

Why Gridlight?

Customer Narrative

Thank you for your interest
in exploring this content.

Carahsoft is the **Trusted Government IT Solutions Provider**® supporting a broad portfolio of industry-leading technologies through a wide range of contract vehicles.

As the **Master Government Aggregator**®, Carahsoft connects government agencies, industry partners, and technology providers to deliver innovative, mission-focused solutions.

In partnership with Gridlight, we provide technology solutions that drive modernization, strengthen operations, and ensure compliance with evolving government standards.



To learn more about how Carahsoft can support your technology needs, please visit carahsoft.com



Explore More Resources:
carah.io/learn



Join Events & Webinars:
carah.io/events



Discover Technology Solutions:
carah.io/Gridlight



Learn About Procurement:
carah.io/contracts



Connect With Our Team:
Gridlight@carahsoft.com



CUSTOMER NARRATIVE • 2026

WHY GRIDLIGHT

Enterprise AI adoption is outpacing enterprise control.

Gridlight is the private AI control plane for the workloads that cannot leave your building — one governed operating layer across your models, your knowledge, and infrastructure you already own.

01

Control

Your operating boundary,
enforced in architecture.

02

Continuity

Work keeps running when the
network does not.

03

Capacity

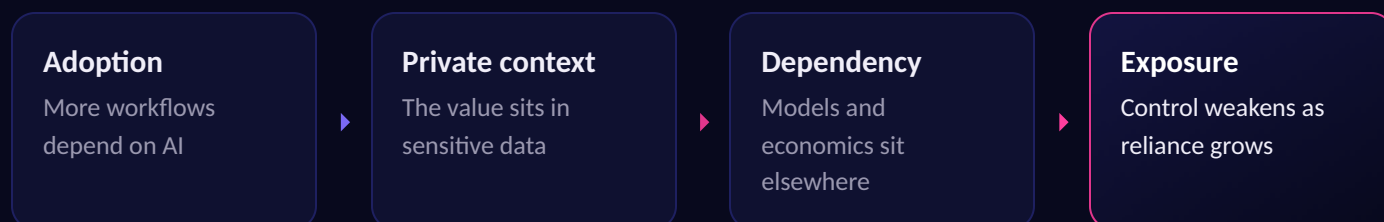
Cost tied to provisioned
compute, not a meter.

How much of your AI estate do you actually control?

As AI moves into software development, service, operations, and decision support, your business becomes dependent on models, infrastructure, availability, and economics controlled elsewhere. Yet your highest-value workloads run on private operational knowledge, intellectual property, regulated records, and data generated in factories, vehicles, field sites, and secure facilities. For many of them, the operating boundary itself must stay yours.

Waiting does not remove the risk — employees adopt consumer tools, departments launch incompatible pilots, and spend spreads across vendors. Building directly on open-source runtimes moves the burden rather than removing it: you assemble and operate the entire system around every deployment.

HOW EXPOSURE COMPOUNDS



THREE STRUCTURAL TRUTHS

The model is not your moat

Models improve and leaders change. Your durable advantage is the private knowledge, policy, workflow, and operating history that stay under your control.

Cloud privacy is not boundary control

Hyperscale providers offer real protections. Gridlight is for work where you must also control where inference runs, where context lives, and whether the system operates disconnected.

