Institute for Physical AI @ Bailey Military Institute · The Charlot Lab & The Hiner Lab

SITING THE COMPUTATION THAT SERVES EMBODIED SYSTEMS

Technical Report TR-2026-33
Survey / Review · Preprint v1
5 August 2026

SITING THE COMPUTATION THAT SERVES EMBODIED SYSTEMS

Economic and Environmental Impact of Distributed AI for Physical AI

Centralised hyperscale facilities are the default location for the computation that serves physical AI. This review surveys the alternatives, grades the measured evidence for and against each, and computes the physical limits that bound the most distributed end of the range.

Grant Markhart, Industrial Research Fellow · The Charlot Lab & The Hiner Lab

92 numbered sources, each carrying a verification grade

5 tables · verified / reported / self-published / modelled

Every computed figure shown with its inputs and its grade

No forecasts and no siting recommendation to any operator

Physical AI raises a siting question. The computation that serves a machine moving through the world can be carried by the machine, placed in a nearby facility, embedded in a network operator's infrastructure, or concentrated in a centralised hyperscale campus, and the choice carries economic and environmental consequences that are argued about more often than they are measured. This review surveys the published evidence for each option and grades every figure it uses. The centralised baseline is large and imprecisely known: US data centres consumed 176 TWh in 2023 [1], a bottom-up model built from equipment shipment counts rather than a metered total, and the widely quoted 74 to 132 GW capacity figure is that model's energy range divided by an assumed 50% average utilisation [2,92]. Latency is the argument usually made for distribution. Measured 5G round trips are 17.27 ms to a node inside the operator's network against 44.08 ms to a

cloud region, with a radio access term near 7 ms that does not move when the server moves [23]. Published vision-language-action policies run inference every 0.5 to 0.8 seconds behind chunked action sequences [26], so a regional round trip leaves substantial headroom for policy inference and none at all for a safety-rated separation function [28]. Environmental comparison turns on utilisation and on the choice between marginal and average emission factors, which differ nationally by a factor of 1.6 to 2.0 [22,92]. The review then derives a general limit on ambient-powered computation: at a 100 cm² aperture and measured road-surface photovoltaic yields, harvest is 8.61 to 106.2 mW [70,72,92], which sustains tinyML-class inference continuously and a 7 W embedded inference module at a duty fraction between 0.12% and 1.52% [59,80,92]. The Institute's research position, the limitations of the review, and the measurements that would settle each open question are stated.

1. Introduction: the siting question and why it remains open

A machine that moves through the world produces sensor data and consumes decisions. The computation that turns one into the other has to happen somewhere, and the plausible locations span roughly four orders of magnitude in distance: on the machine itself, in a cabinet in the same building, in a metro facility a few kilometres away, inside a mobile network operator's aggregation point, or in a hyperscale campus several hundred kilometres distant. The default today is the last of these. US data centres consumed 176 TWh in 2023, 4.4% of national electricity, up from about 76 TWh in 2018 [1], and global data centre electricity was around 415 TWh in 2024 [3]. Capital formation runs at a comparable scale: three of the four largest US operators spent $245.69bn on property, plant and equipment in the first half of calendar 2026 alone, on their own audited filings [6,92].

The question this review addresses is where the computation that serves physical AI should physically sit, and what that choice costs economically and environmentally. The question is open for three reasons that are specific to embodied systems and do not arise for text and search workloads. First, the consumer of the inference is a machine with a control loop and a stopping distance, so latency is a physical quantity with physical consequences rather than a quality-of-experience metric. Second, the machine already carries a computer for its own safety and control functions, so the marginal cost of local inference is not the cost

of a new device. Third, the environmental accounting changes character at small scale: for a centralised accelerator the dominant term is operational electricity, while for a device the dominant term is the carbon embodied in manufacture, and the crossover between the two depends on utilisation and on the carbon intensity of the grid rather than on any property of the hardware.

The literature that bears on this question is scattered across four disciplines that rarely cite one another. Energy accounting sits in national laboratory reports and life-cycle assessment. Latency measurement sits in networking venues. Control-rate requirements sit in robotics papers and machine-safety standards. Cost sits in operator tariffs and regulatory dockets. Each body of work is internally careful and each answers a different question, and the siting question falls in the gaps between them.

This review makes four contributions. It consolidates the measured evidence into one graded evidence base, so that a reader can see which figures are metered, which are modelled, and who published each. It sets out a taxonomy of distributed topologies together with their deployment record, including the withdrawals, which are the part of that record most often omitted. It performs the environmental and economic arithmetic explicitly, showing every input, so that a reader who disagrees with an input can substitute their own and recompute. And it derives a general physical bound on the most distributed end of the range: what computational duty a device powered only by harvested ambient energy can sustain at a small aperture, given measured harvesting yields and measured per-inference energy.

The review does not recommend a siting decision to any operator, does not forecast the size of any market, and does not attempt to settle the question. The central finding is that the question is not currently settleable from published evidence, and that the reason is specific and nameable: the two quantities that would decide it, measured accelerator utilisation by facility class and joules per delivered inference on a named embodied task, are not published by anybody.

2. Scope, method and evidence grading

The survey covers the siting of inference for embodied systems, with the centralised baseline characterised in enough detail to serve as a comparator. Training is treated only where the published evidence about it bears on the siting of inference, principally through fleet utilisation and through the energy accounting that mixes the two. Geographic coverage is United States centred, because the United States is

where facility-level energy, interconnection and tariff data are least incomplete, with European material included where it is primary and where it establishes a regulatory driver that does not exist in the United States. Prices, tariffs and queue figures are dated to 5 August 2026 and move.

Excluded from scope: model architecture and the question of which model should run at all; the allocation of compute between training and inference; data sovereignty, security and jurisdiction; land use, zoning and labour; and the substrate question, which is treated in a companion report [89] and appears here only where a measured substrate multiplier changes the arithmetic of section 8.

Every quantitative figure carries one of four grades. Verified means peer-reviewed, independently replicated, or drawn from a primary government or national laboratory series such as LBNL, EIA, IEA or EPA. Reported means a named analyst house, standards body, operator or preprint stated it without independent confirmation. Self-published means the party publishing the figure is the party selling or operating the thing being measured. Modelled means the figure was computed rather than observed, whether by a cited source or by this author.

Two conventions govern how grades compose, both taken from the Institute's prior methodological report [90]. Under the tie-break rule, a figure that originates with an interested party keeps the self-published grade regardless of which third party relays it, so a vendor number quoted by a trade publication remains self-published. Under the weakest-provenance rule, a figure composed from inputs of different grades takes the weakest of them, so a ratio computed from one verified measurement and one modelled assumption is modelled. The Institute's own prior reports are graded self-published under the same tie-break [89,90,91], because the alternative is to exempt the author's institution from a standard applied to everyone else.

Arithmetic performed for this review carries reference index 92 and is graded modelled. A modelled figure is never written as a measurement. Where a computation takes an input that is an assumption rather than a source, the assumption is named in the sentence that carries the result, and section 8 in particular is written so that the seven assumptions it rests on can each be replaced by a reader with a single multiplication. All working was executed and checked programmatically [92].

Grades are source criticism and not quality assessment. A verified figure can be wrong, and a self-published figure can be exactly right. What the grade records is

who was in a position to be wrong in a particular direction, and whether anyone independent has checked. Two consequences follow that are worth stating in advance. Bottom-up national models are graded verified when published by a national laboratory, and they remain models: LBNL states plainly that direct facility energy data are unavailable and that this limits its analysis [1]. And absence of a figure from a distributed operator is not evidence against distribution, because hyperscale operators disclose more than small operators do, so the evidence base is asymmetric in a direction that flatters the incumbent topology.

This review was assembled from public sources with AI assistance, following Institute convention. Several primary documents returned HTTP 403 to retrieval and are cited through relays; each such case is flagged in the reference entry, and section 10 lists them together. The weakest-provenance item in the entire report is the hyperscale construction cost per MW [42], and no conclusion rests on it.

3. The centralized baseline and the uncertainty in it

The most cited figures for centralised consumption are outputs of one bottom-up model. LBNL's congressionally mandated report reconstructs national data centre electricity from equipment shipment counts, analyst market data and assumed operating hours, arriving at 176 TWh and 4.4% of US electricity for 2023 against about 76 TWh and 1.9% for 2018 [1]. That pair reproduces the report's stated 18.3% compound annual growth rate exactly [92], which confirms the report's internal arithmetic and is not independent evidence of the growth rate, because both endpoints are outputs of the same model. LBNL states directly that direct facility energy data are unavailable and that the lack of data availability significantly limits the analysis [1]. There is no metered national total, and no federal agency maintains a facility register with location, capacity, metered load or water withdrawal; the nearest available list is a commercial directory of about 3,778 US facilities with a self-selected listing [16].

Two widely quoted capacity figures need to be read as what they are. Dividing the 2023 energy total by 8,760 hours gives 20.1 GW of average continuous draw [1,92]. That is an average, and an annual energy total cannot support anything else: it does not establish fleet peak, and it does not establish contribution to system peak, which depends on coincidence with local load. The forward figure of 74 to 132 GW for 2028 is LBNL's 325 to 580 TWh scenario range divided by 8,766 hours and then divided again by an assumed 50% average capacity utilisation, which the report

states as an assumption in its own executive summary [2]. Reproducing the derivation returns 74.2 and 132.3 GW, matching the published range. Re-running the same arithmetic at 70% utilisation instead returns 53.0 and 94.5 GW [2,92]. A twenty-point change in one quantity nobody measures moves the headline capacity number by roughly 28%, and that capacity number is the one that appears in interconnection and siting debate.

The utilisation inputs underneath are assumptions carried forward in part from LBNL's 2016 report: internal and small data centres at 11% operational time in 2014 rising linearly to 20% by 2027, colocation 21% to 35%, hyperscale 45% to 50%, AI servers doing training held constant at 80% and AI servers doing inference held constant at 40% across the entire projection period [2]. This is the single most load-bearing input in the national estimate. Operational time is also not the same quantity as FLOP utilisation: a processor can be busy 80% of the time and deliver 20% of peak throughput, and the two must not be conflated.

Two production measurements exist and neither settles the question. An NVIDIA-authored study of a fleet NVIDIA operates reports measured training model FLOPs utilisation averaging approximately 20% over a two-week window against an expected range of 35 to 50%, and states that only about 20% of fleet workloads had been onboarded to utilisation reporting at all, leaving 80% of GPU-hours with no measurement [18]. The finding runs against the authors' commercial interest, which is a reason to take it seriously and not a reason to treat it as independent. Separately, a measurement of 11,791 long-running jobs on an academic cluster attributes 19.7% of in-execution time and 10.7% of in-execution energy to execution-idle, meaning processors allocated to a running job and drawing power while doing no compute [19]. That bounds one waste channel on one cluster whose scheduling and tenancy differ materially from a hyperscale inference fleet.

Facility efficiency is in the same condition. LBNL's stock-weighted national average PUE of 1.4 for 2023 is a simulation output across weather stations and modelled facility configurations, and the report states that its simulated values assume systems are commissioned and operate as designed, that this is rarely the case, and that actual values will likely not be as good as estimated [2]. An operator survey reports an industry average of 1.56 [20], and one hyperscale operator self-reports a fleet-wide 1.09 [17]. These three numbers are not in conflict and should not be presented as though they were: the survey is an unweighted average of self-reported facilities that over-represents older and smaller sites, LBNL's is weighted by installed server stock which is concentrated in hyperscale and colocation, and the

third is one company's own campus fleet. Which is the right comparator for a siting calculation depends on which facility the compute would otherwise have run in.

The most granular production disclosure located reports a median text prompt at 0.24 Wh, 0.26 mL of water and 0.03 gCO2e, decomposed into 0.14 Wh of active accelerator, 0.06 Wh of host CPU and DRAM, 0.02 Wh of provisioned idle machines and 0.02 Wh of facility overhead [17]. Two ratios follow: 41.7% of delivered energy per unit of work sits outside the accelerator, and 8.3% is attributable to provisioned idle capacity [17,92]. The first is the number any efficiency argument has to beat, and it holds even at a self-reported PUE of 1.09. The second contradicts the common assertion that most data centre energy is wasted on idle machines. It is one vendor, one workload class, a median rather than a mean, self-reported, with training, storage and external network transfer explicitly outside the measurement boundary, and it should be read as one data point.

Projections diverge and their authors say why. EPRI puts US data centres at 9% to 17% of national electricity by 2030 and states that these projections are about 60% higher than its own analysis of eighteen months earlier, while warning that announced nominal capacity should be treated as a pipeline indicator rather than a near-term peak forecast [5]. The IEA projects global consumption rising from around 415 TWh in 2024 to roughly 945 TWh by 2030 [3]; a further IEA statement of a 50% surge in AI data centre electricity during 2025 could not be checked against the primary and no claim here rests on it [4]. Aggregate 2026 capital expenditure guidance of roughly $690bn to $725bn is a statement of intent by the companies doing the spending [7], and where a number must bear weight the audited filings should be used instead [6].

Interconnection queues are frequently quoted as demand and they are not. ERCOT was tracking approximately 474.7 GW of large-load interconnection requests as of June 2026, of which 90.2% were data centres, against an all-time system peak of 91,089 MW: the queue is 5.21 times the largest load the system has ever served [9,92]. No transmission provider publishes the duplicate or speculative fraction, so the queue cannot be converted into expected load. The nearest available analogue is from a different queue entirely: of generator interconnection requests made between 2000 and 2020, 19% of projects and 13% of capacity had reached commercial operation by end-2025, and the median project built in 2025 spent 61 months waiting [8]. That series covers generator interconnection only and explicitly excludes large loads, so it does not measure how long a data centre waits, and no national large-load series exists. Meanwhile PJM's 2027/2028 capacity auction

cleared at its administrative cap of $333.44/MW-day and fell 6,623 MW short of the reliability requirement, with nearly 5,100 MW of a 5,250 MW increase in the peak load forecast attributed to data centre demand [11]. A price at a cap reports scarcity without measuring how scarce.

