AI Everywhere:

Key Considerations for Supporting AI at the Edge

Nina Turner
Research Director, Semiconductor Technology Supply Chain Intelligence; Enabling Technologies: AI and Automotive Semiconductors, IDC

Jennifer Cooke
Senior Research Director, Cloud and Edge Services, Worldwide Infrastructure Research, IDC

Table of Contents

CLICK BELOW TO NAVIGATE TO EACH SECTION IN THIS DOCUMENT.

In This InfoBrief 3

AI Will Be Prevalent from the Datacenter to the Edge 4

Why AI Needs Edge Infrastructure 5

Leading Edge AI Use Cases Require Real-Time Performance 6

Edge Infrastructure for AI Has Increased System Requirements 7

How Organizations Plan to Support Edge AI Workloads 8

Key Considerations for Edge Server Solutions 9

Edge Locations Often Unprepared to Support AI/ML 10

Essential Guidance/Key Takeaways 11

About the IDC Analysts 12

Message from the Sponsor 13

In This InfoBrief

Performance, latency, and data control and security needs are driving AI workloads closer to where data is created — at the edge. This "AI Everywhere" shift requires the ability to meet the unique needs of digital infrastructure deployed outside of traditional datacenters. This IDC InfoBrief provides insight into the key challenges organization face and the different paths they can take to embrace a highly distributed digital platform.

AI algorithms require additional processing and memory capabilities at the edge, with key considerations being the size of the AI model in conjunction with the required functionality. Complexity and required functionality determine whether the AI algorithm will be run on the edge device or edge infrastructure.

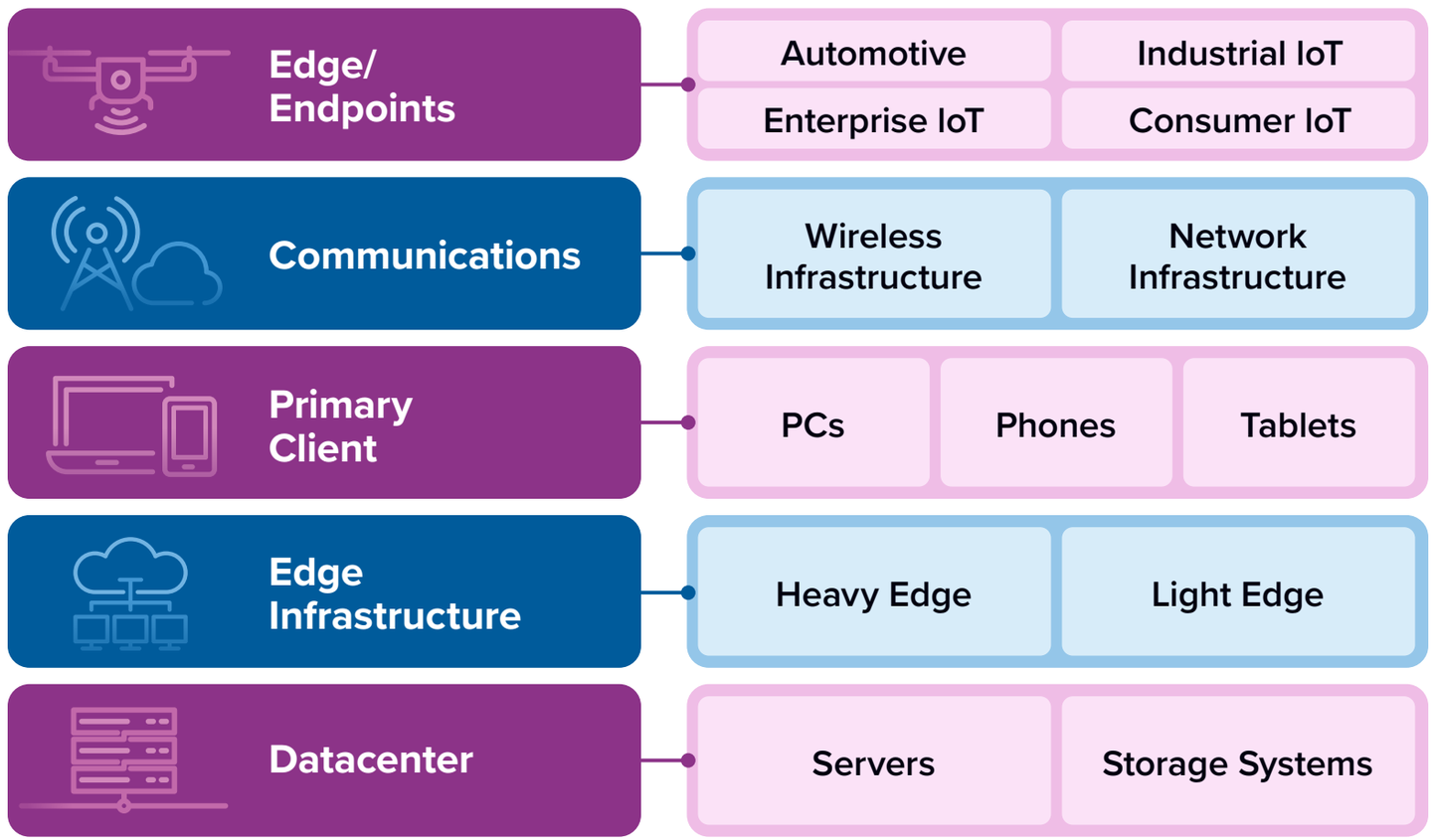
- This InfoBrief defines the system landscape where AI can be distributed as well as the key considerations for the systems handling these distributed workloads as AI migrates toward the edge and edge infrastructure.
- This InfoBrief discusses key AI use cases for the edge and leading vertical use cases and implications for edge infrastructure.

