

WHITE PAPER

Intelligent Edge for Retail

Using the Power of AI, Cloud Infrastructure, and
Edge Computing to Achieve the Store of the Future
Today

By Jim Frey, Principal Analyst
Enterprise Strategy Group

November 2024

Contents

The Modern AI-enabled Store	3
Planning an Edge Compute Deployment	5
Options for Retail Transformation via the Intelligent Edge	6
Option 1: Buy/Build It Yourself	6
Advantages	6
Disadvantages	6
Option 2. Turnkey Edge Solutions	7
Advantages	7
Disadvantages	7
Option 3: AI-optimized Intelligent Edge Platform	7
Advantages	7
Disadvantages	8
The Google Distributed Cloud Solution for Intelligent Edge, Powered by Intel.....	8
Conclusion	9

The Modern AI-enabled Store

Competition. Profitability. Constantly changing customer expectations. These pressures are ever-present in the retail industry, and for those who have brick-and-mortar store sites as part of the mix, the stakes are even higher. How can retailers get an edge on business success? Beyond carefully managing product offerings, supply chain, and inventory, there are other transformational opportunities that promise sustainable advantages. Artificial intelligence (AI) has the potential to transform virtually every industry in today's economy, and retail is no exception. But capturing the promise of AI requires putting the organization in position to take advantage as the technology evolves and matures.

Getting there starts with a solid understanding of the business objectives and then assessing and selecting the technology options that can provide the best and fastest pathway to success. Top business priorities for managing retail stores will commonly include the following:

- Experiences that delight customers, building loyalty and repeat business.
- Optimizing and protecting store operations, ensuring resilience and cost-efficiency.
- Minimizing fraud and theft, reducing operating margin erosion.
- Standardization and rapid agility across hundreds or thousands of locations, ensuring consistent customer experiences and repeatable business success measures.
- Keeping capital and operating costs in check without undercutting essential flexibility for the future.

Preparing for AI technologies to transform retail locations into “AI-enabled stores” means paying attention to enabling local technology, such as edge computing infrastructure, combined with AI-optimized cloud infrastructure and access to AI models and adapted applications. TechTarget's Enterprise Strategy Group research has identified a number of ways in which the retail sector, as a whole, is expecting to leverage transformational AI, including:¹

1. Enhancing customer experience and satisfaction.
2. Better understanding and predicting customer behavior.
3. Improving the speed and accuracy of decision-making.
4. Improving employee engagement and satisfaction.
5. Filling skills gap/job vacancies.
6. Improving risk management and compliance.
7. Reducing costs.

While many of these objectives apply from the top down—from the corporate level—most require data gathering across the value chain, from manufacturing through shipping and delivery, and from store locations, such as customer behavior, customer satisfaction, and operational results. This data commonly drives local recommendations and decision-making.

Beyond broader, corporate-level objectives, there are also on-site store management and optimization needs that can benefit from more intelligent local applications powered with AI, such as:

- Fast, reliable point-of-sale transactions, including “scan as you shop” and self-checkout.
- Fraud and theft prevention, including automated digital security video analysis.

