SYMBEEO

PUBLIC

DESIGN STUDY

# Edge inference at 500 retail stores on a single GPU per site

Complete design on current platforms

| CAPABILITY | INDUSTRY | SETTING |
|---|---|---|
| AI & Data | Retail | Grocery chain |
| **DOCUMENT** | **VERSION** | **DATE** |
| AI-006 | 1.3 | Oct 10, 2026 |

WRITTEN BY

**Diksha Shah**, Principal AI Infrastructure Architect



© 2026 Symbeeo. All rights reserved.

## Document control

| | | | |
|---|---|---|---|
| Document | AI-006 | Version | 1.3 |
| Issued | Oct 10, 2026 | Status | Issued by Symbeeo |
| Classification | Public | Supersedes | 1.2, Sep 27, 2026 |
| Capability | AI & Data | Industry | Retail |
| Author | Diksha Shah | Role | Principal AI Infrastructure Architect |
| Credentials | NVIDIA Certified Professional: AI Operations, NVIDIA Certified Associate: AI Infrastructure and Operations, Google Cloud Professional Cloud Architect, AWS Certified Data Engineer | | |
| Approval | Approved by Edwin Ozer, Founder, Oct 10, 2026. Issued by Symbeeo, Oct 10, 2026. Design study: every figure is vendor-published, a design target or a marked assumption. | | |
| Sign-off | "An engagement under this design is signed by the practice lead before any build proceeds." | | |

## Contents

1 · Scope 3
1c · Public implementations this design follows 4
2 · Requirements register 4
3 · Architecture overview 5
4 · Design by subsystem 5
5 · Failure behavior 6
6 · Limits 7
7 · Lessons learned 7
8 · The author 7

© 2026 Symbeeo. All rights reserved. · Public

CONVENTIONS

**Source tags.** [S] figure read from the vendor's own current page on Sep 18, 2026 · [S*] figure from the vendor's page summary where the page itself could not be opened · [A] design assumption, to be confirmed by the signing engineer · [T] design target taken from the study record or derived from it. No figure comes from any prior engagement, file or asset. A row with more than one tag is counted once: as a published figure if any tag is [S] or [S*], otherwise as an assumption if any tag is [A], otherwise as a design target.

**Table 1.** Revision history

| Date | Version | Change |
|---|---|---|
| Sep 18, 2026 | 1.0 | First issue. |
| Sep 20, 2026 | 1.1 | Public implementations added; figures reconciled with published sources. |
| Sep 27, 2026 | 1.2 | Figures and platform details rechecked against published sources. |
| Oct 10, 2026 | 1.3 | Approval line now reads: approved by the Founder, issued by Symbeeo. Diksha Shah remains the author. |

## 1 · Scope

**Who it is for.** The head of store technology of a grocery chain who wants shelf-gap and checkout-queue analytics in every store, cannot send camera video off site, and has no technician in the store to touch a server.

**Typical buyer [A].** A grocery chain with 500 stores: 500 sites, 16–40 cameras per store, 2 models per store (shelf availability, queue length), one central data center and one hyperscaler tenant already in use for the office estate, a store WAN of 100 Mb/s–1 Gb/s per site with no guaranteed uptime.

**What is in.** One GPU server per store; the store-side inference stack; the central model registry and the push path to 500 sites; the rule that no video leaves the store; the failure behavior. **What is out.** The camera network and cameras themselves; the store WAN design (NET design studies); the analytics dashboards fed by the results; store identity and access.

**Assumptions.** Dell is the server vendor, NVIDIA the GPU vendor, Microsoft the management plane [A]. Each store runs a single-node Kubernetes cluster connected to Azure Arc [A]. Models and containers are pushed by GitOps from a central registry; only counts and events leave the store [T].

**Related documents, read Sep 18, 2026.** Dell PowerEdge R7715 installation and service manual, GPU kit (https://www.dell.com/support/manuals/en-us/poweredge-r7715/r7715_ism_pub/gpu-kit?guid=guid-76f16d0c-dddd-4ac3-a1f0-cc1d2c20436c&lang=en-us) · Dell PowerEdge R7725 product page (https://www.dell.com/en-us/shop/ipovw/poweredge-r7725) · NVIDIA L40S product page, read Sep 27, 2026 (https://www.nvidia.com/en-us/data-center/l40s/) · Microsoft Learn, deploy applications consistently at scale using Flux v2 configurations and Azure Policy (https://learn.microsoft.com/en-us/azure/azure-arc/kubernetes/use-azure-policy-flux-2) · Microsoft Azure Architecture Center, manage and deploy Kubernetes in Azure Arc, read Sep 20, 2026 (https://learn.microsoft.com/en-us/azure/architecture/hybrid/arc-hybrid-kubernetes).