The institutional response is running ahead of the measurement. A Texas directive of 3 August 2026 orders an audit of every data centre in the state interconnection queue and threatens to deny grid access to projects that fail to disclose ownership, financial, water and community-impact information, citing non-compliance with a mandatory state usage survey [10]. That some operators did not comply with a mandatory survey is itself the finding: disclosure is not currently enforced. Private trackers report more than 500 local moratorium instruments across 42 states, on definitions that mix binding ordinances, temporary permit pauses and bills that never passed, and no government body maintains an authoritative registry [15]. The direction is established and the number is not.

The baseline is therefore known to within a factor that matters for the comparison this review is about. The energy total is a model, the capacity figure divides that model by an unobserved assumption, the utilisation input underneath is an assumption carried across fifteen years, and the one production measurement of utilisation that exists reports a figure roughly half the low end of what the same authors expected. Any claim that distributed siting is worse or better than the centralised baseline inherits all of that uncertainty before it adds its own.

4. Latency requirements of embodied systems, and where the latency argument fails

Latency is the argument most often advanced for distributing computation toward the machine, and it is the argument with the best measured evidence behind it. It is also stated imprecisely more often than not, in a way that overstates the requirement by more than an order of magnitude for the workload that actually matters.

Peer-reviewed measurement on a commercial mmWave 5G network gives end-to-end round-trip times of 17.27 ms to a node inside the operator's network, 38.15 ms to a carrier-neutral metro node and 44.08 ms to a cloud region, with the radio access network component holding at 7.02 to 7.60 ms regardless of where the server sat [23]. The invariance of that radio term is the durable result; the absolute values are specific to one carrier and to 2022-23 deployments. Two ratios follow. Of

the best-case 17.27 ms round trip, 59.4% is transport and core, which is the portion a siting decision can attack. Moving a workload from a cloud region to the in-network node removes 60.8% of the round trip, or 26.81 ms [23,92]. That is a genuine reduction of roughly 2.5x and it is the strongest quantitative case the telco edge category has.

Treating an entire round trip as a control period gives an upper bound on closed-loop rate: 57.9 Hz over the in-network node, 26.2 Hz over the metro node, 22.7 Hz over the region, and an absolute ceiling of 142.4 Hz set by the radio alone even if the server sat at the base station [23,92]. Those figures allocate the whole period to network transit and give nothing to sensing, inference or actuation, so achievable rates are lower. What they establish is that a 1 kHz joint-torque loop and a 250 Hz visual servoing loop are outside the reach of every network-sited option measured, including the one physically inside the operator's network, and that the binding term is the radio rather than the distance to the facility.

Distance is not the problem. Light in single-mode fibre travels at about 2.0 x 10 ^ 8 m/s, giving 10 microseconds per kilometre of round trip. New York to the Northern Virginia campuses is about 330 km great circle, so even at a 1.6 route factor the pure propagation round trip is about 5.3 ms [92], roughly one ninth of the 47.69 ms P99 measured to a nearest cloud region on a production 5G network [25]. In that same production measurement, moving from the nearest cloud region to the operator's own multi-access edge platform bought 3.07 ms, from 47.69 ms to 44.62 ms, while a remote site with only 4G coverage measured about 660 ms to every region [25]. The same study found the cheapest hosting was the most distant region, with annual container cost of 4,292 NOK in a neighbouring country against 6,231 NOK in the nearest domestic region, so latency and price moved in opposite directions within one provider's own footprint [25].

Reachability from a dense edge estate is better than from cloud regions, and the gap narrows with the deadline. Measured from 8,456 vantage points, 55% of end-users reached a content-delivery edge server within 10 ms and 82% within 20 ms, against 27% and 62% for all five major cloud providers treated as one [24]. Those edge servers are caches with no accelerator, so the measurement establishes where a packet can reach and not where inference can run, and the probes sit disproportionately on wired connections, which flatters both figures relative to a mobile robot.

The requirement itself is where most of the confusion lies. Published vision-language-action policies do not run inference at the control rate. One open policy

reports image encoders at 14 ms, an observation forward pass at 32 ms and ten flow-matching steps at 27 ms, giving 73 ms on-board and 86 ms off-board including 13 ms of network, and it executes action chunks of horizon 50, re-running inference every 0.5 s for a 50 Hz robot after executing 25 of the chunked actions, and every 0.8 s for a 20 Hz robot [26]. The network consumes 15.1% of the off-board inference call and 2.6% of the 500 ms cadence, and the whole off-board call consumes 17.2% of it. A round trip of 100 ms, roughly eight times the measured one, would still leave 327 ms of headroom [26,92]. A second published humanoid foundation model shows the decoupling more starkly: a single forward pass takes about 160 ms against a 33 ms control period, so inference is already about five control periods long before any network is involved [27]. Quoting a robot's 50 Hz or 1 kHz control rate as the latency requirement for its policy inference is therefore wrong by more than an order of magnitude for this class of architecture.

On that evidence, a round trip to a regional facility can serve chunked policy inference, perception running at a comparable cadence, mapping, and fleet-level learning. It cannot serve a joint-torque or balance loop, which runs one to two orders of magnitude faster than any measured network path and is executed on-board in every published system located. It also cannot serve a safety-rated function, and this is where the latency argument fails in a way that milliseconds do not describe. Machine-safety practice converts reaction time into guarded distance at an approach speed constant of 2,000 mm/s [28], which is 2 mm of additional clearance per millisecond: 26 mm for a local wireless hop, 95 mm for a nearest-region round trip over 5G, and 1.32 m at a remote 4G site [28,92]. That arithmetic is not a statement about floor area. A compliant cell would not place a public network inside a safety-rated function at any latency, because such a function must be certified to a rated performance level with demonstrated diagnostic coverage, and a public network carries no such rating. The approach-speed formula in [28] does not itself impose that requirement; it sits in the functional-safety standards, ISO 13849-1 and IEC 62061, which this review does not survey and whose numeric requirements it therefore does not establish. The binding constraint on the safety loop is certifiability, and no reduction in round-trip time relaxes it.

One further limitation applies to every figure in this section. All of them are means or single percentiles from synthetic probes at fixed vantage points. What determines whether a control-relevant task can be offloaded is the tail and the outage rate, and no operator or fleet publishes the 99.9th or 99.99th percentile, the jitter, or the frequency and duration of link loss for a robot on a production

network. A 500 ms deadline with 327 ms of headroom is comfortable against a mean and says nothing about the hour in which the link drops.

Path	Measured round trip	Closed-loop ceiling (modelled, [92])	Grade	Source
Radio access network component alone	7.02 to 7.60 ms, invariant to where the server sits	142.4 Hz	verified	23
Device to node inside the operator network	17.27 ms (+/- 1.31)	57.9 Hz	verified	23
Device to carrier-neutral metro node	38.15 ms (+/- 1.83)	26.2 Hz	verified	23
Device to cloud region	44.08 ms (+/- 3.04)	22.7 Hz	verified	23
Production 5G, operator edge platform (P99)	44.62 ms	22.4 Hz	reported	25
Production 5G, nearest cloud region (P99)	47.69 ms, a 3.07 ms penalty against the operator platform	21.0 Hz	reported	25
Production 4G-only remote site	about 660 ms to all regions	1.5 Hz	reported	25
Fibre propagation floor, New York to Northern Virginia	about 5.3 ms at a 1.6 route factor	not applicable	modelled	92
Published policy, off-board inference call	86 ms total, of which 13 ms network	policy cadence 0.5 s against a 20 ms control period	self-published	26
		6.25 Hz	reported	27

Published humanoid model, single forward pass	about 160 ms against a 33 ms control period			

Table 1. Measured round trips by network tier, with the closed-loop rate each would permit if the entire period were allocated to network transit. The ceiling column is author's arithmetic throughout and is graded modelled [92]; the grade column applies to the measured round trip in the adjacent column. Ceilings are upper bounds and allocate nothing to sensing, inference or actuation.

5. A taxonomy of distributed topologies and the deployment record

Five topologies are distinguishable by the property that determines their reachability, their cost structure and their environmental accounting. A sixth, orbital compute, is included in the table because it is proposed in the literature, and it is a feasibility study rather than a deployment.

Telco edge places compute inside a mobile network operator's infrastructure. Its defining property is that the compute is reachable only by that operator's mobile subscribers, which constrains the addressable market before any technical question is reached. The largest such footprint located is 33 zone entries across seven carrier partners, of which 20 are one carrier's zones in the United States [29]. A second hyperscaler's equivalent product lists exactly three live sites and its documentation set is archived with a last content update of August 2024, which should be read as a product whose documentation is no longer maintained rather than as a confirmed shutdown [31]. An operator-consortium platform founded in 2018 was sold in April 2022 after its own website and social presence had been taken down [32]. No party publishes capacity in MW, utilisation, revenue or customer count for any of these, so scale reached can only be counted in sites, which is the metric most flattering to the category. Set against the radio estate the compute is meant to serve, 20 zones against 447,605 US cell sites is about one zone per 22,380 sites [29,91,92], an upper bound because the cell-site count covers all carriers. That ratio is consistent with the measured finding of section 4: telco edge compute sits at metro aggregation points, hundreds of cell sites behind the radio, which is exactly why the radio term does not move when the server moves.

The carrier-neutral metro tier is architecturally different and larger, at more than 30 metropolitan areas with seven announced, reachable over any network rather than one [30]. The relevant contrast with the telco tier is structural rather than numerical.

Micro and modular facilities place small containerised halls at aggregation points such as cell towers. The best-funded example planned 400 tower sites with a six-rack, 48 kW module, raised about $11 million, deployed roughly three sites and went into liquidation [33]. Had the full plan been built it would have totalled 19.2 MW of IT capacity, which is 1.6% of a single 1.2 GW AI campus under construction today [33,46,92]. That comparison does not establish that distributed capacity is uneconomic, because the two serve different workloads. It establishes that the distributed category never reached a scale at which its unit economics could be compared to hyperscale on equal terms. A surviving operator in the same class publishes a market map showing 8 markets live of 37 listed, with 26 described as customer ready within 90 days of an order, which is a sales statement rather than a deployment [34].

On-device inference places the computation on the machine. Vendor specifications give 2,070 FP4 TFLOPS sparse, or 1,035 dense, in a 40 to 130 W configurable envelope for the current physical-AI module family [58], and configurable 7 W and 15 W modes for the lower-power module family [59]. These are peak-throughput specifications at a precision few deployed models use, and no vendor publishes joules per policy inference on a named robot task, nor a module price on its own specification pages, which blocks both the environmental and the economic comparison at the same point.

Colocation at generation places the load next to the plant. The flagship case was rejected in its behind-the-meter form: FERC declined by 2-1 to accept the amended interconnection agreement that would have expanded a co-located nuclear-adjacent data centre load from 300 MW to 480 MW, finding the grid operator had not justified the non-standard provisions [43]. The parties restructured in June 2025 as a front-of-the-meter retail power purchase agreement of up to 1,920 MW running to 2042, explicitly to remove the need for FERC approval [44]. The physical adjacency stayed and the regulatory structure did not, which is the substantive point for siting analysis. In December 2025 FERC found the same grid operator's tariff unjust and unreasonable for lacking clear terms for co-located load and directed three service options into compliance filings, leaving rates to a paper hearing [45]. A related case is the fleet of roughly 425 modular units sited at wellheads on otherwise flared gas, sold in March 2025 while the operator redirected itself to a single 1.2 GW campus

[46]. The environmental credit claimed for that class of siting rests on a counterfactual: the operator states 99.9% combustion efficiency against a 91.1% flare average [47], and the independent airborne measurement that establishes the 91.1% baseline is peer-reviewed [48], so the baseline is better evidenced than the installation.

Heat-reuse siting places the facility where its waste heat can be sold. About 98% of the energy a data centre consumes leaves as heat [54], so the constraint is the existence of an offtaker within pipe distance and the temperature at which the heat is available, rather than the quantity of heat. The most developed market in the world connects more than 30 data centres across 16 providers and supplies about 3.5% of a capital city's heat against a 10% target, after roughly a decade [51], which is 35% of its own target [92]. At the price paid there, roughly EUR 170,000 per MW per year, heat sales offset something of the order of EUR 19.4 per MWh of electricity, on an ambiguity about whether the MW denotes heat delivered or IT load that the source does not resolve [51,92]. Regulation is beginning to make siting partly a heat-offtake decision: German law requires facilities above 300 kW entering operation from July 2026 to reach PUE at or below 1.2 and an energy reuse quota of 10%, rising to 15% in 2027 and 20% in 2028 [49]. Against a 98% heat fraction, a 20% reuse quota is a requirement to sell about 20.4% of the heat produced [49,54,92], which is not physically demanding on the heat side and is entirely demanding on the offtaker side. The EU framework above it requires reporting from 500 kW and sets no binding minimum [50]. Measured delivery is the missing quantity throughout: for the largest hyperscale recovery project located, three different annual figures circulate in public sources for what may be different scopes, all of them design capacities or contracted volumes rather than audited delivery [52], and the smallest-scale schemes have the thinnest measured record of any option surveyed [53]. A validated simulation of a single-rack building-integrated unit reports server air outlet temperatures of 35 to 45 degrees C [56], below the 68 degrees C supply temperature of the reference district heating network [54], so heat pumping is required and its electricity must be charged back against the recovery. The principal peer-reviewed synthesis on this topic was located but could not be retrieved, and no figure from it is asserted here [55].

The withdrawals belong in the record with their stated causes, because a taxonomy built only from surviving deployments overstates what has been demonstrated. The telco edge platform's failure is attributed by an edge-computing analyst house to company-specific factors rather than to the category, which is a self-interested reading and is treated as one, and no party disclosed investment, headcount,

revenue or operator deployments [32]. The micro data centre operator's liquidation has no published cause and its executives declined to comment [33]. The archived documentation set carries no formal retirement notice [31]. Two well-documented failures of distributed roadside infrastructure have causes on the record and neither is energy: the regulator that had designated a short-range vehicular radio standard twenty years earlier found it had not been meaningfully deployed and that the spectrum had largely been unused for decades, reallocated it, and set a sunset date, citing spectrum use, technology transition and deployment economics [82]; and a $42 million federal connected-vehicle pilot across three sites recorded its difficulties as procurement, interoperability and installation, on units that were grid- or pole-powered [83]. One survey reports that only 15% of operators rank the network far edge as the top location for future AI inference, from a paywalled source whose sample and question wording could not be established, and it carries no weight here [35].