AI Will Be Prevalent from the Datacenter to the Edge

AI use cases are growing rapidly, with AI adoption across the entire system landscape.

AI requires increased system semiconductor content:

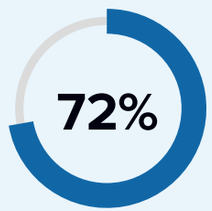
- AI-capable processors
- Increased number of sensors
- Connectivity
- Increased storage and memory



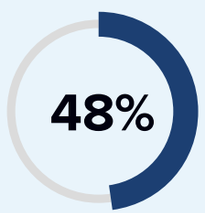
Why AI Needs Edge Infrastructure

In a digital, data-driven world, location matters.

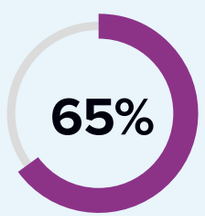
- **Latency and performance:**
  Physics of moving data places limitations on performance. Edge reduces the need to transmit data, enabling near-real-time operations.
- **Edge infrastructure to augment edge AI capabilities:**
  Edge infrastructure enables more complex AI algorithms and consolidates endpoint data for more comprehensive operational analysis.
- **Security and sovereignty:**
  Gathering, analyzing, and storing data at the edge establishes control data residency. Industries such as healthcare and financial services are required to adhere to data residency regulations.
- **Cost:**
  The expense of transmitting data back to core or cloud datacenters can be prohibitively expensive. By processing and analyzing data on site, this cost can be greatly reduced.
- **Data pipelines:**
  Access to data is at the core of all AI/ML models. GenAI speeds the process of gaining access to, cleaning, and organizing data.



of organizations say **GenAI edge workloads** are highly critical or critical.



of organizations say they need edge IT to **meet performance needs and reduce latency** for GenAI workloads. **37%** running GenAI require edge to **reduce network costs.**



say that **AI-assisted quality control edge workloads** are highly critical or critical.

Leading Edge AI Use Cases Require Real-Time Performance

Compliance and data control driving edge investment.

Retail

- Market basket analysis
- Dynamic pricing
- Inventory management

Energy/Utilities

- Pipeline inspection
- Smart metering
- Oil field/refinery optimization

Financial Services

- Fraud detection
- Transaction processing

Transportation/Logistics

- Machine vision to improve speed and quality of rail transportation
- Robotics
- Autonomous ships, cars

Healthcare

- Imaging diagnosis
- Patient monitoring
- Predictive health
- Personalized medicine