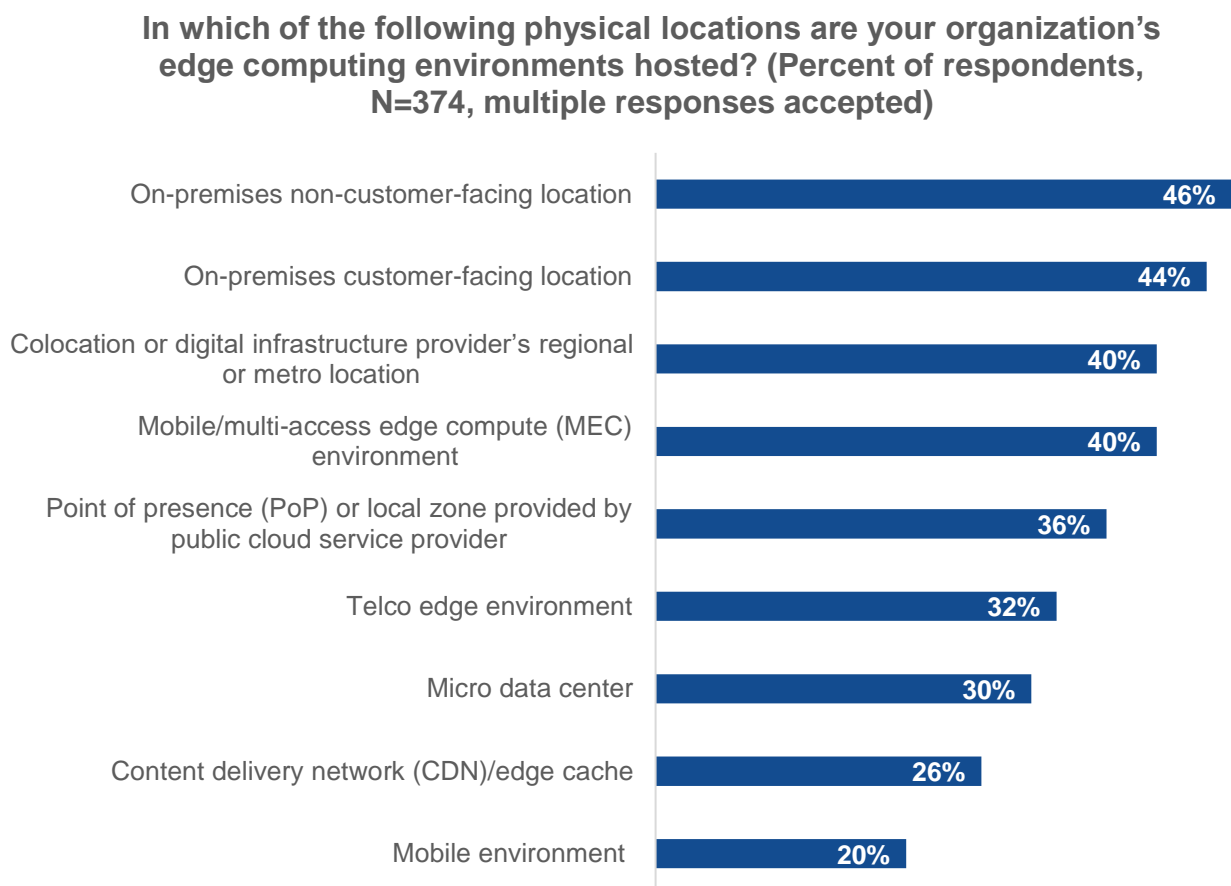
¹ Source: Enterprise Strategy Group Complete Survey Results, [Navigating the Evolving AI Infrastructure Landscape](#), December 2023.

- Inventory management, pricing, and promotional updates.
- In-store sales activity and profitability analytics.
- Customer personalization and delight via in-store kiosks, loyalty rewards programs, inventory search, and AI assistants.
- Local access security, including device authentication and identity management for store workers.

Rethinking On-site Computing to Support AI-enabled Store Transformation

While AI models and AI-enabled applications will change and evolve rapidly, the infrastructure they run on will need to follow typical technology refresh cycles. It will be necessary to deploy computing capacity on-site to support the transformation to AI-enabled operations, along with storage, networking, and integrated security. Such infrastructure initiatives are common examples of “edge computing” projects and, in this case, must include all necessary technology stacks while also being optimized for AI. Across the industry, customer-facing locations, such as retail stores, are one of the top types of deployments of edge computing (see Figure 1).²

Figure 1. Common Edge Computing Deployment Locations



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

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

Planning an Edge Compute Deployment

Establishing the right edge computing infrastructure to open the door to the modern AI-enabled store might seem straightforward, but there are many considerations to take into account and many choices to be made along the way. The bulk of the decision points can be aligned into three categories: technically oriented, resource-oriented, and business-oriented. On the technical side, here are some elements to consider:

- **Does the solution cover all the technical domains?** At a minimum, it will need compute, storage, and networking, both internal to the compute cluster (if there is more than one server and/or storage unit) and external to other systems and devices within the store, as well as to external internet access.
- **Is the infrastructure designed in a way that is optimized for AI?** Not only will it need to support existing software, but it must also be expandable or upgradeable to support AI-native applications, including AI inferencing at the edge. Ideally, the platform will facilitate modern cloud-based DevOps approaches for central development, coupled with automated deployment of software upgrades and patches.
- **Is there enough infrastructure capacity at store sites to support edge configurations?** In many cases, there might be no dedicated space for edge compute within many stores, and each store layout can be relatively unique. As a result, a wide range of environmental conditions can exist, ranging from air-conditioned offices with server racks, to warehouse spaces with little or no environmental control, and even to floor-standing servers located under desks or counters. The most important element to understand here is whether power, environmental, and network bandwidth upgrades would be required for a particular edge infrastructure solution.

