Enterprise Strategy Group by TechTarget

SHOWCASE

MAY 2024

Optimizing Edge Environments for AI Inference

Scott Sinclair, Practice Director; and Monya Keane, Senior Research Analyst

Abstract: Challenges managing AI workloads and container-based applications at edge locations fuel the need for better tools, automation, and management across environments. Google and Intel have partnered to offer a solution based on the open source OpenVINO, which optimizes the deployment and management of AI models across various hardware, with advanced compression and performance capabilities. The solution supports a wide range of AI models to simplify AI projects and enhance their efficiency at the edge.

Introduction

AI initiatives are on the rise. The increase in their popularity is also shining a light on the importance of inference at the edge—a process in which an AI platform uses data that it gathers from devices located in edge locations to generate new output and insights.

Beyond AI use cases, the importance of edge infrastructure is growing overall as well. Unfortunately, these increasingly important edge locations are also presenting significant challenges in terms of management. According to research by TechTarget's Enterprise Strategy Group, 67% of senior IT decision-makers say that managing data at the edge is significantly burdening their organization's IT operations.[1] Additionally, in a separate study of organizations that maintain large edge application environments, 91% of the respondents agreed that their organization would benefit from more consistency in the area of edge application and infrastructure management.[2]

Management of application environments at the edge is often fraught with complexity, especially for developers. Application development teams tend to struggle a bit more with outdated infrastructure that prevents them from identifying, isolating, and resolving application-related problems at remote or edge locations. That is the last thing developers need when they are trying to establish an AI platform that could change the business.

The partnership between Google and Intel, which is, in large part, focused on developing open source AI solutions and tools optimized for Intel processors, Google Distributed Cloud, and edge locations, should come as good news for these professionals and their organizations in general.

Challenges in Managing Artificial Intelligence Initiatives at the Edge

Enterprise Strategy Group research points to a rise in AI-based retail analytics (i.e., for insights into customers' buying trends and other behavior) that relies on leveraging data from the edge. When it comes to the rise of AI initiatives like these, adoption rates are high: Generative AI is the top data science and AI/machine learning technology targeted for purchases in 2024 by a wide margin, with 46% of organizations planning significant investments in GenAI tools. The other top technologies in this category are enterprise machine learning platforms,

[1] Source: Enterprise Strategy Group Complete Survey Results, *Navigating the Cloud and AI Revolution: The State of Enterprise Storage and HCI*, February 2024.

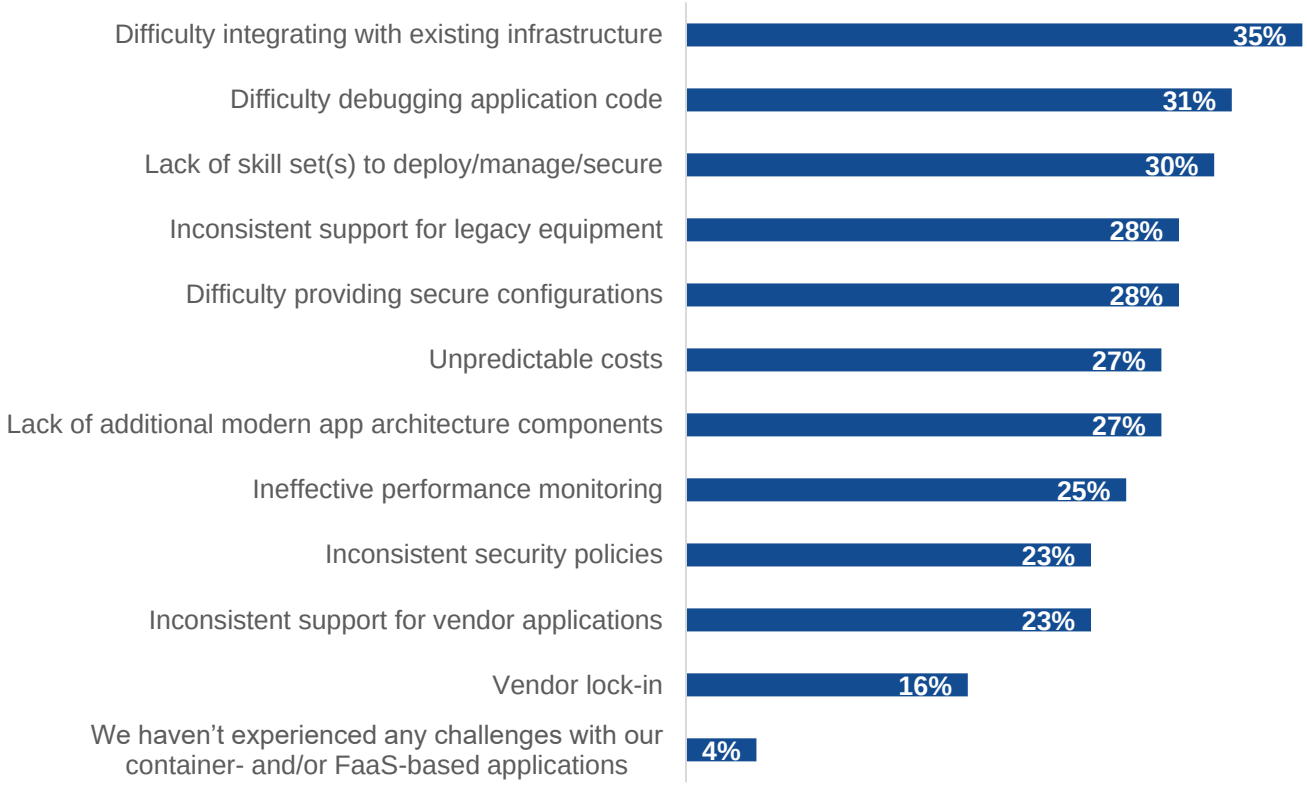
[2] Source: Enterprise Strategy Group Research Report, *Unleashing the Edge: Use Cases, Challenges, and Requirements in Edge Infrastructure and Environments*, March 2024.

