

The AI your site needs, running there.

AI inference at the site, as a service

Delivery of AI inference at a site where the cloud or WAN is slow, costly, risky or absent: the models the site needs, running there, delivered and operated by Norseman on the hardware the site already runs.

Client profile. An operations or analytics owner, federal or commercial, who has been handed an AI mandate for the cameras, sensors or documents at a site, with a date attached and no team to build or run the infrastructure.

Practice. Applied AI & Data Analytics, cross listed under Cloud & Edge Platforms

CHALLENGE

The mandate arrives before the infrastructure, and the link cannot carry the data to the cloud that could run the model.

- Worldwide edge spending reached \$265 billion in 2025 and is expected to nearly double by 2029, with AI among the fastest growing domains because models need low latency inference where data is created (IDC, February 2026).
- Retail and services is the largest edge spending sector, driven by video analytics and real time operational optimization; manufacturing follows with AI quality control and predictive maintenance; financial services is the fastest growing (IDC).
- A site's cameras and sensors produce more data in a week than its link can move in a month. Raw video does not travel to a cloud model over satellite, cellular or a shared circuit.
- Edge kits, orchestration tools and cloud extensions each hand the site one layer. The models, the applications, the security evidence and the operations stay with the buyer, who has none of the three people needed to carry them.

APPROACH

Norseman delivers the inference the site needs as a service and runs the site it lives on.

- Start with one outcome: the model, the data it watches (video, sensors, documents), the decision it informs, and the latency and privacy the site requires.
- Deploy the Odin's Edge profile that fits, Mobile, Site or Region, with GPU compute sized to the model, integrated and tested in Norseman's lab on the hardware the site already runs or on new OEM hardware, shipped staged.
- Deliver the inference stack hardened: model serving, a vector store where retrieval is needed, video pipelines, signed images with an SBOM, and defined behavior when the link is down.
- Send results up and keep raw data home: detections, summaries and alerts cross the link; video and sensor streams stay at the site; models refresh on a schedule.
- Operate it: 24/7 monitoring, patching under SLA, model refresh and drift checks. Or hand it to the customer's team with the same console and runbooks.

TYPICAL OUTCOMES

The AI the mandate asked for runs at the site on the mission's date, and the buyer never owns an AI infrastructure program. Items marked target are design goals until a pilot or lab timing reports.

- Decisions in milliseconds at the site instead of seconds over the link (target; measured in the lab per model).
- Backhaul falls because detections travel instead of frames (target; measured as GB per day sent before and after).
- Live in weeks (target; measured order to live on the first timed build).
- No IT staff at the site; the analytics owner keeps the outcome and gives up the infrastructure.
- The security team receives the configuration baseline, test reports and data flow on install day.
- The same site hosts the next workload, continuity, storage or OT visibility, without a new project.

PROCUREMENT PATHS

- Federal: NASA SEWP VI, ITES-4H, GSA MAS and CIO-CS, with a hardware line and an associated service line per site so the money is classified the way the program needs.
- Commercial: per site subscription or direct purchase from Norseman, with financing, as operating expense.
- A budgetary estimate on your vehicle comes with the first conversation.

Compare vehicles and ordering steps at norseman.com/how-to-buy.

PARTNER TECHNOLOGY EXAMPLES

LAYER	EXAMPLES
Accelerators and servers	AMD, Intel, NVIDIA; Dell Technologies, Hewlett Packard Enterprise, Supermicro
Platform	SUSE Rancher and Harvester, or Spectro Cloud Palette
Inference	NVIDIA AI Enterprise, Red Hat OpenShift AI, vLLM, Ollama and Triton (open source)
Video and vision	Milestone XProtect, Genetec, Frigate (open source)
Retrieval and data	Milvus or Qdrant (open source), NetApp
Security	Palo Alto Networks, SentinelOne, Keycloak (open source)

FEATURED PRODUCT

Odin's Edge: the operated site

Norseman runs a secure cloud at your site. You run the mission.

Odin's Edge is Norseman's managed edge platform. It puts compute, storage and network at the site under one controlled architecture, on the hardware you already run, integrated and tested in our lab before it ships. Applications arrive hardened, with the configuration baselines and test reports your security team needs. Norseman operates it around the clock or hands it to your team with the same console. Buy it as one line on your vehicle or as a per site subscription.

The first thing it delivers is the AI your site needs, running there.

See it run disconnected in a 30 minute demo. Request a budgetary estimate on your vehicle. Review the security brief and boundary diagram with our engineers.

This is a representative use case. To request a briefing or a quote, contact Norseman Defense Technologies at 410.579.8600 or sales@norseman.com. Sources: IDC Worldwide Edge Spending Guide press release, February 2026 ([idc.com](https://www.idc.com)).