Two structural observations follow from the table. First, where a cause is on the record it is commercial, procurement or regulatory; two of the located withdrawals carry no published cause at all. None of the located withdrawal accounts attributes the outcome to a physical limit. Second, the disclosure asymmetry is severe: the centralised topologies publish tariffs, filings and regulatory dockets, while the distributed topologies publish site counts and readiness claims. Absence of a distributed figure is therefore weak evidence about distributed performance.

Topology and instance	Defining property	Deployment record located	Grade	Source
Telco edge, hyperscaler A	Compute inside one operator's network, reachable only by that operator's mobile subscribers	33 zone entries across 7 carriers, 20 of them one carrier's US zones; no capacity, utilisation or customer count published	self-published	29
Telco edge, hyperscaler B	Same class, single carrier partner	3 live sites; documentation archived, last content update August 2024, no formal retirement notice located	self-published	31

Telco edge, operator consortium platform	Operator-founded platform intended to span member networks	Founded 2018, sold April 2022 after its website and social presence were taken down; code open-sourced	reported	32
Carrier-neutral metro	Metro facility reachable over any network	More than 30 metros available, 7 announced; no capacity or utilisation published	self-published	30
Micro and modular, tower-sited	Six-rack containerised module, 48 kW IT capacity	About 3 sites deployed against 400 planned; liquidation after about $11M raised	reported	33
Micro and modular, metro operator	Metro micro data centres across US markets	8 markets live of 37 listed; 26 described as customer ready within 90 days	self-published	34
On-device, current physical-AI module	Compute carried by the machine	2,070 FP4 TFLOPS sparse (1,035 dense) in 40–130 W; no price and no joules per policy inference published	self-published	58
On-device, lower-power module	Same class at a smaller envelope	Configurable 7 W and 15 W module power modes; vendor envelope, no measured workload figure	self-published	59
Colocation at generation, behind the meter	Load inside the plant fence	300 to 480 MW expansion rejected by FERC 2-1, November 2024, on inadequate justification of non-standard provisions	verified	43

Colocation at generation, restructured	Same adjacency, different regulatory structure	Restructured June 2025 as a front-of-the-meter power purchase agreement of up to 1,920 MW to 2042	reported	44
Colocation at generation, tariff status	Rules for co-located load in the largest US RTO	FERC found the tariff unjust and unreasonable, December 2025, directing three service options; rates left to a paper hearing	verified	45
Colocation at generation, wellhead	Modular units on otherwise flared gas	About 425 modular units and 270 MW sold March 2025; operator redirected to one 1.2 GW campus	reported	46
Heat-reuse siting, district heating market	Facility sited to sell heat into a network	More than 30 data centres supplying about 3.5% of one capital city's heat after roughly a decade, against a 10% target	self-published	51
Heat-reuse siting, regulatory driver	Siting made partly a heat-offtake decision by statute	Germany: PUE at or below 1.2 and reuse of 10% (2026), 15% (2027), 20% (2028), above 300 kW	verified	49
Heat-reuse siting, building-integrated	Single rack inside an occupied building	Simulated 12 kW rack, 35–45 degrees C air outlet, below district heating supply temperature; no kWh, CO2 or reuse factor reported	modelled	56
Orbital	Compute in dawn-dusk sun-	Feasibility study; two prototype satellites planned by early 2027; cost	self-published	57

	synchronous orbit	parity conditional on a launch price that does not exist		
Withdrawn: short-range vehicular radio estate	Distributed roadside radios under a designated national standard	Regulator found it had not been meaningfully deployed, reallocated the spectrum, set a sunset date; causes stated as spectrum use, technology transition and deployment economics	verified	82
Withdrawn: federal connected-vehicle pilots	Grid- and pole-powered roadside units at three sites	$42 million across three sites from 2015; documented difficulties were procurement, interoperability and installation	verified	83

Table 2. Topologies for siting computation that serves embodied systems, with the deployment record located for each. One row per item so that each row's grade matches the single source it cites. Counts of sites are the metric the category itself publishes and are the least informative available; no party in the distributed rows publishes capacity in MW, utilisation or customer count.

6. Environmental accounting: embodied against operational, and the utilisation that decides it

The environmental comparison between centralised and distributed siting turns on one structural fact and three accounting choices. The structural fact is that operational carbon scales with electricity consumed while embodied carbon is paid once at manufacture and amortised over whatever work the hardware performs, so the balance between them is set by utilisation and by the carbon intensity of the electricity, and not by any property of the hardware itself.

The clearest demonstration available uses one published server footprint. A vendor life-cycle assessment for a current rack server reports manufacturing at 638, transport at 177, end of life at 25 and use at 2,583 kgCO2e over a four-year life at 4,844.28 kWh per year, totalling 3,424 kgCO2e [66]. Embodied terms sum to 840 kgCO2e, which is 24.5% of the published total, and the use-phase figure implies a

grid intensity of 133 gCO2e/kWh behind the vendor's own assumption [66,92]. Substituting the US national average of 373 gCO2/kWh at a PUE of 1.0 moves the embodied share to 10.4%; adding LBNL's modelled national PUE of 1.4 moves it to 7.7%; and substituting a 50 gCO2e/kWh grid at PUE 1.0 moves it to 46.4% [2,22,66,92]. The crossover, where embodied carbon equals operational carbon for this server, falls at a grid intensity of about 43 gCO2e/kWh [66,92]. The embodied share of a server is therefore a property of the grid it is plugged into, swinging by a factor of six across grids that exist today, and it rises as grids decarbonise. Two cautions attach. The vendor's own uncertainty band on the total runs from 1,976 to 23,261 kgCO2e, wider than the central estimate [66], so every share computed from it inherits that band. And this configuration carries no accelerator, so it bounds the general shape of the relationship and not the magnitude for AI hardware.

For the distributed side the most directly relevant peer-reviewed comparison measured per-inference energy for six generative models on a handset NPU against a data centre accelerator, finding 262.26 J against 3,803.71 J for a 7B language model, a 93% reduction, with similar margins on image and text-recognition models [61]. The comparison is not equal-accuracy and not equal-throughput: the device runs 4-bit and 8-bit quantised models while the cloud runs unquantised or differently quantised models without batching, which removes the largest efficiency lever a data centre has, and the authors say so. Cloud latency is lower in every row, so the energy advantage is bought with time.

The life-cycle outputs from that study are modelled rather than measured, and this review uses only the single row for which every input is published, so that the arithmetic is re-derivable. Per 1,000 queries of the 7B model, the cloud accelerator total is 484.29 gCO2, of which 412.07 is operational and 72.22 embodied, and the handset total is 75.45 gCO2, of which 28.41 is operational and 47.04 embodied [62]. Embodied carbon is therefore 14.9% of the cloud total and 62.3% of the device total, a ratio of 4.2 [62,92]. That is the low-utilisation multiplier made numerical inside one consistent case, and it confirms that the distributed side is an embodied-carbon problem while the centralised side is an operational-energy problem. It does not establish that the edge is worse overall, because the edge total in the same row is 84% lower. It also depends on the authors setting cloud utilisation to 1.0, which is above every utilisation figure LBNL models for any facility class [2], and which therefore understates the cloud embodied share.

The obvious objection is that the device is idle most of the time. The same row answers it. Because the modelled embodied term scales as the reciprocal of

utilisation while the operational term does not, the product of embodied carbon and utilisation is invariant, and the break-even is reached when 28.41 plus 8.9376 divided by utilisation equals 484.29. That gives a utilisation of 1.96%, about 28 minutes of active use per day [62,92]. On this dataset, for this workload and this hardware, the device would have to be active less than 2% of the time before the low-utilisation penalty erased a 93% operational advantage. The result does not extend to a device built solely to host an accelerator, where the embodied term is large relative to the compute delivered and the same algebra runs the other way.

Comparing published utilisation figures directly gives a smaller gap than the surrounding rhetoric suggests. Against LBNL's modelled operational times, a handset at 0.19 carries a 2.1x embodied penalty relative to an AI inference server at 0.40, 4.2x relative to a training server at 0.80, 1.8x relative to a colocation server at 0.35, and 1.05x relative to a server in a small on-premises facility at 0.20 [2,62,92]. This is an order-of-magnitude bound and should be read as one: LBNL's operational time counts hours of active operation while the handset figure counts screen-on activity, the two constructs are not identical, and no source publishes both on a common definition. Two further studies bear on the same point from opposite directions. A workshop study finds edge device carbon dominated by embodied terms at over 80% for mobile devices, and models an 8x net carbon reduction from offloading one accelerator's work to 69 smartphones, falling to 6x with communication carbon included; that result rests entirely on assuming the devices would have been bought anyway, and the paper is explicit that the spare capacity is treated as free [63]. And the counter-case is the cleanest statement of the penalty in the literature: amortising the manufacturing footprint of a 2018 handset requires running a small vision model continuously for three years, beyond the device's typical lifetime [60].

The first accounting choice is marginal against average emission factors. The regulator's own equivalencies calculator uses two different national numbers for two different questions: an average output rate of 823.1 lb CO2/MWh, which converts to 373 gCO2/kWh, for emissions attributable to consumption, and a marginal rate of 603 gCO2/kWh for the avoided emissions of a demand-side change [22]. The marginal figure is 1.62 times the average, or 1.99 times on an earlier revision of the same calculator [22,92]. Siting a new load is a marginal question, and every published comparison located in this literature substitutes an average factor, including the one this review relies on most [62]. No regulator publishes a marginal factor shaped to a continuous, flat, always-on load at balancing-authority resolution, so the correct factor for a siting decision does not exist in any public

series. The gap between the two is larger than most of the efficiency differences being argued about.

The second is attribution. LBNL assigns every facility the annual average of its balancing authority and states that it incorporates no power purchase agreements and no behind-the-meter generation, because facility-level data do not exist [1]. A facility-level study of 403 hyperscale facilities on the same locational basis puts their electricity-weighted carbon intensity at approximately 545 gCO2/kWh against a contemporaneous national average of 370 [21], which is opposite in sign to LBNL's 2023 finding that the data-centre-weighted factor sat marginally below the US average. The two differ in period, sample and attribution method, and the divergence is recorded here as a measure of how unsettled the attribution question is rather than asserted as a contradiction.

The third is water, where the aggregate answer is unambiguous and the local answer is not. Direct on-site water consumption by US data centres was 66 billion litres in 2023 against nearly 800 billion litres consumed indirectly at the power plants supplying them, so on-site water is 7.6% of the total attributable footprint and the grid carries 92.4% [1,92]. Distribution can therefore remove at most 7.6% of the water, and only if the workload does not consume more electricity elsewhere. Everything else follows the electricity wherever it is consumed. A consistency check on the source passes: 66 billion litres over 176 TWh implies 0.375 L/kWh, just above LBNL's stated national site WUE of just over 0.36 [1,92]. Recomputing the peer-reviewed edge comparison's 95% water saving with LBNL's national factors instead of the study's own returns 93.6% [1,61,92], which shows that the saving is driven almost entirely by the 93% energy reduction and not by the absence of cooling water: the on-site term was never load-bearing. None of this addresses local scarcity, which is a distributional question about where the 7.6% falls. One legislature-commissioned study found most data centres in its state using about as much water as an average large office building or less and current use sustainable though poorly managed across competing local uses [12], while a journalistic analysis reports roughly two-thirds of facilities built since 2022 sited in water-stressed regions on a threshold judgement whose method was not inspected [13]. One operator reports 10.9 billion gallons consumed in 2025 with 78% replenished [14], and replenishment is not the same as returning water to the watershed that supplied it.

Three gaps close off the comparison rather than narrowing it. Embodied carbon for AI accelerators is unreconciled: the only accelerator-level figure located is

approximately 1,312 kgCO2e per unit published by the party selling the accelerator [67], while the peer-reviewed comparison used above independently assumes 518 kgCO2 per cloud GPU [62], a factor of 2.5 apart, and no independent life-cycle assessment of an accelerator has been published. The unit of account on the distributed side is unstable: production carbon between simple and complex edge devices varies by more than a factor of 150 [65], and among five off-the-shelf microcontroller boards it ranges from 0.52 to 2.59 kgCO2e, with the most energy-efficient board failing to be the carbon-optimal one for short deployments [64]. And the accounting convention itself has no home for the problem: the software carbon intensity specification amortises embodied carbon by time-share and resource-share, so hardware time that is idle is charged to nobody, which means the embodied carbon that low-duty-cycle distributed hardware fails to amortise disappears from every ledger built on that convention [69]. End of life is similarly unresolved: global e-waste reached 62 million tonnes in 2022 with 22.3% documented as formally collected, and the authoritative monitor does not break out accelerators, servers or on-device AI hardware as a category [68], so no distributed-versus-rack comparison can be made from it. Finally, network transport energy per delivered inference is outside the measurement boundary of every source located, including the most detailed production disclosure [17] and the most relevant life-cycle comparison [62]. That is the term a siting argument most needs and the term nobody publishes.

