IOTech

2026 will be the year that AI is deployed at the Edge

Brad Corrion, CTO, IOTech, provides his top "edge" predictions for 2026

It’s an exciting time to work at the edge. Our relatively quiet, yet critical, sphere of computing is waking up. Why? It’s due to the demand for useful data, the newfound usability of AI and the continued success of open source and commercial edge platforms finding customer deployments.

Looking forward into 2026, I see several drivers that will significantly impact and propel edge computing forward. These will be pulled by progress in usable AI and the need for operational efficiency.

Below are six predictions for the edge computing market in 2026. These are drawn directly from our team’s work, customer investments and trend-spotting.

Prediction 1: The conversation will start with “AI managing the edge” before shifting to “AI at the edge”.

In 2026, we will move past the debate over whether we can use AI at the edge. We have first-hand knowledge that the usage and deployment is ramping up. That is not the prediction. The prediction is that the earliest production use cases will focus on the day-to-day operational work required to keep large edge deployments running smoothly, rather than the high-concept applications people tend to imagine.

Unlike process control data, AI managing the edge can begin with a cloud host, which avoids an up-front capital investment in edge hardware. It also doesn’t require the heavy validation cycles that traditional AI projects demanded, where the verification work often outweighed the benefit. And it sidesteps the site-by-site customization that used to consume more effort than the original technology development.

The biggest advantage, though, is that this kind of automation can be adopted in small steps. Organizations can build confidence and maturity as they go, using AI to monitor and address edge health issues, verify configuration correctness, surface security vulnerabilities, and handle routine maintenance tasks such as log analysis, upgrades, and device triage.

The most successful adopters will be experimenting with agentic (autonomous) capabilities to further streamline operations.

Prediction 2: 2026 is the year that messy industrial data finally becomes usable at scale.

There is a little wordplay here. Obviously, data is already being used at scale today. But the prediction is that we've reached a tipping point where we can finally unlock messy data for secondary use cases.

Part of any effort to break down data silos has struggled to find and label intrinsic meaning that can be understood beyond a given silo. This, in turn, hampered early efforts to create cost effective, multi-source analytical and AI tools, digital twins, or even the simple act of data consolidation to explore other uses.

These efforts have historically been hampered by the extremely high manual cost of identifying, sorting, and tagging this data to make it consumable by additional systems. This has been an accumulation of technical debt for so long that breaking through the debt seemed perpetually aspirational.

However, the same generative AI technologies transforming other sectors are being applied to identify meaning in device and point names that would typically require extensive manual evaluation. It's not a perfect panacea: when there is no signal in the noise, even a skilled human will struggle with the task. Yet once expectations are settled, then even reducing the manual workload by 80% is transformative and justifies continued investments in the capability.

This attack on technical debt will finally advance aspirations around ontologies and tagging, connecting systems and enabling us to build more valuable use cases from the unlocked data.

Prediction 3: 2026 is going to be the year of the small model at the edge.

This prediction is based on practicality and economics. We'll let the cloud parties keep building new atomic power plants to fuel the training of frontier LLMs. At the edge, we can't build out this kind of capability on spec, but if we reduce our expectations and tighten the scope of the problems being addressed, we find that small models are proving far more valuable (see prediction #2, for example).

Remembering the requirements that drove the edge, namely, keeping data local for privacy, secrecy, or legal considerations, means that we can't really send all of this data to a cloud-hosted LLM. Nor can we afford the (upfront) high costs of GPUs to run frontier models at the edge. It is also difficult to address the limitations in available power and heat dissipation, which make it quite unreasonable to deploy hardware sufficient for the largest models when the return on investment is not yet proven.

There is a natural fit between solving a known problem—such as parsing log data or analyzing error messages—with a small language model trained on a single input modality. We don't need audio, speech, graphics, and video to debug a dataflow or to scan the health of a container. These smaller models can be trained with cloud resources and then deployed on modest, if not existing, hardware investments. They deliver immediate value and help break the daunting return-on-investment paradoxes that have plagued AI deployments for years.

Prediction 4: Edge platforms will serve as the substrate, the foundational layer that future AI deployments require.

In the semiconductor world, the substrate is the piece of fiberglass embedded with wiring to map the highly valuable chip to the motherboard to which it will be attached. It's boring, but it's important because it connects the engine to all of the I/O that drives the CPU.

To extend the analogy to our domain, the edge platforms (basically Linux servers of various sizes) need to be able to map the AI workloads to the data and the organizations that need it. These platforms are the picks and shovels to fuel the AI gold rush. Edge platforms need to be able to deploy, update, manage versions, debug, diagnose, and otherwise function across a dynamic, rapidly changing environment as the frontier expands forward.

Organizations need to consider the substrate they are building on. They need to ensure they choose a path that makes it easy to normalize and access data, to repeatedly configure dataflows, workflows, and analytics engines and AI models. They also need to ensure that they are resting their future on a solid foundation.

Prediction 5: Organizations will stop tolerating multiple, duplicate edge platform investments and move toward company-wide standardization.

There is a growing awareness (driven by team size challenges) of the inefficiency caused by multiple teams owning duplicate platforms to handle the foundational aspects of edge computing. It isn't a criticism of how they ended up there, since each team made a series of reasonable cost and benefit decisions along the way. But the sum result is often a waste of resources to maintain duplicate platforms instead of using those resources to build value-added capabilities.

We've seen repeated instances of this pattern and feel confident that it will continue. Organizations will shift their resources and investments to modular platforms that make efficient work of deploying edge workloads — using open solutions like IOTech's Edge Central or EdgeX Foundry. This allows these organizations to focus instead on building the higher-level, differentiating value that their customers expect.

Prediction 6: AI will become a practical ‘genius persona’ knowledge source for technicians, helping to bridge the knowledge gap left by retiring experts.

We are suffering a knowledge gap as individuals with decades of experience are leaving the workforce. Combining some of the predictions shared above, the pieces are coming together to deliver AI models that combine large volumes of textual training data, as well as live data from running systems. Therefore, a technician can query the model as if chatting with a senior teammate.

This offers an optimistic rebuttal to the situation in which senior technologists gain productivity by using LLMs like a “team of interns”—but not actually training a team of interns—thereby depriving future team growth. The need is for models to augment the senior technicians as a knowledge source for training people on how to succeed in their daily jobs.

This is also an interesting twist to the long story of digital twins, whose exactness made them too expensive to deploy in a messy world with messy data. But now they can start to express their purpose via interactive industrial AI models.

Conclusion

In short, 2026 won't be defined by putting massive AI models everywhere. It will be defined by using targeted, practical AI to finally make sense of messy data, stabilize our platforms, and close the expertise gap. The organizations that focus on strengthening their edge foundations, standardizing their platforms, and adopting AI where it removes daily friction will see the biggest returns. Those investments set the stage for the higher-value applications everyone talks about but haven't yet been able to deliver.

IOTech

For additional information on our products, contact us at
info@iotechsys.com

Visit our website **www.iotechsys.com**