Resource-oriented concerns are focused on the human factor: Who will take responsibility for various needs across the lifecycle of the edge compute infrastructure deployment, and who has the requisite experience and skill to tip the balance toward success? Internal and/or external personnel will need to take on the following tasks:

- **Conducting a requirements analysis and designing a solution.** This includes anticipating current and future needs across all included technology domains and ensuring that chosen components are designed to work together.
- **Testing and validating technology components that will comprise the solution.** Both lab testing and field testing is needed to ensure viability as planned and viability as deployed. Testing will be most needed during the initial planning phases of an edge infrastructure project, but it is also necessary during later turns of the lifecycle as components age out and must be replaced or as the solution design is changed for other technical or commercial reasons.
- **Sourcing and purchasing the solution and/or its components at scale.** This includes making sure that the necessary components are available in sufficient quantity for both deployed systems and maintenance spares, while taking into consideration expected lifespan and projected end of life, as well as multi-sourcing supply if there is a need for large quantities. Another consideration is logistics: getting components shipped to and successfully received at the appropriate locations for staging and/or deployment. Finally, it will be necessary to determine and coordinate lead times for securing delivery of the various components or systems so that the final infrastructure builds can take place efficiently.
- **Deploying and configuring the solution.** This involves going on-site to deploy the edge solution across tens, hundreds, or thousands of locations, as quickly as possible and with minimal (if any) demand on store personnel. It might also mean setting up a staging process to prepare the solution, pre-load software, preconfigure components, etc. to reduce the amount of work required upon arrival at the store site.
- **Monitoring ongoing operational health and stability.** After production deployment, edge systems need to be constantly observed to ensure continuous availability and performance in order to avoid any interruption to store operations. This will have to be done remotely so that one central operations team can keep an eye on all sites, regardless of location, region, or time zone.

- **Servicing the hardware.** All hardware systems require a means for conducting periodic maintenance, replacing obsolete components, or expanding/upgrading to accommodate growing needs.
- **Servicing the software.** All software systems require a means for patching and upgrading existing, deployed software systems, as well as adding new applications and/or removing obsolete packages.

There are also financial and strategy elements to take into account, which can help in placing technical and resource topics into the context of a total business case. Some considerations include:

- Can the solution of choice support multiple functions and software applications today and for new potential uses in the future so that the organization gets as much leverage as possible out of its investment?
- How much effort from internal staff will be required, and which aspects can and should be covered via consultants or partner/supplier services?
- What is the timeframe for deployment, both in terms of first store deployed and last store deployed, and how much (if any) disruption is likely to occur at each store during the deployment process?
- How much will need to be budgeted for ongoing maintenance, and does that maintenance include upgrades and major software releases?

Collectively, these considerations can be used to assess and select the best path forward and to evaluate which approaches might lead to the most advantageous ROI.

Options for Retail Transformation via the Intelligent Edge

Let's take a closer look at the three most common paths for executing an edge compute infrastructure deployment project: buy/build it, turnkey edge, and AI-optimized intelligent edge platform. Each provides certain advantages and disadvantages against the consideration criteria outlined above.

Option 1: Buy/Build It Yourself

With this approach, the organization's IT team decides what to buy and how to put it together and test it. The IT team also undertakes the deployment; provides all operational support before, during, and after deployment; and performs all recurring maintenance and upgrades.

Advantages

- **Ultimate control.** This approach provides complete control over which technologies are selected, how they are sourced, and how they are integrated, deployed, managed, and monitored. This means technology selection can be harmonized with other systems and components that might already be in use by the organization, as much as is possible and practical, to simplify sourcing and reduce training curves for adopting new vendors and technologies.