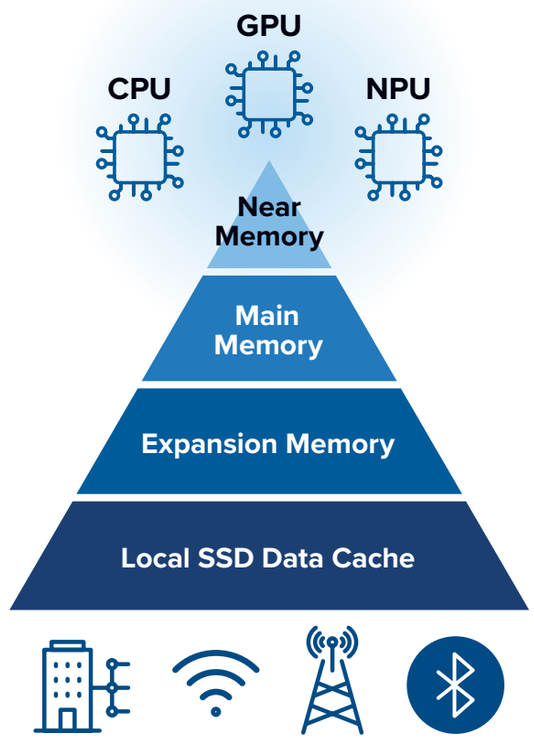
Manufacturing

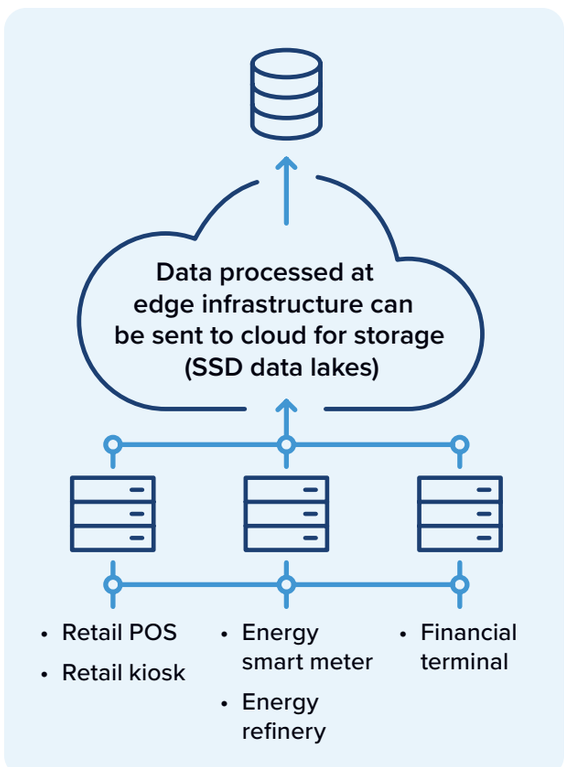
- Asset optimization
- Visual inspections and QA
- Warehouse optimization
- Robotics

Edge Infrastructure for AI Has Increased System Requirements

Compute demands for AI drive increased demands on technology.

- **Processing:** AI requires increased processing resources to handle large or heterogeneous data sets and complex AI algorithms. Depending on workload, edge infrastructure may require additional AI accelerators, such as GPUs or NPUs.
- **Memory and Storage:** AI can be memory intensive, especially when loading and running models without latency issues. A variety of memory types, from near memory such as HBM3E and DDR5, to NVMe caches and SSD storage data lakes are required to address different workloads and manage data.
- **Energy Efficiency:** Energy-efficient hardware for edge infrastructure can reduce the heat dissipation and cooling requirements, allowing for placement of infrastructure in more varied locations.
- **Connectivity:** Choice of connectivity, from wired networks to wireless connectivity, enables heterogeneous communication to a wide variety of edge devices.





How Organizations Plan to Support Edge AI Workloads

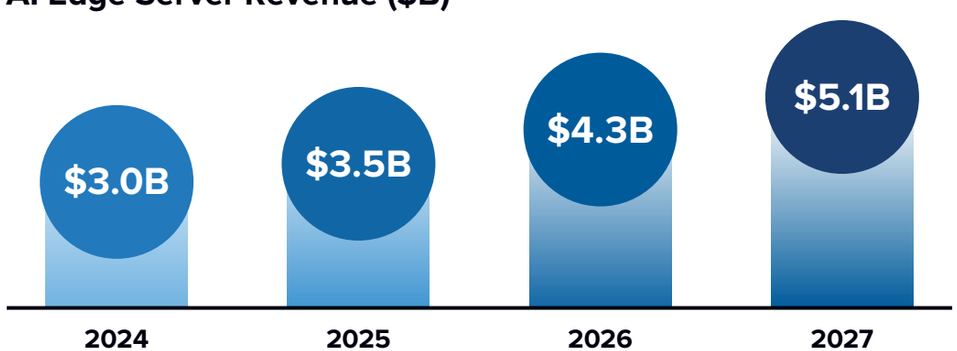
Edge infrastructure enables more capability than edge devices while reducing latency compared to the cloud.

What does this mean for processing, storage, and memory?