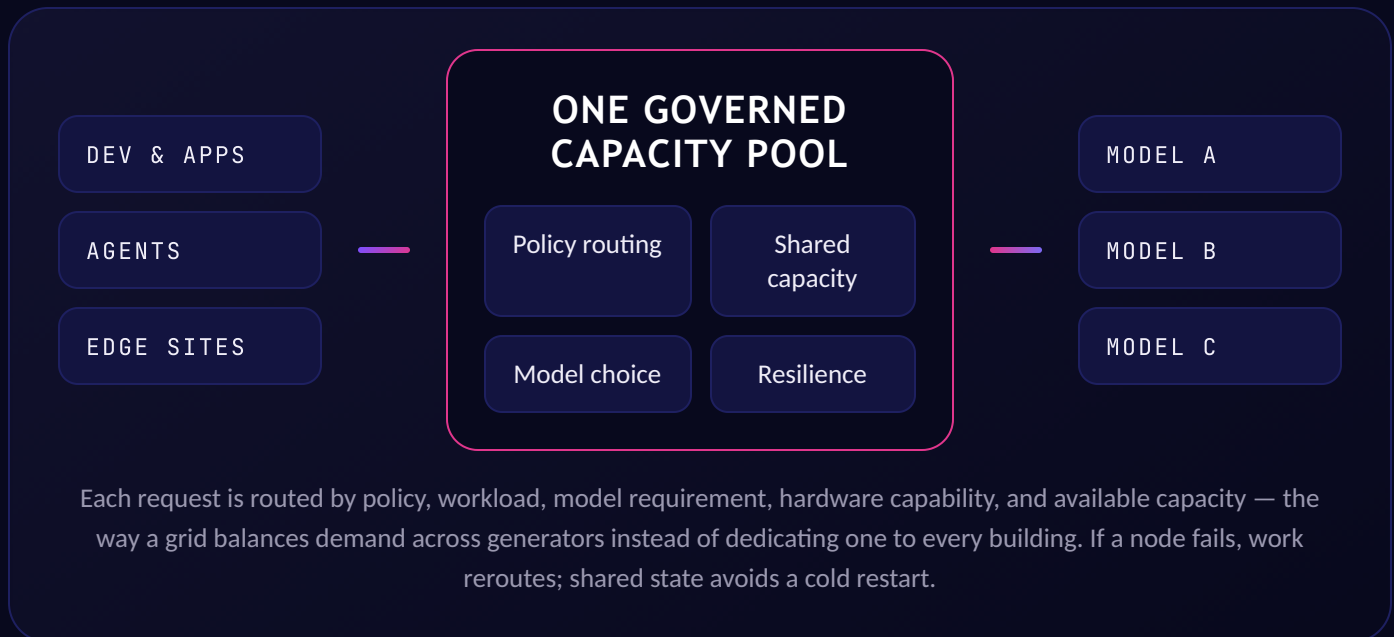
A runtime is not a platform

vLLM, llama.cpp, and NVIDIA NIM are excellent serving components. You still need identity, knowledge, policy, audit, lifecycle, resilience, and fleet operations around them.

Gridlight is not intended to run every workload. Use frontier services where their capability and managed experience matter most. Use Gridlight where sovereignty, continuity, model freedom, private context, or predictable capacity is non-negotiable — and let policy connect the two.

Stop building one AI stack for every workload

Most enterprise AI is assembled one application, one model, and one infrastructure stack at a time. The result is stranded capacity, duplicated controls, and applications that become expensive to change when models, vendors, or hardware move. Gridlight turns the infrastructure and knowledge you already own into a governed platform that evolves as they do.



WHAT CHANGES FOR YOU

Use more of what you own	Eligible systems and chip types contribute capacity instead of sitting isolated in separate deployments.
Change without rebuilding	Applications hold a stable operating layer while models and infrastructure stay replaceable.
Govern once	Identity, policy, audit, routing, and operating controls extend across every workload.
Plan capacity	Economics align with provisioned compute and expected utilization, not an unconstrained per-token meter.
Operate where required	The same control model extends from connected environments to local, edge, and air-gapped deployments.

One private platform — create, infer, operate

Your applications connect to Gridlight rather than to a single model, provider, or machine. **Gridlight Work** gives teams one place to generate content, work with private knowledge, build skills, and automate workflows. **Gridlight Manager** gives IT deployment, access, health, usage, and reporting — from one unit to an entire fleet.

