

Vehicle Powered Edge AI

*A Feasibility Framework
for Distributed Inference on Mobile Power*

Jacob Willson

Executive Summary

AI is running into the grid. A second infrastructure already generates power, manages its own heat, and computes nothing.

The International Energy Agency puts global data center electricity consumption at roughly 415 TWh in 2024 and projects it to double to about 945 TWh by 2030. The United States accounts for 45 percent of that total and data centers will drive close to half of all US electricity demand growth this decade. Siting and power delivery, more than chip supply, now set the pace of AI deployment.

More than 280 million registered vehicles operate on US roads. Each carries electrical generation, a thermal path, a weatherproof enclosure, and usually a data connection. Those are the components of an inference node, already built and already maintained.

This paper proposes a feasibility study of vehicle-powered edge AI, and the claim it makes is narrow. Vehicle power is not free, and distributed vehicles will not displace data centers. But a growing set of inference workloads cannot reach centralized compute at any price, vehicles are well suited to serve them, and no validated framework exists for deciding which deployments qualify. Building that framework, and the measurement protocol behind it, is the work proposed here.

1 The Problem

Data center electricity consumption has grown about 12 percent per year since 2017, four times faster than total electricity demand, and it concentrates rather than spreads. Nearly half of US capacity sits in five regional clusters, so the pressure lands locally and hard.

The industry's own view of where inference belongs points the same direction. A McKinsey survey of automotive OEM, Tier 1, and semiconductor stakeholders (n = 50 approx.) ranked offline availability first among reasons to move AI execution onto the vehicle, cited by 39 percent, followed by reduced latency at 35 percent and data privacy at 20 percent. Network traffic cost came last, at 6 percent.

That ordering is the important part. The people closest to the problem are not chasing cheaper inference. They want inference that works when the network does not.

The latency gap behind that ranking is measurable. The same study clocked cloud voice-assistance tasks at 1,000 to 2,200 ms against 300 to 700 ms on the edge, roughly threefold on identical work. Capital is moving accordingly: the automotive market for sub-20 nm microcomponents is projected to grow from \$5.1 billion in 2024 to \$18.2 billion by 2030, raising its share of automotive semiconductor spend from 8 to 18 percent.

2 Vehicles as Compute Platforms

A modern vehicle is a mobile power plant with a cabin attached. Its electrical system handles peak simultaneous demand from lighting, climate, entertainment, and safety systems, a peak it rarely sustains. For scale, a headlamp pair draws about 50 W and a premium audio system several hundred watts under load. An NVIDIA Jetson Orin Nano runs on 7 to 25 W. An NVIDIA DGX Spark draws about 240 W.

The hardware proposed here therefore sits well inside loads vehicles already carry. The open engineering problems are integration, thermal behavior, and graceful degradation.

The concept has also moved from speculation to earnings calls. On Tesla's Q3 2025 call, Elon Musk floated organizing idle vehicles into a “*giant distributed inference fleet*,” estimating that 100 million cars at a kilowatt each would yield 100 gigawatts of inference with power and cooling already solved.

Tesla's version depends on compute already installed in its own EVs. The approach proposed here targets the far larger installed base of conventional vehicles, which have electrical and thermal headroom but no onboard inference hardware. The addressable fleet is bigger by more than an order of magnitude, the hardware is vendor-neutral, and nothing depends on a single manufacturer's roadmap.

3 Viability Criteria

Start with an honest premise. Vehicle power is not free. A vehicle's electrical system generates on demand, and added load gets paid for in fuel. There is no reservoir of stranded energy waiting to be collected.

That sharpens the question. Whether vehicle-generated electricity beats grid electricity on price is already settled, and the answer is no. The useful question is where the *delivered cost of compute* competes, counting what it would otherwise cost to serve the location at all.

Where backhaul is reliable, centralized inference wins on cost and vehicle nodes have no business competing. The opening is in deployments where backhaul is unavailable, unreliable, expensive, or prohibited; where the latency budget rules out a round trip; or where the vehicle is already running anyway. In those settings the benchmark is fixed infrastructure that costs far more to install, or no compute at all.