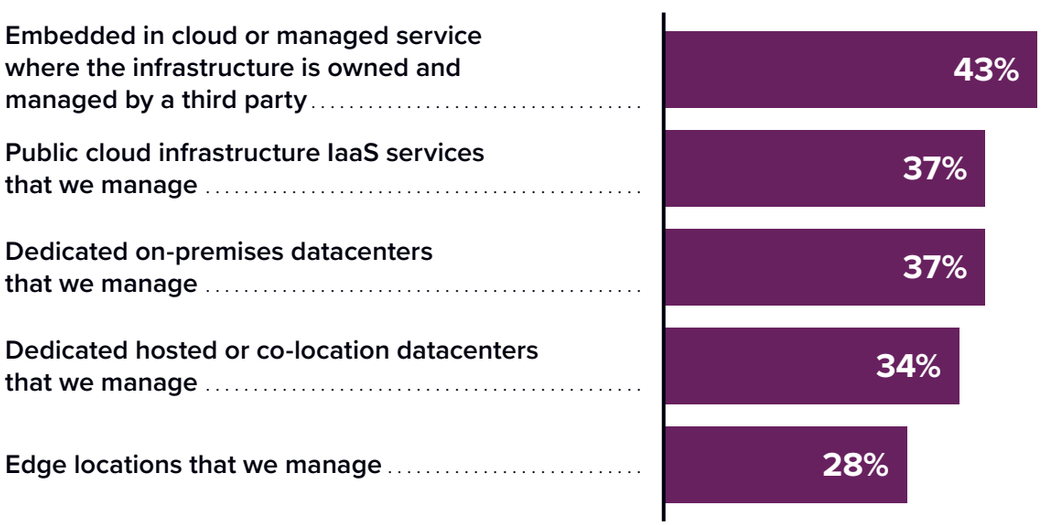
GenAI inferencing will be performed using many locations and IT operating models.

Managing data and infrastructure across many locations and operating models will challenge the majority of organizations.

AI Edge Server Revenue ($B)

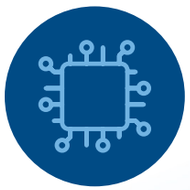


Over the next 18 months, what will be the primary approach by which your organization deploys and manages infrastructure resources for Gen AI inferencing? What will be your secondary approach?

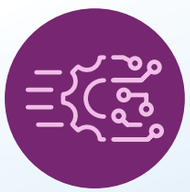


n = 800; Source: IDC EdgeView 2024

Key Considerations for Edge Server Solutions



Compute processors and accelerators sized for AI application



High-performance memory in edge AI infrastructure required for efficient compute

Flexible memory capacity and access to storage for data and AI models

Need for lower power consumption for more environmentally demanding applications

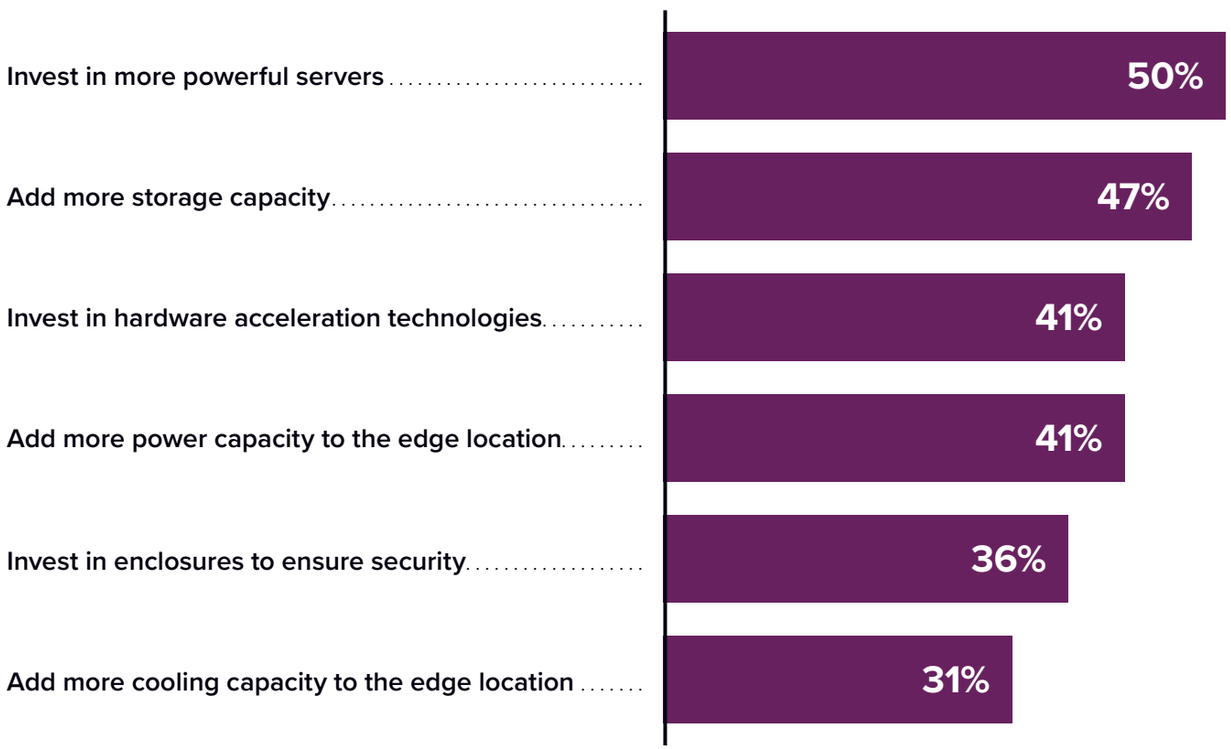
Support for multiple platforms and operating models