in which 31% of organizations plan to make significant investments, and data science/AutoML platforms at 28%.[3] In a separate study, Enterprise Strategy Group found that 19% of organizations identified the edge as the primary physical location for their AI infrastructure deployments.[4]

Edge data's value has always been recognized, but now it is accessible to put through AI analysis—and 91% of organizations with edge environments agree that analyzing data generated at the edge provides their organization with a competitive advantage.[5] This edge data is giving these organizations the ability to better understand and engage with their customers, improve operational efficiency at the edge, and control costs.

Again, though, managing applications at the edge can be tricky. Figure 1 highlights the top challenges managing container-based applications residing at edge locations.[6]

Figure 1. Top Challenges Encountered Managing Container-based Applications at the Edge



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

As the development of AI-based applications often incorporates the use of containers, the development and management of AI workloads performing inference operations at the edge often faces similar challenges. In the case of both containers and AI, IT organizations need better tools and management capabilities. They would particularly benefit from better automation, with 30% of respondents reporting they are dealing with a lack of deployment and management skills in-house. They also need help with CI/CD workflows, Kubernetes, and

[3] Source: Enterprise Strategy Group Research Report, *2024 Technology Spending Intentions Survey*, February 2024.
[4] Source: Complete Survey Results, *Navigating the Evolving AI Infrastructure Landscape*, December 2023.
[5] Source: Enterprise Strategy Group Research Report, *Unleashing the Edge: Use Cases, Challenges, and Requirements in Edge Infrastructure and Environments*, March 2024.
[6] Ibid.

application portability—27% said they lack modern application architecture components, and 31% are finding it hard to perform root-cause analysis and debug issues.[7]

As mentioned, all of those capabilities are also essential for AI workloads, which likewise will benefit from better tools and better integration with infrastructure across all locations, especially at the edge. Such improvements will ensure good performance, provide a better usage experience, and optimize the workload for the infrastructure.

Google and Intel Collaborate to Simplify Inference at the Edge

The partnership between Google and Intel has produced a combined solution that embodies a great level of business value for performing AI inference and analytics at the edge. It centers on the Google Distributed Cloud, which uses a cloud-backed control plane that provides a consistent management experience at scale.

This is a hardware and software solution with multiple infrastructure options designed for AI-related data requirements at the edge. Within the Google Distributed Cloud hardware, Intel® chips are used inside Intel's Xeon™ scalable processors.