Quantity	Value	Grade	Source
Rack server life-cycle phases, as published	Manufacturing 638, transport 177, end of life 25, use 2,583 kgCO2e over 4.0 years at 4,844.28 kWh/yr	self-published	66
Vendor's own uncertainty band on that total	3,424 kgCO2e central; 1,976 at the 5th percentile and 23,261 at the 95th	self-published	66
Grid intensity implied by the vendor's use phase	133 gCO2e/kWh	modelled	66,92
Embodied share at that implied intensity	24.5%	modelled	66,92

Embodied share at the US average grid, PUE 1.0	10.4%	modelled	22,66,92
Embodied share at the US average grid, PUE 1.4	7.7%	modelled	2,22,66,92
Embodied share at a 50 gCO2e/kWh grid, PUE 1.0	46.4%	modelled	66,92
Grid intensity at which embodied equals operational	About 43 gCO2e/kWh	modelled	66,92
Measured energy per inference, 7B model, device against cloud	262.26 J against 3,803.71 J, a 93% reduction, not equal-accuracy and unbatched on the cloud side	verified	61
Modelled life-cycle carbon per 1,000 queries, cloud accelerator	484.29 gCO2 total = 412.07 operational + 72.22 embodied; embodied share 14.9%	modelled	62,92
Modelled life-cycle carbon per 1,000 queries, handset NPU	75.45 gCO2 total = 28.41 operational + 47.04 embodied; embodied share 62.3%	modelled	62,92
Ratio of the two embodied shares	4.2	modelled	62,92
Device utilisation at which its life-cycle carbon per query equals the cloud's	1.96%, about 28 minutes per day	modelled	62,92
Embodied penalty from utilisation alone, handset against server classes	2.1x against AI inference at 40%, 4.2x against training at 80%, 1.05x against a small on-premises server at 20%	modelled	2,62,92
	603 against 373 gCO2/kWh, a ratio of 1.62; 1.99 on	modelled	22,92

Marginal against average US grid emission factor	an earlier calculator revision		
Direct share of the US data centre water footprint	7.6% on site, 92.4% at the power plant	modelled	1,92
Water saving of the same edge comparison under LBNL factors	93.6% against 95% as published	modelled	1,61,92
Hyperscale electricity-weighted carbon intensity, 403 facilities	About 545 gCO2/kWh against 370 national, on locational attribution	modelled	21
Spread in production carbon between simple and complex edge devices	More than 150x	verified	65
Accelerator embodied carbon in circulation	About 1,312 kgCO2e per unit against 518 kgCO2 assumed elsewhere, unreconciled	self-published	67
Network transport energy per delivered inference	Outside the measurement boundary of every source located	modelled	17,62,92

Table 3. Environmental accounting for the siting comparison. Rows computed for this review are graded modelled and carry index 92 alongside the sources of their inputs, under the weakest-provenance rule stated in section 2. The single 7B-model row is used for the embodied-share arithmetic because it is the case for which the source publishes every input.

7. Cost structure of the siting choice, and the denominators that are not published

One operator publishes a machine-readable tariff that prices the siting choice directly, holding instance type and billing model constant so that location is the only variable. In a New York metro zone, four unrelated CPU instance families each carry exactly a 25.00% premium over the parent region and two GPU families each carry 34.97%. In a Los Angeles metro zone, seven instance families carry 19.8% to 20.0% and one GPU family carries 37.38%. Inside a carrier's network in New York,

CPU instances carry 34.6% to 34.9% and one GPU instance carries 75.13% [36,92]. The uniformity is the informative part: a multiplier of exactly 1.2500 applied across four unrelated instance families is an administered price rather than a cost pass-through, so these figures price what one operator charges for edge siting and not what edge siting costs it. The two metro zones carry different multipliers for the same instance families and the operator publishes no reason. Only four instance types are offered in the carrier zone at all [36], which suggests availability may bind before price does.

Savings figures for distributed and decentralised compute are denominated in a single unit, the hyperscaler on-demand list price, and range from 45% to 90% [88]. That denominator is traceable and re-extractable: the eight-accelerator instance lists at $55.04 per hour, which is $6.88 per accelerator-hour [36,92]. The same operator discounts the same instance by up to 62.40% under its own published three-year commitment, reaching an effective $2.5869 per accelerator-hour [36,92]. A claim of 60% off list therefore implies $2.7520, which lands 6.38% above the operator's own committed price [36,88,92]. A distributed alternative priced against on-demand rates is measured against a rate 62.40% above the incumbent's own published three-year committed rate, before any account is taken of the cost of capital tied up in a $543,866 prepayment, of negotiated enterprise discounts the operator does not publish at all, or of the networking, storage, availability and support the incumbent bundles.

The utilisation dependence of amortised distributed hardware is the strongest economic argument against distribution in this survey, and it rests on an unmeasured input. With identical hardware and identical service life, capital cost per unit of delivered work scales as the reciprocal of utilisation. Against LBNL's modelled operational times, a small facility at 20% must recover its capital over 2.5 times fewer productive hours than a hyperscale facility at 50%, a 150% penalty, and a colocation facility at 35% carries 42.9% [2,92]. The 150% penalty is four to seven times larger than the 20% to 35% premium the tariff above charges for metro edge siting. Two qualifications are heavy enough to reverse the reading. LBNL's figures are model assumptions rather than measurements of any fleet, and they describe conventional servers. For accelerators, LBNL assumes a flat 40% inference operational time regardless of siting [2], and if that assumption is right then siting does not change accelerator utilisation at all and the penalty vanishes entirely. Nobody publishes measured accelerator utilisation by facility class, so the question cannot currently be settled in either direction.

For owned on-device hardware the same trade appears in a different form. At an assumed device price of $3,000, which is this author's assumption and not a sourced figure because no vendor publishes a module price on its own specification pages [58,59], a three-year service life of 26,280 hours breaks even against a rented single-accelerator instance at $0.8048 per hour at a duty cycle of 14.2%, or about 3.4 hours a day every day [36,92]. That establishes the shape of the trade rather than its location, and two omitted effects matter in opposite directions. Against ownership, one rented instance can be time-shared across a fleet, so its effective utilisation is the fleet aggregate rather than any one machine's, and fifty machines at 10% duty each fill a single instance. For ownership, the machine must carry a computer anyway for its safety and control loops, so the marginal cost of using it for policy inference is far below the full device price. A serious version of this calculation requires both terms and a published device price, and neither exists.

Transport pricing does not support the argument usually built on it. Data transfer into a hyperscale region is $0.0000 per GB at published tariff [37], so the backhaul-cost argument for edge inference cannot be made at the cloud boundary: sending sensor data up costs nothing there. Egress runs $0.090 per GB for the first 10 TB per month falling to $0.050 above 150 TB [37], which is 583 times and 324 times a theoretical wholesale transit floor derived from $0.05 per Mbps per month for 100 GigE in the most competitive markets [37,38,92]. That comparison is deliberately unfair in one direction and the unfairness must be stated: it assumes a transit port filled 100% of the time, which no buyer achieves, and real 95th-percentile billing on a partly filled port typically yields an effective cost several times the floor, so a fairer ratio is on the order of 81 to 146 times. What the comparison establishes is that cloud egress pricing is a commercial tariff and not a measurement of what transport costs. The cost that actually bears on an embodied system sits on the last mile, where published retail cellular data runs $0.03 per MB, which is $30 per GB [39], and where no carrier publishes the rate a high-volume machine connection would pay.

Workload shape can flip the sign of the comparison entirely. Comparing published list prices for the same open model on a distributed network and on a centralised provider, the distributed network charges $0.293 per million input tokens and $2.253 per million output, against $1.04 for both at the centralised provider. The break-even falls at 0.616 output tokens per input token: below that the distributed network is cheaper at list, above it the centralised provider is [40,92]. Embodied workloads are input-heavy, with images and proprioception in and a short action chunk out, which places them on the side where the distributed list price wins. This compares prices and not costs of production, the two serving stacks are not

equivalent, and the distributed provider does not state where its inference physically runs, so the result cannot be attributed to siting.

Several denominators that would settle these questions are not published by anyone. There is no capital cost per kW for a metro edge site drawn on the same accounting boundary as hyperscale construction, and the hyperscale figure itself, roughly $10.7 million to $12.0 million per MW, is the weakest-provenance item in this review and should be treated as an order of magnitude only [42]. There is no published electricity tariff spread between a hyperscale campus on a negotiated industrial contract and a small distributed cabinet on a commercial account; national averages of 8.71 cents per kWh industrial against 13.54 cents commercial [41] bracket the question without answering it. There is no published per-GB carriage price for a machine fleet, no published device price, no published measured accelerator utilisation, and no published cost of production for an inference token from any provider at any tier, so every dollar-per-token comparison in circulation compares margins as much as it compares physics.

Finally, the grid-side externalities of siting are being litigated rather than measured. A capacity auction cleared at its administrative cap with nearly 5,100 MW of a 5,250 MW load-forecast increase attributed to data centres [11]. A legislature-commissioned study modelled residential generation and transmission costs rising by $14 to $37 per month in constant dollars by 2040 while simultaneously finding that data centres currently pay their full cost of service and that current rates allocate costs appropriately [12], so the cost shift in that study is a forward risk rather than a present finding. And FERC left the rates for co-located load to a paper hearing precisely because whether such load shifts cost onto other ratepayers is not established [45]. None of these is a measurement of the siting choice's cost, and each is evidence that the accounting for it does not yet exist.

Item	Published or derived figure	Grade	Source
On-demand list, eight-accelerator instance, parent region	$55.04 per hour, which is $6.88 per accelerator-hour	reported	36
Same operator's three-year All Upfront commitment	$543,866, an effective $20.6951 per hour, $2.5869 per accelerator-hour	reported	36

That commitment as a discount on the same operator's own list	62.40%	modelled	36,92
Savings claimed against on-demand list by distributed compute networks	45% to 90%	self-published	88
Where a 60%-off-list claim lands against the incumbent's committed price	6.38% above it	modelled	36,88,92
Metro edge premium, New York zone	+25.00% on four CPU families, +34.97% on two GPU families	modelled	36,92
Metro edge premium, Los Angeles zone	+19.8% to +20.0% CPU, +37.38% on one GPU family	modelled	36,92
Telco edge premium, New York carrier zone	+34.6% to +34.9% CPU, +75.13% on one GPU family; only four instance types offered	modelled	36,92
Cloud ingress from external networks	$0.0000 per GB	reported	37
Cloud egress to the internet	$0.090/GB for the first 10 TB per month, $0.050/GB above 150 TB	reported	37
Wholesale IP transit, most competitive markets, Q2 2025	$0.05 per Mbps per month at 100 GigE	reported	38
Egress tariff against a theoretical fully-filled transit port	324x to 583x; on the order of 81x to 146x at realistic port fill	modelled	37,38,92
Retail cellular data, self-service tier	$0.03 per MB, which is $30 per GB	reported	39
Capital penalty from utilisation, small facility against hyperscale	2.5x, a 150% penalty; 1.43x for colocation	modelled	2,92

Break-even duty cycle for an owned device at an assumed $3,000 price	14.2%, about 3.4 hours a day	modelled	36,92
Output-to-input token ratio at which the distributed list price stops winning	0.616	modelled	40,92
Hyperscale construction cost	Roughly $10.7M to $12.0M per MW; weakest-provenance item in this review	reported	42
US average retail electricity, May 2026	13.83 cents/kWh all sectors, 8.71 industrial, 13.54 commercial	verified	41

Table 4. Published prices bearing on the siting choice, and the figures derived from them. Derived rows are graded modelled under the weakest-provenance rule and carry index 92. Every price is a list or tariff price rather than a transaction price, and no party publishes a cost of production at any tier.

8. Physical limits of ambient-powered computation at small apertures

The most distributed end of the range is a device with no wire, powered only by energy harvested from its immediate environment. The question this section answers is general and physical: what computational duty can such a device sustain, given measured harvesting yields and measured per-inference energy, at the aperture a small fixed module actually presents? Two generic form factors are used, a module embedded in a pavement surface and a module mounted on a pole, and the analysis is about the physics of the aperture rather than about any product.

Step 1, the aperture. The pavement-embedded case is taken at 100 cm2, or 0.01 m2, and the pole-mounted plate at 0.02 m2. Both are this author's assumptions and are graded modelled [92]. They are stated explicitly because every shortfall figure below divides by them, and because the result scales linearly in aperture: a reader who substitutes a different area rescales every harvest figure and every shortfall by one multiplication.

Step 2, yield for a trafficked pavement surface. Three field installations give measured yields, and all three were eventually removed or shut down. A tile

installation generated 52.397 kWh in six months across 13.9 m2, an annualised 7.54 kWh/m2/yr or 0.861 W/m2 of mean power [70,92]; against its 1.529 kW rated capacity that is a capacity factor of 0.78% [70,92], which is the arithmetic signature of an installation in which roughly 75% of panels were broken before installation and 25 of 30 malfunctioned in the first week. A cycle path installation reported first-year specific yields of 73 and 93 kWh/m2/yr for two versions, giving up to 10.62 W/m2, declining to about 41 kWh/m2/yr or 4.68 W/m2 as the top layer's light transmission degraded [72,92]. A carriageway installation designed for 790 kWh/day delivered about half of that before part of the road was demolished for wear damage [71]. The working band used below is 0.861 to 10.62 W/m2, with 0.861 read as a floor and 10.62 as a first-year best case that degraded in every installation measured. General soiling literature gives 3% to 5% of annual production for tilted, cleanable arrays [73] and contains no figure for a horizontal, trafficked, tyre-abraded surface, which would be expected to be worse.

Step 3, the pole-mounted case is a different yield class and this survey cannot bound it. A pole-mounted plate is not a trafficked surface: it suffers neither tyre abrasion nor the degradation of a load-bearing transparent top layer, which were the two mechanisms that drove the declines above. It also carries an unfavourable orientation, roadside shading and soiling for which no measured yield at small aperture was located. The honest position is that the trafficked-surface band above must not be applied to it, that its true yield lies somewhere above that band and below a conventional fixed array, and that the gap is unmeasured. The arithmetic that follows is therefore stated for the pavement-embedded case only, and the pole-mounted case appears in the open problems.

Step 4, harvest at aperture. At 0.01 m2, the photovoltaic band gives 8.61 to 106.2 mW of mean power [70,72,92].