Ability to work in heterogeneous operational environments

Greater autonomy, ability to deploy and manage remotely

Edge Locations Often Unprepared to Support AI/ML

Will your organization need to invest in any of the following to prepare your edge locations to run AI and ML workloads?



n = 800; Source: IDC EdgeView 2024

- Most IT organizations plan to run AI workloads at the edge.
- Running AI/ML at the edge will require significant investment for many.
- Half of IT organizations will need to deploy more powerful servers.
- 40% require new hardware acceleration technologies.
- Coordination between IT and edge facility operations is crucial to keeping projects on time. Delays are common due to lack of power or cooling capacity.

Essential Guidance/Key Takeaways

- **Build a business case for edge AI projects.**
  Compute demands for AI drive increased demands on technology. Most organizations will require additional hardware investments to support demanding AI workloads.
- **Consider the unique needs of edge AI.**
  Varied environments, industries, and data sovereignty needs can derail even the most experienced IT organizations. Power capacity and energy efficiency become key decision criteria in edge infrastructure decisions.
- **Facilitate IT/OT connections.**
  Bring together business unit leaders, IT, and operations experts to uncover new ways to leverage and extract value from edge data with AI/ML. Using edge for inferencing requires a high degree of coordination across traditional organization siloes of data.
- **Embrace a distributed IT mindset shift.**
  Edge AI requires ability to secure and manage distributed data and applications with new workflows and processes. Success is dependent upon the ability to ensure data security across all locations.
- **Budget for investments in high-performance compute and memory.**
  AI applications are memory intensive, especially when loading and running models that require low latency.

About the IDC Analysts

Nina Turner
Research Director, Semiconductor Technology Supply Chain Intelligence; Enabling Technologies: AI and Automotive Semiconductors, IDC

Nina Turner is a research director on IDC's Enabling Technologies and Semiconductor team. Her core research includes the Semiconductor Applications Forecaster (SAF) product, automotive technology, and energy and smart building research. Ms. Turner has extensive experience in various technology industries, supply chain management, product development and management, and market and technology strategic assessment.

More about Nina Turner

Jennifer Cooke
Senior Research Director, Cloud and Edge Services, Worldwide Infrastructure Research, IDC

Jennifer Cooke is senior research director within IDC's worldwide infrastructure research organization and part of the cloud and edge services practice. Jennifer leads IDC's research on edge (computing) trends and strategies. Jennifer's research provides insights into the ecosystem of physical infrastructure, software, and services that support secure and resilient operations at the edge. With a background in datacenter research and a 25+ year career as a technology analyst, she has a keen interest in the evolving role of technology in supporting efficient operations and innovation.

More about Jennifer Cooke

Message from the Sponsor

Micron is a leading provider of innovative memory and storage solutions that transform how the world uses information and data.

With over 45 years of experience, Micron has been instrumental to the world's most significant technology advancements, delivering optimal memory and storage systems for a broad range of applications. From smartphones and tablets to PCs and the datacenters delivering services to these devices, Micron memory can be found fueling the applications you rely on every day.

Micron understands that a robust cloud-to-edge AI strategy can reduce latency by distributing key workloads to edge and on-premises servers where data can be most effectively collected, analyzed, and used instantaneously to deliver real-time results. This leads to optimized GPU utilization, improved data security and a reduction in the cost and power associated with transporting data to the cloud.

Learn more at micronpg.com/edgeAI

IDC Custom Solutions

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. This IDC material is licensed for external use and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

IDC Research, Inc.
140 Kendrick Street, Building B, Needham, MA 02494, USA
T +1 508 872 8200

idc.com @idc @idc

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.

©2024 IDC. Reproduction is forbidden unless authorized. All rights reserved. CCPA