Disadvantages

- **Ultimate responsibility.** The buy/build option also puts organizations on the hook for every aspect of success or failure. This means that design, technology evaluation, integration testing, sourcing, staging, timely deployment, and ongoing operational management will all fall upon the corporate IT team to effectively plan and execute, managing budget as well as operational and security risks along the way.
- **High internal resource intensity.** This approach will likely place the greatest demands on internal resources for supporting the edge infrastructure across the lifecycle. Third-party resources can certainly be employed as well—and likely will be at certain stages. But the high degree of reliance on internal resources also means that it will be necessary to build up skill levels among team members to either execute the work or properly supervise the efforts of third-party resources.
- **Flexibility challenges.** Unless great care is taken during the design process, with great forethought of future expandability, internally designed solutions will often come with limitations. For instance, managing

deployments at scale, both for operational monitoring and materials management, can become daunting. Further, it might be very difficult to properly accommodate the rapid evolution and maturity of important technologies, such as AI, that will be so important for the future of the modern AI-enabled store.

Option 2. Turnkey Edge Solutions

The turnkey edge approach will typically involve a specific technology vendor or service partner delivering a preconfigured edge compute platform, commonly with specific applications pre-installed, as a means for accomplishing specific solution goals.

Advantages

- **Reduced technology decisions.** Turnkey solutions will come with a predetermined and pre-qualified assemblage of technology components, often pre-loaded with software solutions that will be used in store operations. This removes much of the design, selection, testing, and sourcing load from the IT organization.
- **Available deployment and monitoring services.** Many turnkey offerings will include optional services to assist with scaling and speed for deploying across store sites, as well as ongoing monitoring of some or all of the edge infrastructure deployment. This eliminates internal resource needs during the rollout and post-deployment operational phases, helping to scale up effectively. Care must be taken to understand which of these services are priced as one-time versus recurring.

Disadvantages

- **Fit for purpose.** Special attention should be paid to whether or not the turnkey solution being considered can be deployed without requiring expensive and slow infrastructure upgrades in stores, such as power, internal network, or cooling. Further, not all edge infrastructure solutions have been designed with an AI future in mind. If a turnkey solution is only built to support limited or application-specific AI, it might limit future AI enablement across the span of operational objectives and needs.
- **Flexibility challenges.** Many turnkey solutions are primarily built to deliver one or a few specific applications, and the ability to serve as a multi-purpose platform is an afterthought. As a result, they are not extensible and end up as solution “islands” that cannot be reused for other purposes. Additionally, not all turnkey solutions are built to be fully connected back to cloud development environments, so future deployment of new or upgraded applications (other than those initially designed into the solution) and use of new or evolving AI models become limited, at best.
- **Financial cost structure.** While some turnkey solutions are designed to be consumed as a service (Opex), others might require significant capital investments (Capex) up front. Others will lean heavily on services costs during deployment and configuration, which might hide system complexity and long project runs that delay ROI. In many cases, upgrades to the edge technology stack, including both the hardware and software it hosts, are cost-add items that result in significant additional expense.

Option 3: AI-optimized Intelligent Edge Platform

The third option, AI-optimized intelligent edge, applies and leverages architectures used in the cloud to deliver an extended “cloud-in-the-store” platform environment, completely consumed as a service. This includes edge computing infrastructure that has been optimized for AI and designed specifically to enable and support both current and future AI models and technology. These solutions operate standalone as needed but are also connected directly to the cloud so that barriers to deploying new and updated solutions can follow well-established cloud development workflows. What differentiates them from other turnkey solutions is that this is a platform approach designed for extensibility and flexibility rather than around one or a few specific applications.

Advantages

- **Reduced technology decisions.** These solutions provide the same level of advantage as turnkey solutions, coming with a predetermined, pre-validated, and pre-integrated set of edge infrastructure hardware and

software technologies, greatly reducing the efforts required for design, selection, validation, integration, and sourcing.

- **Integrated deployment and monitoring services.** Intelligent edge solutions will typically include a full lifecycle-as-a-service approach, covering design, testing, staging, deployment, and production monitoring, all included in the core licensing structure. It might also include software and hardware maintenance, further simplifying total cost of ownership and operations.
- **Fully flexible.** The hallmark of this type of solution is adaptable compute infrastructure that can run whatever applications might be chosen, whether internally developed, open source, or licensed from an independent software vendor. Further, the platform approach will naturally address developer flexibility, supporting modern DevOps regimes for building and developing in the cloud and then deploying updates quickly and easily *en masse* across hundreds or thousands of stores. Ideally, it will also support mixes of older applications, perhaps deployed on VMs, as well as newer applications deployed via containers. This can extend the ROI from legacy applications while focusing clearly on the future for new deployments.
- **AI-ready.** Intelligent edge solutions that are designed as AI-optimized will have specifically been prepared for supporting the use of AI technologies, including access to the latest AI and GenAI models.