Step 5, the other candidate sources on the same 100 cm2 basis, each scaled from a measured device output. A thermoelectric generator sustaining about 10 mW for 8 hours a day from a 40.96 cm2 module gives 8.14 mW as a 24-hour average when scaled to 100 cm2, and it requires a hot side in the pavement and a cold side in air or flowing water, which a sealed pavement-embedded module does not have because it is close to isothermal with the pavement [76,92]. A hygroelectric film averaging 3.0 microwatts/cm2 over three months outdoors gives 300 microwatts [78,92]. A pavement-embedded piezoelectric harvester producing about 50 mWh per month from 16 transducers across a 0.16 m2 footprint under real traffic gives 4.28 microwatts [74,92]; the same literature reports 2,381 mW of peak instantaneous

output under a passing axle [75], which is a different quantity from a time average and cannot be substituted for one, and a vendor pilot reporting lane-scale figures without a time base cannot be converted to a power density at all [87]. Ambient radio frequency at the best measured outdoor density of -7 dBm/m2 gives 1.76 microwatts incident on a modelled effective antenna aperture of 0.00883 m2 at 900 MHz, or 0.53 microwatts after an assumed 30% rectifier efficiency, both of which are this author's assumptions [77,92]. Vehicle-induced airflow reaches useful power only at turbine scale, up to 48 W from a swept area on the order of a square metre [79], which is two orders of magnitude larger than the apertures considered here and is recorded to mark the boundary rather than to be included.

Step 6, these sources are not additive. They compete for one exposed surface: a photovoltaic cell, a thermoelectric cold-side exchanger, a hygroelectric film and an antenna cannot all occupy the same 100 cm2 without taking area from one another. A sum is therefore an upper bound and never an operating point. That upper bound runs from 17.1 mW to 114.6 mW including a thermoelectric with a working cold side [92]. Photovoltaics alone supply 96.6% to 99.7% of the total excluding the thermoelectric, and 50.5% to 92.6% including it [92]. Photovoltaics dominate the sum at every point in the band except its lowest, where a thermoelectric with a working cold side is comparable. A sealed pavement module has no such cold side, so for that form factor the harvest is a solar budget with small correction terms, and the correction terms do not change the conclusion in either direction.

Step 7, duty against measured per-inference energy at the tinyML tier. Under a common energy harness, a visual wake words inference completed in 22.2 microjoules and a keyword-spotting inference in 35 microjoules [80], with a vendor separately reporting an always-on hotword detector under 280 microwatts at a 3.3% duty cycle [81]. At raw harvest the 22.2 microjoule figure supports 388 to 4,784 inferences per second, and the 35 microjoule figure 246 to 3,034 [70,72,80,92]. Applying a deliberately harsh 90% system derate for power conversion, storage round trip, leakage, housekeeping and radio, which is this author's assumption and not a measurement, leaves 39 to 478 inferences per second [92]. The envelope closes for this class of work: 39 to 478 inferences per second after the derate, against the one inference per second that the coin-cell comparison in [80] uses as its reference duty, on commodity silicon measured today.

Step 8, duty against an embedded inference module. Taking 7 W and 15 W, the configurable module power modes of a current low-power embedded family [59], continuous operation requires 0.66 to 8.13 m2 of collector at 7 W and 1.41 to 17.4

m2 at 15 W [59,70,72,92]. Against a 0.01 m2 aperture that is a shortfall of 66x to 813x at 7 W and 141x to 1,740x at 15 W [92]. No pavement-embedded module at this aperture can continuously power an embedded inference module of this class from a trafficked road surface, at any measured yield, by a wide margin.

Step 9, intermittent operation with a buffer, which is the reading the continuous shortfall obscures. A module with an energy store need not run continuously; it can accumulate and burst. The sustainable duty fraction is simply harvested power divided by load power: 0.12% to 1.52% at 7 W and 0.057% to 0.71% at 15 W [59,70,72,92]. In operating terms that is 4.4 to 54.6 seconds of full-power operation per hour at 7 W, or one second of operation every 66 seconds at the best measured yield and every 13.6 minutes at the worst [92]. The reciprocal of the duty fraction and the continuous shortfall factor are the same number, which is worth stating because the two are often presented as different findings. The buffer must hold at least one burst's energy, and its round-trip efficiency, leakage and cold-start cost sit inside the 90% derate assumed above rather than being measured.

Step 10, season. Annual-mean yields overstate winter availability substantially, and no seasonal series for any road-surface photovoltaic installation was retrieved, so a winter floor cannot be computed from the measured record. The only seasonal measurement located anywhere in this survey is thermal rather than solar: a road thermoelectric field system peaked at 162.33 mW in summer against 34.63 mW in winter, a ratio of 4.7 [76,92]. Applying an explicitly assumed winter-to-annual-mean ratio of one third, which is this author's assumption and the weakest step in this section, gives a winter harvest of 2.87 to 35.4 mW, a tinyML rate of 13 to 160 inferences per second after the 90% derate, and a 7 W duty fraction of 0.041% to 0.51% [92]. If the true ratio is one fifth rather than one third, every winter figure falls by a further 40%. The tinyML tier survives the winter floor with headroom; the embedded-module tier does not change character, because it was already intermittent.

What the envelope admits, at a 0.01 m2 pavement-embedded aperture: continuous inference at the tinyML tier, at 39 to 478 inferences per second on the annual mean and 13 to 160 at the assumed winter floor. What it excludes: continuous operation of a 7 to 15 W embedded inference module, by two to three orders of magnitude in aperture. What it admits conditionally: the same module operated in bursts against a buffer, at a duty fraction between roughly one part in a thousand and one part in seventy. Restating the aperture dependence, continuous 7 W operation at the best

measured yield would require 0.66 m2, which is 66 times the assumed aperture and remains far above any small fixed module.

Three limits on the reading. First, the tasks the tinyML figures cover are small: 96x96 grayscale person detection, ten-keyword spotting, small-image classification and anomaly detection [80]. None is a perception or control workload of the kind an embodied system needs, and no comparable measured per-inference energy series exists for the tier above, so this bound describes what ambient power admits at the sensing tier and is silent about the tier between sensing and an embedded module. Second, none of this covers the radio. Every figure here is an inference-energy figure, and for a module that must also communicate, the radio is the term most likely to dominate; no measured energy per delivered bit for such a link at these duty cycles was located. Third, alternative substrates move the aperture requirement by their measured multiplier and no more. Reversible logic has demonstrated energy-recovery factors of 1.77 and 1.41 in 22 nm silicon, which is 43.5% and 29.1% of energy saved, and its recovery improves only by slowing the clock [84,92]. Neuromorphic spiking shows measured advantages spanning roughly 2x to 145x, strongly workload-dependent, clustering at 2x to 20x on dense work, with the largest figures on temporally sparse event-driven input [85], and with a measured accuracy cost of 2.4 percentage points in one published comparison for a 67% energy reduction [86]. A substrate multiplier m divides the aperture requirement by m: closing a 66x gap would require the top of the neuromorphic range on a workload suited to it, and closing an 813x gap exceeds every measured substrate result located. Every neuromorphic figure cited is also core-level, with no sensor, analog-to-digital conversion, radio or storage inside the measurement [85], which is the composed-path omission the Institute's methodological report identifies as the most consequential in this area [90].

Step	Input or result	Grade	Source
Aperture, pavement-embedded module	100 cm2 (0.01 m2); author's assumption, stated because every shortfall divides by it	modelled	92
Aperture, pole-mounted plate	0.02 m2 assumed; yield class not established by any measurement located	modelled	92
Measured trafficked-surface photovoltaic		modelled	70,92

yield, weakest installation	7.54 kWh/m2/yr, 0.861 W/m2 mean, a 0.78% capacity factor against rated		
Measured trafficked-surface yield, best first year	93 kWh/m2/yr, 10.62 W/m2 mean	modelled	72,92
Same installation after top-layer degradation	41 kWh/m2/yr, 4.68 W/m2	modelled	72,92
Photovoltaic harvest at 0.01 m2	8.61 to 106.2 mW	modelled	70,72,92
Thermoelectric at 0.01 m2, cold side required	8.14 mW; unavailable to a sealed pavement module	modelled	76,92
Hygroelectric at 0.01 m2	300 microwatts	modelled	78,92
Piezoelectric at 0.01 m2, field-measured device	4.28 microwatts	modelled	74,92
Ambient radio frequency at 0.01 m2, assumed 30% rectifier efficiency	0.53 microwatts	modelled	77,92
Sum of all sources, upper bound only since they compete for one surface	17.1 to 114.6 mW	modelled	92
Photovoltaic share of that sum	96.6% to 99.7% without a thermoelectric; 50.5% to 92.6% with one	modelled	92
Measured energy per tinyML inference	22.2 microjoules (visual wake words), 35 microjoules (keyword spotting)	reported	80
Sustainable tinyML rate at raw harvest	388 to 4,784 per second at 22.2 microjoules	modelled	70,72,80,92
Same after an assumed 90% system derate	39 to 478 per second	modelled	92

Continuous aperture required, 7 W embedded module	0.66 to 8.13 m2, a 66x to 813x shortfall at 0.01 m2	modelled	59,70,72,92
Continuous aperture required, 15 W module	1.41 to 17.4 m2, a 141x to 1,740x shortfall	modelled	59,70,72,92
Sustainable duty fraction with a buffer, 7 W module	0.12% to 1.52%; one second of operation per 66 s to per 13.6 min	modelled	59,70,72,92
Sustainable duty fraction with a buffer, 15 W module	0.057% to 0.71%	modelled	59,70,72,92
Winter floor at an assumed one third of annual mean	2.87 to 35.4 mW; 13 to 160 tinyML inferences per second after derate; 0.041% to 0.51% duty at 7 W	modelled	92
Only seasonal measurement located, thermal rather than solar	Road thermoelectric 162.33 mW summer against 34.63 mW winter, a ratio of 4.7	verified	76
Substrate multipliers available to move the aperture requirement	Reversible logic 1.41x to 1.77x recovery; neuromorphic spiking 2x to 145x, clustering at 2x to 20x on dense work	reported	84,85

Table 5. A general bound on ambient-powered computation at a 100 cm2 pavement-embedded aperture. Every derived row is graded modelled and carries index 92 alongside the sources of its inputs. Seven inputs are this author's assumptions and are marked as such: the aperture, the 90% system derate, the 30% rectifier efficiency, the one-third winter ratio, the linear scaling of every non-photovoltaic source with area, the lambda-squared effective aperture used for the radio-frequency term, and the treatment of an 8-hour thermoelectric measurement as recurring daily. The result scales linearly in aperture.

9. Discussion: where the evidence supports distribution, and the Institute's research position

Three findings support moving computation toward the machine. The first is energy and carbon per unit of work. The one peer-reviewed cloud-versus-edge life-cycle comparison located reports 93% less energy and 84% less carbon per query on a device NPU than on a data centre accelerator for a 7B model [61,62], the conclusion

survives substituting national water intensity factors for the study's own [1,61,92], and it survives a utilisation stress test down to a 1.96% duty cycle [62,92]. The comparison is not equal-accuracy, does not batch on the cloud side, and excludes network transport entirely, so it is a directional result rather than a settled margin. The second is the radio. On-device inference removes the 7 ms radio access term that section 4 shows is invariant to every other siting choice [23], and with it removes the outage question that no operator publishes a distribution for. That is the strongest form of the latency argument, and it argues for on-device rather than for any intermediate tier. The third is regulatory: in at least one jurisdiction the siting decision for a facility above 300 kW is already partly a district-heating-connection decision by statute [49,92], which is a driver that operates independently of any efficiency claim.

Three findings do not support it. The telco edge tier buys a real and measured 60.8% reduction in round trip [23,92] and lands on the wrong side of the requirement in both directions: the 17.27 ms that remains is still above the control period of on-board loops, and the published policy cadence of 0.5 to 0.8 seconds [26] means the saving is not needed for the workload that would use it. Its deployment record is consistent with that: three archived metros, twenty zones at one carrier, one consortium platform sold after its site was taken down, and one liquidation [29,31,32,33]. The latency argument as usually stated is the second: quoting a robot's control rate as its policy inference requirement overstates it by more than an order of magnitude for chunked-action architectures [26,27], and on that evidence a regional round trip is adequate for policy inference, which converts the siting question from a feasibility question into an economic one. The third is utilisation. Amortised distributed hardware carries a capital penalty of 2.5x against hyperscale at LBNL's modelled operational times, four to seven times larger than the price premium charged for metro edge siting [2,36,92], unless accelerator utilisation is siting-invariant, which LBNL assumes and nobody has measured.

Six measurements would change the answer, and their directions are known in advance. Measured accelerator utilisation by facility class decides whether the capital penalty above is real or an artefact of a fifteen-year-old assumption. A published joules-per-policy-inference figure on a named embodied task decides whether the device energy advantage measured on handset workloads transfers. Measured tail latency and outage rate on a production fleet decides whether the 327 ms of headroom in a chunked policy is comfortable or illusory. A marginal emission factor shaped to a flat continuous load decides whether the carbon comparisons in section 6 are answering the question they claim to answer. An independent

accelerator life-cycle assessment closes a factor-of-2.5 gap in the embodied term. And a measured network transport energy term decides whether offloading inference saves energy at system level or merely relocates it.

The Institute's research position follows from those gaps rather than from a preference among topologies. First, the siting question is not currently decidable from published evidence, and the reason is specific: the two denominators that would decide it, measured accelerator utilisation by facility class and joules per delivered inference on a named embodied task, are not published by any party, and both are measurable today by parties who already have the instrumentation. Second, on the evidence that does exist, the case for moving computation toward the machine is strongest where it removes the radio from the loop entirely and weakest at the intermediate tiers, which buy milliseconds the published policies do not need while paying a utilisation penalty the published prices do not cover. Third, a siting comparison computed on average grid factors and attributional allocation is not an emissions result for a siting decision, because siting is a marginal question, and such a figure should be reported as what it is; the difference between the two factors nationally is 1.6x to 2.0x [22,92], which is larger than most of the efficiency differences under argument. Fourth, the ambient-power bound of section 8 is a real constraint and it is a constraint on a tier rather than on the idea: at hand-sized apertures, ambient harvesting supports sensing-class inference continuously and embedded-module-class inference only intermittently, and that distinction is available today on commodity silicon without any novel substrate. Fifth, the Institute publishes the arithmetic rather than the conclusion, because the conclusion moves when the unpublished denominators arrive, and a reader who substitutes a different aperture, duty cycle, device price or grid factor should be able to recompute every number in this review without asking the author for anything.