© 2026 Symbeeo. All rights reserved. · Public

## 1c · Public implementations this design follows

Deployment rows are real public deployments, each dated. Where no public deployment matches, the row is Vendor documentation: the vendor's document that this design, Symbeeo's own, is built on.

**Table 2.** Public implementations this design follows

| Implementation | Who published it · date | What it settles for this design | URL |
|---|---|---|---|
| **Vendor documentation.** Azure Architecture Center reference architecture: Arc-enabled Kubernetes clusters across datacenters and edge locations, Flux GitOps pulled from a source repository, Azure Policy enforcing a fluxConfiguration on every connectedClusters resource, Azure Monitor container insights | Microsoft · published Jan 31, 2025, updated Sep 17, 2025 | The management pattern of R7: pull-based GitOps per site with policy as the fan-out to many clusters; monitoring of the store clusters, which R13 needs | https://learn.microsoft.com/en-us/azure/architecture/hybrid/arc-hybrid-kubernetes |

## 2 · Requirements register

**Table 3.** Requirements register

| # | Requirement | Value | Source |
|---|---|---|---|
| R1 | Sites | 500 stores, one GPU server each | [T] |
| R2 | GPU per site | 1 × NVIDIA L40S: 48 GB GDDR6 with ECC, 864 GB/s, 350 W, dual slot | [S] NVIDIA L40S |
| R3 | GPU inference and media | 733 TFLOPS FP8 dense; 3 × NVENC, 3 × NVDEC with AV1 | [S] NVIDIA L40S |
| R4 | Server | Dell PowerEdge R7715, single-socket AMD EPYC 9005; L40S listed in the GPU kit; RC4 riser up to 2 double-width GPUs | [S] Dell R7715 manual |
| R5 | Cameras per store | 16–40 streams at 1080p, 15 fps | [A] |
| R6 | Decode budget | 3 NVDEC per GPU; stream count per site checked at first build | [S] NVDEC count; [A] budget |
| R7 | Model push | central registry; GitOps by Flux v2 through Azure Arc; Azure Policy assigns the configuration at scale | [S] Microsoft Learn |
| R8 | Policy limit | 20 parameters per assignment; one kustomization per built-in policy | [S] Microsoft Learn |
| R9 | Data leaving the store | counts and events only; no video, no frames | [T] |
| R10 | WAN loss tolerance | store inference continues 72 hours with no WAN | [A] |
| R11 | Rollout | staged: 5 stores, 50, 500, one week apart | [A] |

© 2026 Symbeeo. All rights reserved. · Public

| # | Requirement | Value | Source |
|---|---|---|---|
| R12 | Hardware service | next-business-day part swap; no technician on site | [A] |
| R13 | Availability | 99 % of stores reporting each day | [A] |

Thirteen rows; six carry a vendor figure, two are design targets, five are design assumptions.

## 3 · Architecture overview

```
Central: model registry · container registry · Git (Flux source) · Azure Arc · Azure Policy
     │  GitOps pull (Flux v2), configuration by policy across 500 connected clusters
     ↓
Store ×500: Dell R7715 (2U) · 1 × L40S (48 GB) · single-node Kubernetes · Arc agent
     │ cameras (16–40) on the store LAN ─→ decode (NVDEC) ─→ shelf model · queue model
     └─→ counts and events only ─→ central analytics       (video stays on the R7715)
OOB: iDRAC per store · remote hands by part swap
```

**Figure 1.** Store edge inference and central GitOps

One 2U server and one GPU per store; the store pulls its own configuration from Git so it never needs an inbound connection; Azure Policy is what makes 500 clusters converge on the same Flux configuration.

**Table 4.** Architecture overview

| Role | Product (exact model) | Qty | Capacity used | Why this unit |
|---|---|---|---|---|
| Store server | Dell PowerEdge R7715 | 500 | 1 of up to 2 double-width GPU slots on RC4 [S] | Single-socket 2U with L40S in the supported GPU kit [S] |
| GPU | NVIDIA L40S, 48 GB | 500 | 2 models, 16–40 streams [A] | The design calls for an L40S-class part; 3 NVDEC and FP8 inference on one 350 W part [S] |
| Store cluster | single-node Kubernetes, Azure Arc-enabled | 500 | — | Arc's connectedClusters resource is the Flux target [S] |
| Management plane | Azure Arc-enabled Kubernetes, Flux extension, Azure Policy | 1 | 500 clusters | Built-in policy installs the Flux extension and the configuration at scale [S] |
| Central registries | model registry and container registry, central data center | 1 each | — | Vendor chosen at first build [A] |
| Central control node | Dell PowerEdge R7725, 2U, 2 × AMD EPYC 9005 | 2 | registries, Git | Current 2U dual-socket Dell node [S] |

## 4 · Design by subsystem

### 4.1 Store server.

Dell PowerEdge R7715 with one NVIDIA L40S in the RC4 riser, which supports up to 2 double-width GPUs [S]; the second slot stays empty for growth [A]. The manual lists L4, L40S, A16, H100 and H200 in the GPU kit [S]. Rejected:

© 2026 Symbeeo. All rights reserved. · Public

R7725 dual-socket per store — a second socket is unused at 2 models [A]. Rejected: L4 — 24 GB and no headroom for a third model [A].