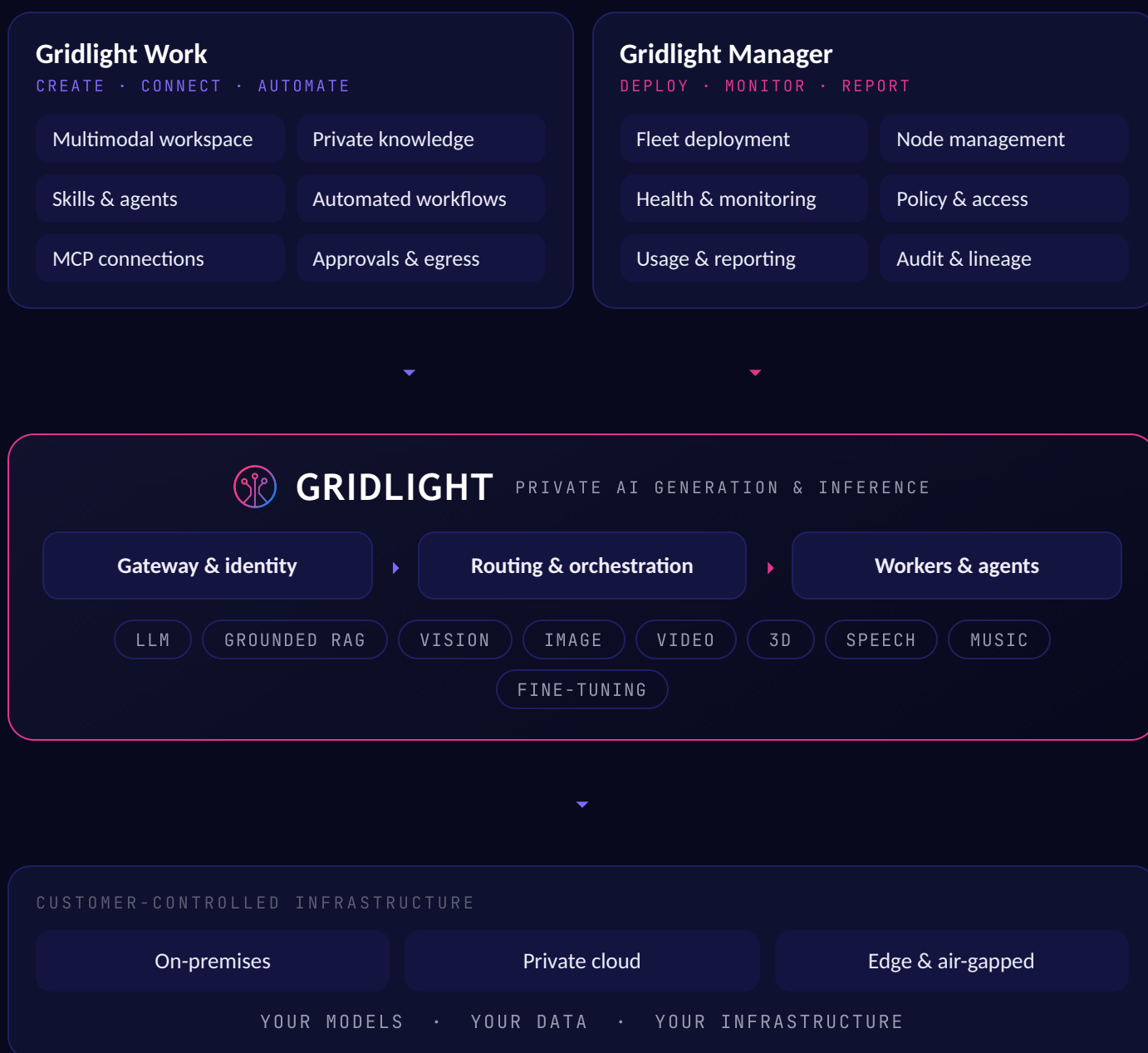


FIGURE 1 — GRIDLIGHT PLATFORM REFERENCE ARCHITECTURE

Gridlight competes on system control

You do not need every local model to beat the best frontier model. You need strong inference performance plus control of the system around your workload: private knowledge, infrastructure choice, governance, lifecycle, resilience, and operations inside your boundary.