10. Limitations of this review

This is not a systematic review. Sources were assembled by targeted retrieval against a set of questions rather than by a registered protocol with defined search strings and inclusion criteria, no reproducible search log is published, and one author assigned every grade with no inter-rater check. A different author working the same questions would assemble an overlapping but not identical evidence base.

Coverage is United States centred. European material appears where it is primary and establishes a regulatory driver, and Asian and other deployments are not

covered at all, which matters because siting economics depend on tariffs, permitting and grid structure that vary by jurisdiction.

Several primary documents returned HTTP 403 to retrieval and are cited through relays or search-index extracts: an IEA news figure [4], a moratorium tracker [15], a paywalled operator survey [35], two trade-press items on withdrawals [32,33], the coverage of one asset sale [46], a regulatory fact sheet corroborated through four independent law-firm summaries [45], the hyperscale construction cost [42], and the principal peer-reviewed heat-reuse synthesis, from which no figure is asserted [55]. Each is flagged at its reference entry. No conclusion in sections 9 or 12 rests on any of them.

The evidence base is asymmetric in a direction that favours the incumbent topology. Hyperscale operators publish tariffs, filings, life-cycle assessments and regulatory dockets, while distributed operators publish site counts and readiness claims. Absence of a distributed figure is therefore weak evidence about distributed performance, and several places where this review reports that something is not established should be read as non-disclosure rather than as a negative finding.

Comparisons across studies use inconsistent boundaries and this review does not harmonise them. Network transport energy is outside the boundary of every source located. Embodied water is unallocated everywhere, because the one study that addresses it states that reliable supply-chain water life-cycle data do not exist [62]. Scope and vintage differ across the utilisation figures that carry most of the weight.

Section 8 rests on seven assumptions that are the author's rather than any source's: the aperture, the 90% system derate, the 30% rectifier efficiency, the linear area scaling of every non-photovoltaic source, which is generous to the small sources, the lambda-squared effective aperture for the radio-frequency term, the treatment of an 8-hour thermoelectric measurement as recurring daily, and the one-third winter ratio. Each is labelled at the point of use and each is replaceable. The result scales linearly in aperture and in the derate, which is why those two are stated as multipliers rather than buried. The winter ratio is the weakest of the four, because no seasonal series for a road-surface photovoltaic installation was retrieved at all. Section 8 also bounds only the inference term of an ambient-powered module and says nothing about its radio, which is likely the dominant term.

No primary measurement was performed for this review. Every measurement quoted is somebody else's, and the contribution here is consolidation, grading and

arithmetic. The arithmetic was executed and checked programmatically [92], which removes calculation error and does not remove input error.

Finally, grades are source criticism and not quality assessment. A verified figure can be wrong and a self-published figure can be exactly right. What a grade records is who was in a position to be wrong in a particular direction and whether anyone independent has checked. Prices, queue figures and regulatory positions are dated to 5 August 2026 and several of them were in active proceedings at that date.

11. Open problems, each with the measurement that would settle it

Measured accelerator utilisation by facility class. Every amortisation argument in this subject, on both sides, rests on this one unpublished number, and the only production measurement located reports roughly 20% mean training FLOPs utilisation on one fleet over two weeks while noting that 80% of that fleet's processor-hours carried no measurement at all [18]. Settled by an audited series of accelerator-hours delivered against accelerator-hours provisioned, by facility class, on a definition that distinguishes operational time from FLOP utilisation rather than conflating them.

Joules per policy inference on a named embodied task, on deployed hardware. Vendors publish peak throughput and a power envelope [58,59] and nobody publishes the energy of one decision. Settled by an energy harness at the module terminals, in the style of the tinyML harness one tier down [80], reporting joules per inference alongside a named task and an accuracy figure.

Network transport energy per delivered inference. This is the term a siting argument most needs and it lies outside the measurement boundary of every source located, including the most detailed production disclosure [17] and the most relevant life-cycle comparison [62]. Settled by an end-to-end measurement from sensor to compute and back, decomposed into radio access, transport and core, at 2026 network intensities.

Tail latency and outage behaviour for a machine on a production network. Every latency figure in section 4 is a mean or a single percentile from synthetic probes at fixed vantage points [23,24,25]. Settled by a published distribution from a deployed fleet including the 99.9th and 99.99th percentiles, the jitter, and the frequency and

duration of link loss, which are the statistics that determine whether an offloaded deadline holds.

A marginal emission factor shaped to a continuous, flat, always-on load, at balancing-authority and hourly resolution. The national marginal and average factors differ by 1.6x to 2.0x [22,92] and neither published series is shaped to this load. Settled by a regulator or national laboratory publishing the series; until then every siting comparison substitutes a factor built for a different question, and should say so.

An independent life-cycle assessment of an AI accelerator. The only accelerator-level figure in circulation is published by the seller [67] and a peer-reviewed study independently assumes a value 2.5 times lower [62]. Settled by a third-party cradle-to-gate study with a published bill of materials and stated confidence by component.

The counterfactual on device manufacture. Whether a handset, laptop or machine would have been built anyway is the single assumption that flips the distributed-versus-centralised carbon result [63], and every study located is attributional and then chooses an amortisation convention [69]. Settled by a consequential life-cycle assessment that establishes the counterfactual rather than assuming it.

Capital cost per kW for a metro edge or micro facility on the same accounting boundary as hyperscale construction, covering land, shell, mechanical and electrical, grid connection and permitting. The published edge premium is an administered multiplier that prices the difference without decomposing it [36,92]. Settled by a disclosure from any party other than a vendor selling the module.

Seasonal harvest from a complete, sealed, road-legal aperture, and a measured yield class for a pole-mounted plate. Every yield in section 8 comes from an installation that was subsequently removed [70,71,72], and the pole-mounted case has no measured basis at all. Settled by a year of metered generation from a deployed unit including soiling, snow cover, tyre abrasion, shading by traffic and freeze-thaw, reported monthly.

Energy per delivered bit for a low-power radio at the duty cycles of section 8, under real link budgets and real interference. Section 8 bounds the inference term and is silent on the term most likely to dominate for any module that must communicate. Settled by a measured link budget from a deployed unit rather than a component datasheet.

Achieved energy reuse in aggregate. Reporting obligations exist in the European register [50] and national quotas are now binding in one member state [49], and no regulator has published an outturn distribution. Settled by publication of the register's per-facility reuse factors, which would replace design capacities and contracted volumes [51,52,53] with delivered energy.

Whether co-located and behind-the-meter load shifts cost onto other ratepayers. FERC left this to a paper hearing precisely because it is unresolved [45], and one legislature-commissioned study found a forward risk alongside a present finding that such load currently pays its full cost of service [12]. Settled by the record in that proceeding, and by cost-of-service studies published on a common boundary across more than one RTO.

12. Conclusions

The centralised baseline against which every distributed proposal is compared is a model rather than a measurement. The 176 TWh figure for 2023 is reconstructed from equipment shipments because facility data are unavailable, the 74 to 132 GW capacity figure is that model's energy range divided by an assumed 50% utilisation, and a twenty-point change in that unobserved assumption moves the capacity figure by roughly 28% [1,2,92]. Claims about distributed siting inherit that uncertainty before adding their own.

The latency case for distribution is real, measurable, and weaker than usually stated. Siting can attack 59.4% of a best-case round trip, and moving from a cloud region to a node inside the operator's network removes 60.8% of it [23,92]. The radio access term of about 7 ms does not move at all [23]. Published policies run inference every 0.5 to 0.8 seconds behind chunked action sequences, so a wide-area round trip of 100 ms, roughly eight times the measured one, still leaves 327 ms of headroom [26,92]. The workloads that genuinely cannot be offloaded are bounded by certifiability rather than by milliseconds. That constraint sits in the functional-safety standards rather than in the approach-speed formula cited here, and no reduction in round-trip time relaxes it [28].

The distributed deployment record is thin and its withdrawals have commercial rather than physical causes. The best-funded tower-edge plan would have totalled 19.2 MW had it been built, which is 1.6% of a single AI campus under construction today [33,46,92]. Telco edge footprints are counted in tens of sites against hundreds of

thousands of cell sites [29,91,92], and no party in the category publishes capacity, utilisation or customers.

The environmental comparison is decided by utilisation and by grid carbon intensity, and not by hardware. The embodied share of a rack server ranges from 7.7% to 46.4% across grids that exist today, with embodied equalling operational at about 43 gCO2e/kWh, so the embodied share rises as grids decarbonise [66,22,2,92]. On the single published case where every input is available, embodied carbon is 14.9% of the cloud total and 62.3% of the device total [62,92], and the device would have to be active less than 2% of the time before that penalty erased a 93% operational advantage [62,92]. On-site water is 7.6% of the data centre water footprint, so distribution can remove at most that fraction [1,92].

The cost comparison is dominated by prices nobody decomposes and denominators nobody publishes. The metro edge premium is an administered multiplier of exactly 1.2500 across unrelated instance families [36,92]. The savings claims for distributed compute are denominated in a list price the incumbent itself discounts by up to 62.40%, so a 60%-off claim lands above the incumbent's own committed rate [36,88,92]. The utilisation-driven capital penalty for a small facility is 2.5x, four to seven times the premium charged for edge siting, and it vanishes entirely if accelerator utilisation is siting-invariant, which is assumed and unmeasured [2,92].

At the most distributed extreme the physics is unambiguous and tier-dependent. At a 100 cm2 pavement-embedded aperture, measured trafficked-surface photovoltaic yields give 8.61 to 106.2 mW, photovoltaics supply 96.6% to 99.7% of the total excluding a thermoelectric, which a sealed pavement module cannot use, and 50.5% to 92.6% including one, and those sources compete for one surface rather than adding [70,72,74,76,77,78,92]. That harvest sustains 39 to 478 tinyML-class inferences per second after a harsh derate, and 13 to 160 at an assumed winter floor [80,92]. It cannot continuously power a 7 to 15 W embedded inference module, falling short by 66x to 1,740x in aperture, though it can operate one in bursts at a duty fraction between 0.057% and 1.52% against a buffer [59,92]. The sensing tier closes on commodity silicon today; the embedded-module tier does not close continuously at any measured yield.

The question the review set out to answer is not settleable from published evidence, and the reason is nameable. Two quantities decide it, measured accelerator utilisation by facility class and joules per delivered inference on a named embodied task, and both are measurable today by parties who already own the instrumentation, and neither is published. Until they are, every siting argument in

this literature, including the ones this review finds most persuasive, is an argument about assumptions. The productive response is to publish the denominators rather than to sharpen the rhetoric.

References

1. [1] Shehabi, A., et al., "2024 United States Data Center Energy Usage Report", Lawrence Berkeley National Laboratory, LBNL-2001637, December 2024. Historical accounting for 2014-2023: 176 TWh and 4.4% of US electricity in 2023 against about 76 TWh (1.9%) in 2018, 66 billion litres direct and nearly 800 billion litres indirect water, 61 billion kg CO2e, and national intensity factors of 4.52 L/kWh indirect water and 0.34 kg CO2e/kWh. Congressionally mandated update to LBNL's 2016 report. The totals are bottom-up from equipment shipment counts, analyst market data and assumed operating hours; the report states that the lack of data availability significantly limits the analysis and that direct facility energy data are unavailable.
   VERIFIED
2. [2] Shehabi, A., et al., LBNL-2001637, December 2024, forward scenarios and model inputs: 325-580 TWh and 6.7-12.0% for 2028, 74-132 GW at an assumed 50% average capacity utilisation, national PUE 1.4 (2023) falling to 1.15-1.35 (2028), site WUE just over 0.36 rising to 0.45-0.48 L/kWh, and server operational time by class (internal and small 11% to 20%, colocation 21% to 35%, hyperscale 45% to 50%, AI training constant 80%, AI inference constant 40%). These are scenario spans and modelling assumptions, several carried forward from the 2016 report, and not measurements of any fleet. LBNL states that its simulated PUE and WUE assume systems are commissioned and operate as designed, that this is rarely the case, and that actual values will likely not be as good as estimated.
   MODELLED
3. [3] International Energy Agency, "Energy and AI", April 2025. Global data centre electricity around 415 TWh in 2024, about 1.5% of global consumption, with the US at 45%, China 25% and Europe 15%. The 2030 figure of roughly 945 TWh is a projection and is treated as modelled wherever used. iea.org returned HTTP 403 to this author; the 2024 figures are corroborated through the Environmental Law Institute's January 2026 fact sheet, which cites the IEA page directly.
   VERIFIED
4. [4] IEA news release stating that electricity use by AI-focused data centres surged 50% in 2025, relayed by trade press. Both the IEA page and the relay returned HTTP 403 to this author and the figure comes from a search-result summary that was not checked against the primary. It is recorded for completeness and no claim in this review rests on it.
   REPORTED