With this architecture, IT teams can leverage the common tools, policies, and processes that they typically use in the cloud for mission-critical use cases, running them on the edge in the same way to simplify and enhance a hybrid cloud environment for AI projects.

OpenVINO Capabilities

The technology simplifies AI projects specifically via OpenVINO, the open source toolkit designed for AI initiatives, such as those leveraging computer vision (i.e., identifying objects and people in videos and images), natural language processing, and other forms of audiovisual recognition.

Whatever their use case is, AI-focused developers get access to:

- Easy-to-leverage PyTorch for building deep-learning models.
- The open source TensorFlow machine learning library.
- TensorFlow Lite for converting and optimizing TensorFlow models to run on mobile and edge devices.
- Open Neural Network Exchange (ONNX) to standardize the representation of the machine learning models.
- Automatic compression and tools for quantization, distillation, and pruning.
- Inference API, an open source toolkit that accelerates AI inference with lower latency and higher throughput while maintaining accuracy, reducing model footprint, and optimizing hardware use.

OpenVINO supports model serving and provides an operator for deploying and scaling inference in Kubernetes clusters. It enables developers to quickly optimize, deploy, and scale AI applications across hardware device types with cutting-edge compression features and advanced performance capabilities. The model server uses the Open Inference Protocol KServe standard to expose industry-standard gRPC and RESTful interfaces.

OpenVINO is able to optimize AI-enabled applications for a variety of hardware options at the edge and can apply compression techniques to reduce infrastructure-related demands while maintaining accuracy.

With OpenVINO, developers are able to optimize and deploy AI models on premises and at the edge on a wide range of hardware, including Intel® Xeon™ and Core™ processors, Intel Arc™ GPUs, Intel Neural Processing Units (NPUs), Arm CPU, and more, to get the best performance from a wide variety of models—including classification, object detection and segmentation, recommenders, GenAI models, and LLMs.

[7] Ibid.

Intel gave Enterprise Strategy Group a look at performance benchmarks for common models such as BERT, EfficientDet, Llama 2, Mistral, MobileNetv2, ResNet-50, Stable Diffusion, and YOLOv8. This information is also available online. See the OpenVINO Performance Benchmarks here, which provide performance data for CPU, GPU, and NPU devices across client and server platforms. New data is published with each release of OpenVINO.

Real-world Benefits

In real-world terms, this solution could be great for supporting retail use cases at the edge, such as accelerating self-checkout, analyzing video for security, mapping a store's typical traffic flow, and understanding or recognizing customer behavior patterns.

In all, the combined Google/Intel solution delivers:

- Simplified and accelerated deployment of AI workloads at the edge through technology that can best utilize Intel-fueled edge infrastructure for peak performance.
- Power optimization and associated savings thanks to the enhanced optimization. Power optimization is crucial at the edge. According to Enterprise Strategy Group research, 45% of organizations with large edge environments identify high-power efficiency as essential for new edge infrastructure purchases.[8]
- Zero-touch provisioning and automation.
- Access to the power of Google's AI data analytics and database solutions to uncover insights and reduce the cost of data movement. IT teams can, thus, send insights and value-added data to the regional cloud without uploading low-value data.
- Better data locality and sovereignty.
- A hybrid cloud experience with Google Distributed Cloud, making it easier to manage and migrate data across environments.

Conclusion

AI inferencing at the edge represents the future of business for many industries—retail, manufacturing, and more. In retail, for example, it could help supermarkets or other retail shops identify which products to place next to one another to generate maximum interest and purchasing. It could also identify the habits of shoplifters and help store security staff reduce theft-related losses before they occur. In manufacturing, inferencing at the edge can obtain information from sensors and determine which data is suited for sending to the cloud and which data is not, thereby improving efficiency and conserving data-movement costs prior to AI analysis.

The key, however, is to attain real-time performance with the various edge devices associated with all of those types of workloads. Intel has provided proof with its benchmarks that its CPU, GPU, and NPU devices across client and server platforms are up to the task of helping AI developers and users get the most value out of a variety of different models.

Businesses want to leverage AI, including at the edge, but they need help with things like compressing models to run on edge hardware and reducing resource consumption. Google, with its Google Distributed Cloud, and Intel, with its renowned processors, are now offering just that kind of help, and in an open source manner to boot.

intel | Google Cloud

[8] Ibid.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com