### 4.2 GPU.

NVIDIA L40S: 48 GB GDDR6 with ECC, 864 GB/s, 733 TFLOPS FP8, 3 NVENC and 3 NVDEC with AV1, 350 W, dual slot, vGPU supported [S]. Decode is the binding budget at 40 streams (R6); the per-site stream limit is set at first build [A].

### 4.3 Store software.

Single-node Kubernetes with the Arc agent; the two models run as containers with the GPU shared by the two inference services [A]. Video is decoded and inferred in memory; no frame is written to disk or sent off the LAN (R9) [T].

### 4.4 Model push.

Flux v2 configurations on Arc-enabled clusters, applied by Azure Policy so that every cluster in scope receives the same configuration; the built-in policy first installs the Flux extension, then assigns the configuration [S]. One kustomization per built-in policy and a 20-parameter limit [S], so the store configuration is one kustomization pointing to one Git path [A]. Rejected: push by SSH from the center — needs inbound access at 500 sites [A].

### 4.5 Staged rollout.

Three policy assignments with different Git refs (branch, tag, semver or commit, exactly one per configuration [S]) scope the 5, 50 and 500-store rings [A].

### 4.6 WAN loss.

The store keeps running on the last pulled configuration and models; results queue locally and flush when the WAN returns [A]. 72 hours of local queue (R10).

### 4.7 Management and out-of-band.

iDRAC on every R7715 over the store WAN; a failed server is swapped as a whole unit and reprovisions itself from Arc on first boot [A].

## 5 · Failure behavior

**Table 5.** Failure behavior

| Failure | Behavior | Source |
|---|---|---|
| Store WAN down | inference continues; events queue locally up to 72 h | [A] |
| Store GPU fails | inference stops at that store; alert to central; part swap next business day | [A] |
| Store server fails | store dark for analytics; unit swap; auto-provision from Arc | [A] |
| Bad model version pushed | ring 1 (5 stores) shows it; Git revert propagates by Flux | [S] GitOps ref; [A] rings |
| Central registry unavailable | stores keep the current version; no new pushes | [A] |
| Azure Arc control plane unreachable | stores keep running; policy and Flux reconcile when it returns | [A] |

© 2026 Symbeeo. All rights reserved. · Public

## 6 · Limits

- The design calls for an "L40S-class" part; this document names the L40S because it is listed in the R7715 GPU kit. On Sep 27, 2026 NVIDIA's L40S page still offers it ("Where to Buy", no end-of-life notice) and gives the R2 and R3 figures unchanged [S], so the design keeps it. A Blackwell successor is not verified in that kit and is not named.
- 40 streams on 3 NVDEC engines is not verified; the per-site stream limit is measured at first build.
- The one-kustomization and 20-parameter limits shape the Git layout; a store needing a different configuration needs its own policy assignment.
- Next-business-day service means a store can be without analytics for a day; no in-store spare.
- No video leaves the store by design; a central review of a specific incident needs a store visit or a separate, consented export path that this document does not design.

## 7 · Lessons learned

None yet — this is a design study; populated after the first build against it.

## 8 · The author

WRITTEN BY

**Diksha Shah**

Principal AI Infrastructure Architect

NVIDIA Certified Professional: AI Operations, NVIDIA Certified Associate: AI Infrastructure and Operations, Google Cloud Professional Cloud Architect, AWS Certified Data Engineer

Signed off by Edwin Ozer, Founder.

© 2026 Symbeeo. All rights reserved. · Public

SYMBEEO

Each vendor's tool shows you their option. We show you all of them against one requirement.

## Compared. Then Delivered.

hello@symbeeo.com · symbeeo.com

© 2026 Symbeeo. All rights reserved.

Product and company names and logos are their owners' trademarks, shown only to identify the lines Symbeeo sources. Naming a vendor or showing its logo does not mean the vendor sponsors or approves Symbeeo's comparisons or any recommendation Symbeeo makes.

Figures are as published by their sources on the dates given in this document; figures marked [A] are assumptions.

AI-006 · Version 1.3 · Oct 10, 2026