5. [5] Electric Power Research Institute, "Powering Intelligence: Updated Scenarios of U.S. Data Center Electricity Use and Power Strategies", 2026. 9% to 17% of US electricity by 2030 against 4-5% today, 177-192 TWh in 2024 growing to 380-790 TWh by 2030, and 56-132 GW of nominal IT capacity by 2030 against 35-44 GW in 2024. EPRI is funded by the electric utility industry, which has an interest in load-growth forecasts, and states that these projections are about 60% higher than its own analysis of eighteen months earlier, which is the honest measure of how firm any of them are. EPRI itself warns that announced nominal capacity should be treated as a pipeline indicator rather than a near-term peak forecast.
   MODELLED
6. [6] US Securities and Exchange Commission, EDGAR XBRL company facts, us-gaap:PaymentsToAcquirePropertyPlantAndEquipment (Alphabet, Microsoft, Meta) and us-gaap:PaymentsToAcquireProductiveAssets (Amazon), 10-K and 10-Q filings through 31 July 2026. Alphabet $91.45bn FY2025 and $80.60bn H1 2026; Amazon $131.82bn FY2025 and $98.41bn H1 2026; Microsoft $115.95bn for the fiscal year ended June 2026, of which $66.68bn falls in January to June 2026; Meta $69.69bn FY2025 and $19.00bn Q1 2026.
   VERIFIED
7. [7] Aggregate 2026 hyperscaler capital expenditure guidance of roughly $690bn to $725bn across Microsoft, Amazon, Alphabet and Meta, with Microsoft attributing about $25bn of its approximately $190bn to component price inflation, relayed by Futurum Group, Yahoo Finance and CNBC. Forward-looking targets stated by the companies doing the spending, graded self-published under the series tie-break regardless of which outlet carried them. Guidance has been revised upward repeatedly within single fiscal years, so it is a statement of intent.
   SELF-PUBLISHED
8. [8] Rand, J., et al., "Queued Up: 2026 Edition", Lawrence Berkeley National Laboratory, June 2026. About 2,061 GW across 8,244 projects actively seeking transmission interconnection at end-2025 against an installed US fleet of 1,374 GW; 19% of projects and 13% of capacity requesting interconnection between 2000 and 2020 in service by end-2025; a median 61 months in queue for projects built in 2025 against 36 months in 2015 and 22 months in 2008; over 750 GW withdrawn against about 600 GW of new requests in 2025, with active gas capacity up 86% year on year.
   VERIFIED
9. [9] Electric Reliability Council of Texas, "ERCOT Update", presentations by CEO Pablo Vegas to the Texas Senate Committee on Business and Commerce, 1 April 2026 and 29 July 2026. Approximately 410 GW of large loads in the interconnection queue as of 26 March 2026, about 87% data centres; approximately 474.7 GW as of June 2026 of which 420.8 GW (90.2%) data centres; all-time hourly peak demand 91,089 MW on 22 July 2026.
   VERIFIED

10. [10] Patel, S. C., "Abbott Orders Full Audit of Texas Data Center Interconnection Queue, Threatens to Deny Grid Access", POWER magazine, 4 August 2026. Trade press reporting of a governor's letter of 3 August 2026 directing the PUCT and ERCOT to audit every data centre in the queue and to deny grid access to projects that fail to disclose ownership, financial, water and community-impact information, citing non-compliance with the PUC's statutory water and power usage survey. The letter itself was not retrieved.
REPORTED

11. [11] PJM Interconnection, "PJM Auction Procures 134,479 MW of Generation Resources", news release, 17 December 2025. The 2027/2028 Base Residual Auction cleared at the FERC-approved cap of $333.44/MW-day and fell 6,623 MW short of the reliability requirement, the first time the entire RTO including FRR areas has done so; nearly 5,100 MW of a 5,250 MW increase in the peak load forecast is attributable to data centre demand; total cleared cost $16.4bn.
VERIFIED

12. [12] Joint Legislative Audit and Review Commission of the Virginia General Assembly, "Data Centers in Virginia", Report 598, December 2024. Unconstrained demand in Virginia would double within ten years with data centres the main driver, and a typical Dominion Energy residential customer could see generation and transmission costs rise by an estimated $14 to $37 per month in constant dollars by 2040. The dollar range is consultant modelling within a legislature-commissioned report and not an observed bill increase, and the same report found that data centres currently pay their full cost of service and that current rates allocate costs appropriately, so the cost shift is a forward risk rather than a present finding. It also found most Virginia data centres using about as much water as an average large office building or less, and current use sustainable though poorly managed across competing local uses, which is included because it cuts against the direction of most water commentary.
VERIFIED

13. [13] Bass, D., et al., "AI is Draining Water From Areas That Need It Most", Bloomberg Technology, May 2025, reporting that roughly two-thirds of data centres built since 2022 are sited in water-stressed regions, retrieved through the Environmental Law Institute, "Data Centers and Water Fact Sheet", January 2026. A journalistic analysis whose method and facility list are behind a paywall and were not inspected by this author; water-stressed is a threshold judgement dependent on the stress index chosen; the ELI fact sheet is itself a secondary compilation.
REPORTED

14. [14] Google 2026 Environmental Report, published 30 June 2026, reporting 10.9 billion gallons of water consumed in 2025, up 34% year on year and more than double the 2021 level, with 7.7 billion gallons (78%) replenished through 165 projects across 97 watersheds. Weak sourcing: the primary report was not retrieved by this author and the figures come from secondary summaries. Google is one of the few operators disclosing anything, which is why the figure appears at all.
SELF-PUBLISHED

15. [15] Good Jobs First, "Data Center Moratorium Bills Are Spreading in 2026", and private trackers including datacenterbans.com, dcmap.us and Green Data Center Guide, July 2026, reporting more than 500 moratorium instruments across 42 states including emergency ordinances in Seattle (June 2026) and Cleveland (July 2026). The count could not be verified: the Good Jobs First page returned HTTP 403 and the trackers do not share a definition of a moratorium instrument, mixing binding ordinances, temporary permit pauses and filed bills that never passed.
REPORTED

16. [16] Data Center Map, "USA Data Centers", accessed 6 January 2026, listing approximately 3,778 US facilities, cited through the ELI water fact sheet of January 2026. A commercial directory with a self-selected listing and not a census. No federal agency maintains a facility register with location, capacity, metered load or water withdrawal.
REPORTED

17. [17] Elsworth, C., et al. (Google), "Measuring the environmental impact of delivering AI at Google Scale", arXiv:2508.15734, 21 August 2025. Median Gemini Apps text prompt in May 2025: 0.24 Wh, 0.26 mL water and 0.03 gCO2e, split as 0.14 Wh active accelerator (58%), 0.06 Wh host CPU and DRAM (25%), 0.02 Wh provisioned idle machines (10%) and 0.02 Wh data centre overhead (8%); fleet-wide average PUE 1.09. Google measuring Google, with no independent replication and no release of the underlying telemetry.
SELF-PUBLISHED

18. [18] Pedersen, C., Ahn, S., Migdal, J., Neale, R., Konyuchenko, D. (NVIDIA), "Instant GPU Efficiency Visibility at Fleet Scale", arXiv:2605.20799v1, 21 May 2026. Measured training model FLOPs utilisation averaging approximately 20% over a two-week window on a production fleet, against a 35-50% range the authors describe as expected, with only about 20% of fleet workloads onboarded to MFU reporting at all, leaving 80% of GPU-hours with no utilisation measurement.
REPORTED

19. [19] Lei, J., Fernandez, D., Kypriotis, A., Skarlatos, D., Strubell, E., Sherry, J., Vosler, P., "The Energy Cost of Execution-Idle in GPU Clusters", arXiv:2604.04745, 6 April 2026. Across 11,791 long-running jobs on an academic GPU cluster, 19.7% of in-execution time and 10.7% of in-execution energy is attributed to execution-idle, meaning GPUs allocated to a running job and drawing power while doing no compute. Preprint, not peer-reviewed.
REPORTED

20. [20] Uptime Institute, "Global Data Center Survey 2024", reporting an industry average PUE of 1.56. An unweighted average of self-reported facilities, which over-represents older and smaller sites. This is not in conflict with LBNL's stock-weighted modelled 1.4 and must not be presented as such: the two weight different populations, and which is the right comparator depends on which facility the compute would otherwise have run in.
REPORTED

21. [21] Guidi, G., Dominici, F., Squartini, S., Sprinkle, J., Gilmour, M., Butler, C., Bell, M., Delaney, S., Bargagli-Stoffi, F., "Assessing the Carbon Emissions and Energy Consumption of U.S. Hyperscale Data Centers", arXiv:2606.05420, 3 June 2026. 403 US hyperscale facilities, May 2024 to April 2025: 68-99 TWh, 37-54 Mt CO2, and an electricity-weighted average carbon intensity of approximately 545 gCO2/kWh against a contemporaneous national grid average of 370, with roughly 54% of attributed generation fossil. Preprint, not peer-reviewed at retrieval.
MODELLED

22. [22] US Environmental Protection Agency, "Greenhouse Gas Equivalencies Calculator: Calculations and References" and its revision history, 2024-2025. AVERT national weighted-average marginal CO2 rate 6.03 x 10^-4 metric tons CO2/kWh (2022 data), against an earlier revision using 7.44 x 10^-4; eGRID national average output rate 823.1 lb CO2/MWh (2022 data), which converts to 373 gCO2/kWh. The AVERT rate is intervention-shaped, published separately for wind, solar, storage and uniform efficiency, and the figure quoted is the wind-displacement profile; the eGRID figure is CO2 only and not CO2e. The two are not a clean like-for-like pair and the gap between them should be read as an order-of-magnitude signal about accounting choice rather than a precise ratio. No regulator publishes a marginal factor shaped to a continuous, flat, always-on load at balancing-authority resolution.
VERIFIED

23. [23] Fezeu, R. A. K., Ramadan, E., Ye, W., Minneci, B., Xie, J., Narayanan, A., Hassan, A., Qian, F., Zhang, Z.-L., Chandrashekar, J., Lee, M., "An In-Depth Measurement Analysis of 5G mmWave PHY Latency and Its Impact on End-to-End Delay", Passive and Active Measurement (PAM) 2023, Springer LNCS, Table 2. End-to-end round-trip time on commercial Verizon mmWave 5G: 17.27 ms (+/- 1.31) to an AWS Wavelength node, 38.15 ms (+/- 1.83) to an AWS Local Zone and 44.08 ms (+/- 3.04) to an AWS Region, with the radio access network component at 7.02-7.60 ms regardless of server siting. One carrier, mmWave only, line-of-sight and high-CQI conditions, and no compute time included. The absolute values are specific to 2022-23 deployments; the invariance of the radio term across siting choices is the durable result.
VERIFIED

24. [24] Charyyev, B., Arslan, E., Gunes, M. H., "Latency Comparison of Cloud Datacenters and Edge Servers", IEEE GLOBECOM 2020, NSF PAR 10184999. From 8,456 RIPE Atlas vantage points to 6,341 Akamai edge servers and 69 cloud locations across five providers: 55% of end-users reached an edge server within 10 ms and 82% within 20 ms, against 3-21% and 22-52% for individual cloud providers, rising to 27% and 62% for all cloud providers treated as one.
VERIFIED

25. [25] Sharma, et al., "Evaluating 5G-connected IoT for Power Line Temperature Prediction: Real-World Latency and Cost Trade-offs Between MEC and Cloud", arXiv:2607.03993v1, July 2026. Measured P99 latency on Telenor's production 5G network at Fornebu, Norway: 44.62 ms to operator MEC, 47.69 ms to the nearest

cloud region, 78.49 ms to a region in the neighbouring country, and about 660 ms at a remote 4G-only site.

REPORTED

26. [26] Black, K., et al., "pi0: A Vision-Language-Action Flow Model for General Robot Control", Physical Intelligence, Table I and Appendix D. On an RTX 4090: image encoders 14 ms, observation forward pass 32 ms, ten flow-matching steps 27 ms, off-board network latency 13 ms, totals 73 ms on-board and 86 ms off-board. Action chunk horizon 50; for 50 Hz robots inference runs every 0.5 s after executing 25 actions, and for 20 Hz robots every 0.8 s after 16.

SELF-PUBLISHED

27. [27] Wei, et al., "Psi-0: An Open Foundation Model Towards Universal Humanoid Loco-Manipulation", arXiv:2603.12263v1, 2026. A single forward pass takes about 160 ms while the control loop runs at 30 Hz (a 33 ms period) and the low-level locomotion thread at 60 Hz, so inference is already about five control periods long before any network is involved. Preprint; the 160 ms is on the authors' own hardware, which they characterise as server-side, so the figure is not portable to on-robot silicon.

REPORTED

28. [28] ISO 13855 (positioning of safeguards with respect to approach speeds), approach speed constant K = 2,000 mm/s for hand and arm intrusion, and ISO/TS 15066 (collaborative robots), protective separation distance Sp = Sh + Sr + Ss + C + Zd + Zr with Sr the robot system reaction time contribution. Read through machine-safety vendor summaries and a peer-reviewed comparative review of ISO 10218 (arXiv:2602.17822) rather than the paywalled primary text.

VERIFIED

29. [29] Amazon Web Services, "AWS Wavelength Zone locations", retrieved 5 August 2026. 33 zone entries across seven carrier partners, of which 20 are Verizon zones in the United States, with the page stating availability in 31 cities. AWS's own page. The 33-versus-31 gap is partly explained by Manchester appearing under both BT and Vodafone; the remainder was not established and is not speculated on.

SELF-PUBLISHED

30. [30] Amazon Web Services, "AWS Local Zones locations", retrieved 5 August 2026, listing availability in more than 30 metropolitan areas with a further seven announced. AWS's own page; this author's extraction of the individual city lists was internally inconsistent, so only the headline count is used. No capacity, utilisation or revenue is published.

SELF-PUBLISHED

31. [31] Microsoft Learn, "Key concepts for Azure public MEC" (archived), page metadata and parent-region table, retrieved 5 August 2026. Exactly three live sites, all with AT&T: Atlanta, Dallas and Detroit. The documentation set carries a /previous-versions/ canonical URL, an is_archived flag and a NOINDEX,NOFOLLOW robots directive, with a last content update of August 2024.