Vehicle-powered edge AI is a compute availability play. Posed that way, the viable deployment classes can be identified and ranked.

4 Deployment Screening Framework

Standard edge-versus-cloud frameworks weigh output speed, latency, scalability, availability, maintainability, security, time to market, and cost. They assume the vehicle consumes AI services. The five criteria below adapt that structure for the vehicle as compute platform. Each one is observable and measurable.

- **Backhaul availability.** Is there reliable, affordable connectivity to centralized compute at the point of work?
- **Latency budget.** Does the workload tolerate a network round trip, or must the decision be made locally?
- **Edge data volume.** How much data is generated at the vehicle's location, and how expensive is it to move?
- **Incidental runtime.** Is the vehicle already operating for a reason unrelated to computing?
- **Alternative infrastructure cost.** What would it cost to serve this location with fixed infrastructure instead?

Applied to candidate deployment classes, the ranking runs counter to most public discussion of the idea.

Deployment class	Backhaul	Latency need	Edge data	Incidental runtime	Alt. infra cost	Fit
Emergency response and disaster recovery	Absent or degraded	Tight	High	High, vehicles stage on scene	Very high	Strong
Defense and expeditionary operations	Contested or denied	Tight	High	High	Very high	Strong
Remote infrastructure inspection (utility, rail, pipeline)	Poor	Moderate	High	High	High	Strong
Precision agriculture	Poor	Moderate	High	High, equipment runs for hours	High	Strong
Campus and municipal facilities fleets	Good	Loose	Moderate	Moderate	Low	Pilot testbed
Long-haul logistics	Variable	Loose	Low	High	Low	Conditional
Rideshare and consumer vehicles	Good	Loose	Low	Low	Low	Weak

The pattern holds across the table. Vehicle-powered edge AI is strongest where connectivity is the binding constraint and weakest where it is abundant. Most speculation about distributed vehicle computing has centered on consumer fleets and aggregate capacity, which score poorly on four of five criteria. Field, emergency, and remote-operations work scores strongly on all five.

Campus fleets appear here as a pilot testbed, not a commercial target. A campus offers controlled conditions, predictable routes, institutional access, and safe failure, which is what validation needs, even though it is not where the technology earns its keep.

5 Reference Architecture

Two tiers, both built from commercially available hardware, which keeps the critical path on integration and measurement instead of silicon. Workload selection follows model size: speech-to-text, text-to-speech, perception, and compact language models deploy at the edge today, while text-to-video, large-scale reasoning, and frontier language models stay cloud-scale. The node tier targets the first group. The hub tier raises that ceiling without a network link.

Tier	Platform	Role
Node	NVIDIA Jetson Orin Nano class, up to 67 TOPS within a 7 to 25 W envelope	On-vehicle inference. Perception, object detection, sensor fusion, and compact language models executed locally with no network dependency.
Hub	NVIDIA DGX Spark, GB10 Grace Blackwell, 128 GB unified memory, about 240 W	Mobile or fixed coordinator. Job scheduling, result aggregation, and models too large for the node tier.
Power	Isolated conditioning stage, buffer storage, supplemental solar	Electrical isolation from vehicle systems, brownout tolerance, clean shutdown. Vehicle primary function always holds priority.
Connectivity	Campus Wi-Fi, LTE and 5G, LEO satellite, store-and-forward	Tiered fallback down to fully autonomous local operation, which is the design case here.

Answering the stated objections

The two objections raised most often against in-vehicle edge AI are ones this design already answers. The first, cited by 46 percent of surveyed stakeholders, is resource constraint: vehicle systems-on-chip offer limited compute, memory, storage, and bandwidth next to data center hardware. The proposed design does not use the vehicle's SoC. It adds a dedicated inference module sized to the workload, which turns a fixed platform limitation into a specification choice. The cost is one more component to install, power, and maintain, and the pilot would measure that burden.

The second objection, raised by 35 percent, is that heavy AI workloads consume too much energy, flagged in the source as a particular concern for battery electric vehicles. It bites hardest where compute competes directly with driving range. It bites far less on the internal combustion platforms targeted here, where stored energy is large next to the compute load and refills in minutes.