Disadvantages

- **Requires selection of software applications.** The intelligent edge platform approach does require the selection and validation of which store management applications will be deployed on the edge infrastructure as would be the case with the build/buy option. Turnkey solutions, alternatively, will commonly come with one or more pre-loaded, preconfigured applications included.

Overall, these options are something of a continuum. At one end is build/buy, and at the other end is AI-optimized intelligent edge platform. Turnkey solutions will fall somewhere in the middle when it comes to simplifying edge infrastructure solution deployments, in terms of both advantages and disadvantages. While each organization must evaluate these options versus their own priorities and capacities, the AI-optimized intelligent edge platform is the most forward-looking approach, aligning best with modern development practices.

Google Distributed Cloud for Intelligent Edge, Powered by Intel

A prime example of an AI-optimized intelligent edge portfolio is Google Distributed Cloud (GDC), which includes a fully managed software and hardware stack that is optimized for AI, cloud-connected, and ready for the oncoming waves of new and future AI models. With this solution, Google Cloud has extended the full power of the cloud development and support model for both infrastructure and applications to the store.

- GDC includes prebuilt, pre-validated compute, storage, and networking hardware, delivering an open sourced platform that automates the management, deployment, and scaling of containerized applications and services, while also supporting legacy VM requirements. Available configurations can be single node or multi-node, supporting stores large and small. All configurations are designed for deployment with no need for expanded power, cooling, or networking, avoiding costly infrastructure updates. The system is cloud-connected but can operate standalone for a week or more if there is any interruption to internet connectivity.
- Computing is powered by 5th Gen Intel Xeon Silver processors to enable optimal, scalable performance with minimal power and cooling demands. Intel Xeon has a long-proven history of successful, reliable deployments in retail settings.

Case Study: National Retail Chain

With 6,000 stores across the U.S., this retailer was facing big challenges from the constant demand for new store-level application capabilities. Having GDC now in place, the retailer can roll updates from the cloud across its entire store network much more efficiently, cutting deployment time from weeks to days.

- The platform-based architecture empowers organizations to develop, choose, and change the applications used to run store operations and to adapt services and customer experience as often as necessary to remain competitive. It provides full developer agility to rapidly respond to new business needs, build in the cloud, and deploy to hundreds or thousands of sites in a consistent, seamless manner.
- Google Cloud supports the entire lifecycle with integrated services. For instance, deployment is accelerated by inclusive professional services to drive completion across large, distributed networks of store locations, eliminating any demands on central IT or local store personnel. Google Cloud also includes 24/7 operational monitoring and maintenance services to make sure store systems are always operating smoothly.

Conclusion

Moving to the modern, AI-enabled store is a must to keep pace with changing customer expectations and maintaining an edge on the competition, all while reducing risks and protecting profitability. The use of edge compute infrastructure is the best way forward to provide distributed, intelligent capacity for supporting store site operations that are connected back to corporate systems for monitoring and support.

Of the available options for deploying edge computing infrastructure, the most compelling is that of an intelligent edge platform that is optimized for AI. This approach is the most complete, providing the greatest degree of flexibility now and in the future, as well as the best potential ROI. Google Distributed Cloud, powered by Intel, represents a best-in-class option for AI-optimized intelligent edge and is available now to help retailers achieve the modern, AI-enabled store today.

Click here to learn more about a [Google Distributed Cloud configuration for your business](#).

©TechTarget, Inc. or its subsidiaries. All rights reserved. TechTarget, and the TechTarget logo, are trademarks or registered trademarks of TechTarget, Inc. and are registered in jurisdictions worldwide. Other product and service names and logos, including for BrightTALK, Xtelligent, and the Enterprise Strategy Group might be trademarks of TechTarget or its subsidiaries. All other trademarks, logos and brand names are the property of their respective owners.


Information contained in this publication has been obtained by sources TechTarget considers to be reliable but is not warranted by TechTarget. This publication may contain opinions of TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of TechTarget, is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact Client Relations at cr@esg-global.com.

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