CAPABILITY	FRONTIER SERVICES CHATGPT · CLAUDE · BEDROCK	OPEN-SOURCE RUNTIMES VLLM · LLAMA.CPP · NIM	GRIDLIGHT PRIVATE OPERATING LAYER
Operating boundary	Provider-operated	Yours to assemble	Yours, enforced — including air-gapped mode
Model freedom	Provider roadmap	Open, per deployment	Swap models behind a stable interface
Private knowledge	Uploaded to the service	Not included	Retrieval, reranking, citations, lineage built in
Existing hardware	Not applicable	Per-node, isolated	Agents advertise measured capacity for routing
Governance & audit	Strong, provider-scoped	You build it	SSO, tenancy, provenance across the estate
Disconnected operation	Requires connectivity	Possible, unmanaged	Local, edge, and air-gapped under one control model
Cost model	Per-token consumption	Hardware plus your labour	Capacity licensing against provisioned compute

PERFORMANCE COMMITMENT

More output tokens per second than both leading open-source runtimes on the NVIDIA Tesla T4 comparison platform — with the governance, resilience, and operational control around it.

+18%
vs llama.cpp

+8%
vs vLLM

Benchmark commitments depend on configuration and workload. Validate using the same model, quantization, prompts, hardware, concurrency, and output-quality checks.

Three workloads that belong inside your boundary

USE CASE 01

Software development and agentic work

THE PAIN

AI-assisted development exposes source code, credentials, and proprietary processes to the AI layer. Agentic loops generate far more requests than a person expects, turning usage into unpredictable expense and new governance gaps.

WHAT GRIDLIGHT CHANGES

Developers and approved agents get a governed path to a defined model ensemble, private knowledge, and connected tools through MCP — with identity, policy, audit, and routing held constant.

WHAT YOU MEASURE

Cost per unit of work against provisioned capacity, utilization across eligible systems, and auditable automation coverage.

USE CASE 02

Aerospace anomaly detection at the edge

THE PAIN

Engines and onboard systems stream temperature, fuel-flow, shaft-speed, and oil-condition data. Insight depends on moving raw telemetry over expensive, intermittent links and waiting for ground processing.

WHAT GRIDLIGHT CHANGES

Analysis runs on the aircraft, vessel, or remote site — combining live signals with approved procedures, flagging deviations locally, and transmitting findings instead of every raw signal. The workflow survives loss of connectivity.

WHAT YOU MEASURE

Transmitted data volume, time to awareness, diagnostic turnaround, and avoidable time out of service.

USE CASE 03

Manufacturing quality and equipment readiness

THE PAIN

Small changes undermine static systems: a new package design triggers false positives, an almost invisible seal deviation becomes a rejected shipment, a subtle equipment trend becomes an unplanned outage.

WHAT GRIDLIGHT CHANGES

Live observations are compared with history and approved production context, uncertain cases route to human validation, and sensor history, fault patterns, manuals, and parts data support earlier intervention.

WHAT YOU MEASURE

False-positive rate, defect escapes, rejected shipments, unplanned downtime, and maintenance lead time.

Prove control on one valuable workflow

Your first engagement should not be a platform transformation. Start with one use case where private data, operational continuity, or utilization creates a real constraint. Together we agree the infrastructure, knowledge source, model, baseline, and success metric before anything is deployed.

01**Select**

One use case with a meaningful constraint and an accountable owner.

02**Deploy**

Run Gridlight in your environment on approved data, models, identities, and tools.

03**Measure**

Compare accuracy, time saved, cost at expected utilization, offline availability, or audit evidence.

04**Decide**

Use the evidence to set where Gridlight expands and what stays with frontier services.

The question is no longer whether you will use AI — it is which parts of your AI environment must remain yours.

Gridlight makes that decision testable on your infrastructure, with your data and your success criteria.

Schedule a focused Gridlight evaluation

`sales@gridlight.ai`

REFERENCE NOTES

Competitive sources reviewed September 2026: NVIDIA NIM deployment and air-gap docs (docs.nvidia.com/nim); AWS Bedrock security and service docs (aws.amazon.com/bedrock); vLLM serving, security, and parallelism docs (docs.vllm.ai); llama.cpp README and server docs (github.com/ggml-org/llama.cpp); OpenAI enterprise privacy (openai.com/enterprise-privacy); Anthropic Claude Enterprise security and administration docs (support.anthropic.com). Gridlight capability statements draw from current product documentation. Performance figures are Gridlight commitments; actual results depend on model, quantization, prompts, hardware, concurrency, output-quality requirements, and utilization.