Two principles govern the build. The vehicle's primary function is never subordinated to computation, so the compute subsystem is electrically isolated and independently sheddable. And the system must work with no network connection, because the target deployments are defined by that condition. Hybrid cloud-and-edge architectures will dominate mainstream vehicles for years to come; this proposal is scoped to the deployments where the hybrid fallback does not exist.

6 Validation Plan

Four phases across roughly six months. Each phase would produce a standalone result and generate the evidence for the next, so the work would still contribute if it stopped early. All raw measurement data would be published alongside results, which both peer review and competitive federal proposals require.

Phase	Focus	What would be measured	Output
1 Months 1 to 2	Single-vehicle baseline	Delivered power envelope under real road and load conditions, against nameplate capacity. Thermal derating curve as a function of ambient and enclosure temperature. Joules per inference by model and task.	Reproducible baseline dataset for one platform.
2 Months 2 to 4	Two-node network	Coordination overhead across the scheduler. Connectivity availability and effective throughput by link tier across instrumented routes, including behavior in coverage gaps.	Working two-vehicle network and documented replication protocol.
3 Months 4 to 5	Field validation in a strong-fit class	Node reliability under vibration, thermal cycling, and repeated power transitions. Tokens per gallon. Fully loaded cost per thousand inferences against a grid-connected baseline.	Screening framework populated with measured inputs.
4 Months 5 to 6	Package and publish	Not applicable.	SBIR-ready technical proposal, open measurement dataset, campus demonstration.

Tokens per gallon. Proposed here as a normalized fuel-to-output unit, allowing direct comparison of energy efficiency across vehicle platforms, hardware configurations, and duty cycles.

7 Funding Pathway

The study is scoped as pre-commercial feasibility work, sized to produce the technical data and economic modeling that competitive federal proposals require. Identified pathways include Department of Energy SBIR and STTR programs, with particular alignment to DOE priorities on energy-efficient AI; ARPA-E transformational energy technology funding; NSF edge computing initiatives; and NYSERDA programs addressing grid modernization and load growth.

The strong-fit classes in Section 04 also map onto federal mission needs in emergency response and expeditionary operations, which widens the set of relevant solicitations well beyond energy-focused programs.

8 Open Questions

Three questions sit outside what the measurement plan alone would settle.

- At what fleet size does coordination overhead exceed the marginal value of adding another node?
- Does a vehicle's duty cycle align with when the workload actually arrives, or does the work show up when the vehicle is off?
- Among the four strong-fit deployment classes, which offers the shortest path to a paying customer?

References

- [1] International Energy Agency. Energy and AI. IEA, Paris, April 2025.
<https://www.iea.org/reports/energy-and-ai>
- [2] International Energy Agency. Key Questions on Energy and AI. World Energy Outlook Special Report. IEA, Paris.
<https://www.iea.org/reports/key-questions-on-energy-and-ai>
- [3] Venus, A., Poltronieri, G., de Jong, M., Kellner, M., Blos, A., and Steiner, F. “The rise of edge AI in automotive.” McKinsey & Company Semiconductor Practice, in collaboration with the Global Semiconductor Alliance, August 2025.
<https://www.mckinsey.com/industries/semiconductors/our-insights/the-rise-of-edge-ai-in-automotive>
- [4] Tesla, Inc. Q3 2025 Earnings Call transcript, October 22, 2025.
<https://www.fool.com/earnings/call-transcripts/2025/10/22/tesla-tsla-q3-2025-earnings-call-transcript/>
- [5] NVIDIA Corporation. “Jetson Orin for Next-Gen Robotics.” Embedded systems product documentation.
<https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/>
- [6] ServeTheHome. “NVIDIA DGX Spark Review: The GB10 Machine.”
<https://www.servethehome.com/nvidia-dgx-spark-review-the-gb10-machine-is-so-freaking-cool/>
- [7] U.S. Federal Highway Administration. Highway Statistics Series, Table MV-1, State Motor-Vehicle Registrations. U.S. Department of Transportation.