions appears nearly essential, with 55% of our poll respondents asserting the need to customize all aspects and an additional 40% needing to adjust key parameters. That leaves only 5% of respondents believing out-of-the-box solutions are sufficient. Organizations clearly require high degrees of control and customization in most or all of their edge AI applications to maximize their results; this highlights the need for flexible tools and platforms and is consistent with the judgment cited below that those currently available could be much better.

What is your current level of satisfaction with the tools and platforms you use for edge-based AI?



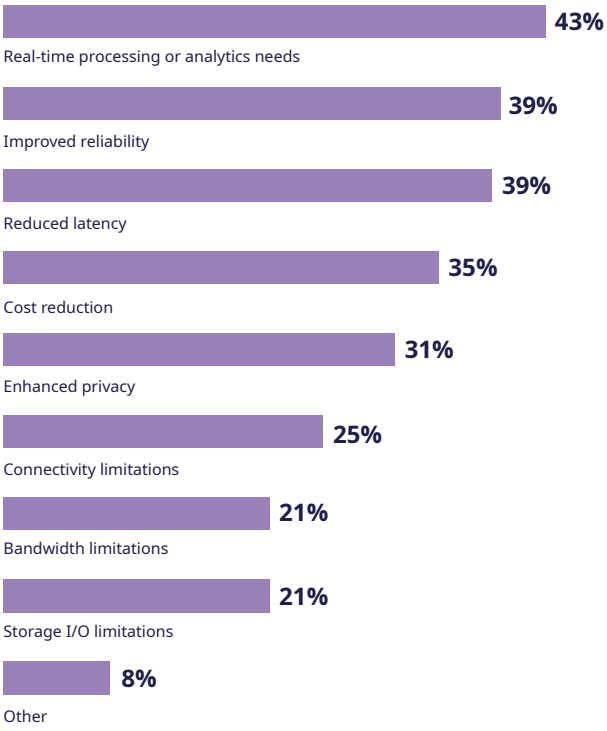
Respondents expressed a significant level of dissatisfaction with current edge-based AI tools and platforms, with only 17% of users being very satisfied and 31% satisfied, making a total of 48% satisfaction—which is to say, 52% dissatisfaction—among edge AI developers and operators. This indicates a notable perceived gap in the market for better, more user-friendly, and effective tools for edge AI and a real demand for improvements in this area.

What types of tools or platforms do you most prefer for developing and/or deploying AI systems—whatever their ultimate production environment (cloud, data center or edge)?



The majority of organizations (56%) prefer cloud-based tools and platforms for developing and deploying AI systems, regardless of the final production environment. A smaller percentage, 27%, prefer environment-specific tools and platforms, such as edge-specific tools. Seventeen percent prefer custom built tools and platforms. This preference for cloud-based tools suggests that organizations want the flexibility and scalability that cloud development environments offer, even when deploying to the edge; and points to a need for tools that integrate cloud-based development with edge deployment.

In your view, what are the primary drivers for using edge-based AI?



The primary drivers for adopting edge-based AI are related to performance and reliability, with real-time processing or analytics needs (43%), improved reliability (39%), and reduced latency (39%) being the most cited reasons. Cost reduction is a factor for some organizations (35%) but less of a driver than performance and reliability. This indicates that organizations are turning to edge AI mainly to address limitations in both cloud- and data-center–based AI regarding speed and responsiveness. Other factors include enhanced privacy, connectivity limitations, and bandwidth limitations.

This Techstrong PulseMeter is sponsored by Latent AI.
To learn more about Latent AI, visit latentai.com.