SELF-PUBLISHED

32. [32] STL Partners, "Google's acquisition of MobiledgeX", 12 May 2022; Light Reading, "MobiledgeX is dead. Long live Google Cloud"; DCD, "Google acquires Edge computing company MobiledgeX". The telco edge platform founded by Deutsche Telekom in 2018 with an operator consortium was sold to Google in April 2022 after its website and social presence had been taken down; the code was open-sourced and the chief executive left within a week. Neither Deutsche Telekom nor Google disclosed investment, headcount, revenue or operator deployments.
    REPORTED
33. [33] DCD, "Report: EdgeMicro has gone into liquidation" (reporting Light Reading), and Data Center Frontier, "EdgeMicro Readies Its Micro Data Center Platform for Edge Growth", 7 March 2019. EdgeMicro planned tower-sited micro data centres across US tier-2 cities, raised about $11 million, deployed roughly three sites (Austin, Tampa, Raleigh) against a stated plan for 400 cell-tower locations, and went into liquidation.
    REPORTED
34. [34] Vapor IO, "Edge-to-Edge connectivity in 36 US markets", retrieved 5 August 2026. The company's own market map shows 8 markets live, 26 described as customer ready within 90 days of an order, and 3 coming soon, against a page headline of 36. No dates, no capacity in MW and no customer count are published; customer ready in 90 days is a sales statement rather than a deployment; and the 36-versus-37 discrepancy between the headline and the list was not resolved. No independent audit of which sites carry live load was located.
    SELF-PUBLISHED
35. [35] Omdia, "Telco Edge Computing Survey 2025: Opportunities, Challenges, Demand Drivers, and Infrastructures", reporting that only 15% of telcos ranked the network far edge as the top location for where most AI inferencing will take place. The Omdia page is paywalled and returned HTTP 403; the figure is quoted from a search-index extract and the sample size, question wording and respondent mix were not established.
    REPORTED
36. [36] Amazon Web Services, Price List Bulk API, offer AmazonEC2, version 20260805185658, published 5 August 2026, regions us-east-1, us-east-1-nyc-1, us-east-1-wl1-nyc1, us-west-2 and us-west-2-lax-1, Linux, shared tenancy, on-demand, no pre-installed software. Includes p5.48xlarge on-demand at $55.04/hour and its three-year Reserved terms under SKU 3D4V8UAYEMB38GU2 ($543,866 All Upfront, $289,290 plus $11.008/hour Partial Upfront, $23.77728/hour No Upfront), g6.xlarge at $0.8048/hour, and the per-instance rates used for the metro-edge and Wavelength premium comparisons. The operator's own published tariff, fully re-extractable by any reader, and a list price rather than a transaction price. Negotiated enterprise discounts are not published, so 62.40% is an upper bound on the published discount and a lower bound on the achievable one. Only four instance types are offered in the Verizon New York Wavelength Zone at all.
    REPORTED

37. [37] Amazon Web Services, Price List Bulk API, offer AWSDataTransfer, current version, retrieved 5 August 2026. Inbound from External to US East (N. Virginia) $0.0000/GB; outbound to internet $0.090/GB for the first 10 TB per month, $0.085 for the next 40 TB, $0.070 for the next 100 TB and $0.050 above 150 TB; Local Zone to parent region $0.0000/GB in both directions; internet egress from the Verizon New York Wavelength Zone $0.108/GB against $0.090 in the parent region. The operator's own tariff.
REPORTED

38. [38] TeleGeography, "IP Transit Pricing in 2025: More Competition, More Price Erosion", 2025. $0.05 per Mbps per month for 100 GigE and $0.07 for 10 GigE in the most competitive markets in Q2 2025, with 100 GigE prices falling 12% compounded annually from Q2 2022. These are the lowest prices in the most competitive hubs and not medians, and they are port prices rather than delivered per-GB prices; real buyers pay 95th-percentile billing on partly filled ports, so the effective per-GB cost is several times the theoretical floor and the ratios computed against it in section 7 are correspondingly overstated in the unfavourable direction.
REPORTED

39. [39] Hologram, IoT pricing page, 2026: $0.03 per MB, which is $30 per GB, plus $1 per SIM per month and $3 per SIM one time. A low-volume self-service rate for small telemetry devices and not the rate a robot fleet would pay. Hologram states that volume discounts exist and does not publish them; no carrier publishes a per-GB rate for a high-volume machine connection, so the real last-mile cost for physical AI is unpriced in public.
REPORTED

40. [40] Cloudflare Workers AI pricing documentation, 2026, listing llama-3.3-70b-instruct-fp8-fast at $0.293 per million input tokens and $2.253 per million output tokens; Together AI pricing page, 2026, listing Llama 3.3 70B at $1.04 per million tokens for both input and output. Two vendors, different hardware, different serving stacks and different availability terms.
REPORTED

41. [41] US Energy Information Administration, Electric Power Monthly, Table 5.6.A, May 2026 data: 13.83 cents per kWh across all sectors, 8.71 cents industrial, 13.54 cents commercial and 18.44 cents residential, up 5.3% across all sectors year over year. National averages across all customers. A hyperscale campus on a negotiated industrial tariff and a small distributed cabinet on a commercial account do not pay the same rate, the spread between them is the number a siting argument actually needs, and EIA does not publish it at that granularity.
VERIFIED

42. [42] dgtlinfra, "How Much Does it Cost to Build a Data Center", quoting roughly $10.7 million to $12.0 million per MW for hyperscale construction, with $11.5M/MW attributed to the Equinix development pipeline and $12.0M/MW and about $1,050 per square foot to the Digital Realty pipeline, and AI facilities at roughly double.

Weakest-provenance item in this review and flagged as such in sections 2 and 7: the source article returned HTTP 403, this author could not read it, and could not verify how the two company figures were derived from those companies' filings. Treat as an order of magnitude only. No comparable published per-MW figure for metro edge sites on the same accounting boundary was located from any party other than a vendor selling the module.

REPORTED

43. [43] Federal Energy Regulatory Commission, order in docket ER24-2172, 1 November 2024, rejecting by 2-1 the amended interconnection service agreement that would have expanded Amazon's behind-the-meter data centre load co-located at Talen's Susquehanna nuclear plant from 300 MW to 480 MW, on the ground that PJM had not justified the non-standard provisions. Reported by Utility Dive, RTO Insider, POWER magazine and ANS Nuclear Newswire.

VERIFIED

44. [44] Talen Energy announcement of 11 June 2025 restructuring the Amazon arrangement as a front-of-the-meter retail power purchase agreement of up to 1,920 MW running to 2042, explicitly to remove the need for FERC approval, with the existing 300 MW co-located load folded in and transition expected in spring 2026, reported by Utility Dive. The approximately $18 billion lifetime revenue figure originates with Talen and is self-published. The ramp (840-1,200 MW by 2029, 1,680-1,920 MW by 2032) is a forward schedule and not delivered capacity.

REPORTED

45. [45] Federal Energy Regulatory Commission, order in docket EL25-49, 18 December 2025, finding PJM's tariff unjust and unreasonable for lacking clear terms for co-located load and directing compliance filings by 20 January and 16 February 2026 offering co-located customers a choice of network integration service on a gross-demand basis, a new firm contract demand service, or a non-firm contract demand service. The FERC fact sheet returned HTTP 403 on fetch; the content is taken from the FERC news title plus four independent law-firm summaries that agree on date, docket and the three options. Rates and terms were left to a paper hearing with briefs running to April 2026, precisely because whether co-located load shifts cost onto other ratepayers is not established. PJM only; other RTOs differ.

VERIFIED

46. [46] CNBC, "Crusoe Energy sells bitcoin mining unit to focus on huge opportunity in AI", 25 March 2025, and DCD, "Crusoe exits crypto operations to focus on AI". Crusoe sold its digital flare mitigation and bitcoin business to NYDIG, transferring 270 MW of generation, more than 425 modular data centres in the US and Argentina and about 135 employees, and redirected the company to a 1.2 GW AI campus at Abilene, Texas. Both pages returned HTTP 403 on direct fetch and the figures come from search-index extracts; re-verify before republication.

REPORTED

47. [47] Crusoe, "Minimizing the harm of methane emissions through flare reduction", company blog: 99.9% combustion efficiency against a 91.1% average for flares, a

68.6% reduction in CO2-equivalent emissions versus flaring, and about 800,000 tons per year of CO2e abated across about 125 modular data centres. The seller's figures for the product being sold.
SELF-PUBLISHED

48. [48] Plant, G., et al., "Inefficient and unlit natural gas flares both emit large quantities of methane", Science, 30 September 2022. Airborne sampling across the three US basins responsible for more than 80% of US flaring found that flares destroy only 91.1% of methane (95% CI 90.2-91.8) against the 98% assumed in inventories, roughly a fivefold increase in effective emissions and 4-10% of total US oil and gas methane emissions, with unlit flares and inefficient combustion contributing comparably.
VERIFIED

49. [49] Energieeffizienzgesetz (EnEfG) section 11, Federal Republic of Germany, primary legislation read directly at gesetze-im-internet.de. Facilities entering operation from 1 July 2026 must reach PUE at or below 1.2 and an energy reuse quota of at least 10%, rising to 15% from 1 July 2027 and 20% from 1 July 2028; electricity must be 50% renewable from January 2024 and 100% from January 2027; the regime covers facilities of at least 300 kW non-redundant rated electrical capacity.
VERIFIED

50. [50] Directive (EU) 2023/1791 (Energy Efficiency Directive), requiring data centres with installed IT power demand of at least 500 kW to monitor and report energy consumption, power utilisation, temperature set points, waste heat utilisation, water use and renewable energy share, while setting no binding minimum efficiency or performance standard.
VERIFIED

51. [51] Eurelectric, "Stockholm Exergi: Data Parks shows the potential for scaling the reuse of waste heat when it becomes a tradable product", 3 June 2026. The Open District Heating programme connects more than 30 data centres across 16 providers and supplies about 3.5% of Stockholm's heat supply against a 10% target, paying operators roughly SEK 2 million (about EUR 170,000) per MW per year. Figures originate with the party running the programme and are relayed by an industry association, so they are self-published under the tie-break.
SELF-PUBLISHED

52. [52] Ramboll, "Meta: surplus heat to district heating" project page, and the Harvard Business School environment blog on Fjernvarme Fyn, January 2024. Meta's Odense heat recovery is designed to deliver 215,000 MWh per year through about 45 MW of heat pumps to more than 12,000 homes, operational from late 2019; figures of 165,000 MWh and about 100,000 MWh per year also circulate in secondary coverage for what may be different scopes.
SELF-PUBLISHED

53. [53] CIBSE Journal, "Making a splash: recovering heat from mini data centres for leisure centres", read directly, and coverage in The Next Web and The Register of

Octopus Energy's GBP 200 million commitment to Deep Green, January 2024. Reported savings of about GBP 22,000 a year at Exmouth and a 62% cut in one pool's gas requirement at Trafford. The case study contains no IT load in kW, no kWh of heat delivered, no percentage of pool heat demand met and no recovery fraction; the savings figures originate with the operator.

SELF-PUBLISHED

54. [54] Tofani, "A case study on the integration of excess heat from Data Centres in the Stockholm district heating system", KTH Royal Institute of Technology, TRITA-ITM-EX 2022:469, 2022. Source for the figure that roughly 98% of the energy consumed by a data centre is dissipated as heat, which the thesis quotes from earlier literature rather than measuring, so the primary should be traced. It also records a network supply temperature of 68 degrees C with 35-55 degrees C return, which is why data centre heat needs upgrading by heat pump before it can be sold.

REPORTED

55. [55] Yuan, Liu, Sun, Lin, Fan, Zhao, Kosonen, "Data center waste heat for district heating networks: A review", Renewable and Sustainable Energy Reviews 219:115863, September 2025. Listed as a located but unread source: the publisher returned HTTP 403 and no quantitative content from this review is asserted anywhere in this report. It is recorded so that a reader extending this work knows where the principal peer-reviewed synthesis sits.

VERIFIED

56. [56] Kuzay, Demirel, Yilmaz, Koirala, Heer, "Maximizing waste heat recovery from a building-integrated edge data center", Scientific Reports, 5 November 2025. A validated simulation of a single-rack, air-cooled, building-integrated edge data centre (32 OCP servers, 12 kW maximum IT load, Empa NEST, Switzerland) reporting server air outlet temperatures of 35-45 degrees C, waste heat recovery improved by up to 17.1% at 3 kW IT load through workload allocation, and cooling load reduced by up to 53.2% through water flow rate optimisation.

MODELLED

57. [57] Google Research, "Exploring a space-based, scalable AI infrastructure system design" (Project Suncatcher), 4 November 2025. Claims that a solar panel in a dawn-dusk sun-synchronous orbit can be up to 8 times more productive than on Earth and that launch prices below $200/kg by the mid-2030s would make space data centre cost comparable to terrestrial, supported by a bench demonstration of an 800 Gbps inter-satellite optical link and TPU radiation tolerance to 2 krad(Si) against an expected 750 rad(Si) five-year dose, with two prototype satellites planned by early 2027. The company's own feasibility study for its own programme. Cost parity is conditional on a launch price that does not exist and a date in the next decade, so it is a forecast rather than a checkable resource claim; the radiation result is a single-part ground test; and thermal rejection in vacuum, the binding physical constraint, is not quantified.

SELF-PUBLISHED

58. [58] NVIDIA, "Introducing NVIDIA Jetson Thor, the Ultimate Platform for Physical AI", developer blog, 25 August 2025, and the Jetson Thor and Jetson Modules product pages, 2026. The T5000 module is specified at 2,070 FP4 TFLOPS sparse (1,035 dense) in a 40-130 W configurable envelope with 128 GB LPDDR5X at 273 GB/s, developer kit $3,499; the T4000 at 1,200 FP4 TFLOPS sparse in 40-70 W. Vendor peak-throughput specifications at a precision few deployed models use, and not measured figures on any named workload. No joules-per-inference figure for any deployed robot policy is published, and no module price appears on NVIDIA's own specification pages. The 40-130 W envelope belongs to this family and must not be transferred to the lower-power modules at reference 59.
SELF-PUBLISHED

59. [59] NVIDIA, Jetson Orin Nano series module specifications, product and module pages, 2026. The Jetson Orin Nano series is specified with configurable module power modes of 7 W and 15 W. This is the specification that carries the 7 W and 15 W envelope used as the embedded-inference reference load in section 8, and it is a vendor envelope rather than a measured figure on any named workload. It is a different and lower-power part than the Jetson Thor family at reference 58, whose envelope is 40-130 W; the two must not be interchanged.
SELF-PUBLISHED

60. [60] Gupta, U., Kim, Y. G., Lee, S., Tse, J., Lee, H.-H. S., Wei, G.-Y., Brooks, D., Wu, C.-J., "Chasing Carbon: The Elusive Environmental Footprint of Computing", HPCA 2021 / IEEE Micro 2022, arXiv:2011.02839. The share of an iPhone's life-cycle carbon attributable to hardware manufacturing rose from 49% (2009) to 86% (2019); at one hyperscaler in 2019, after conversion of its data centres to renewable energy purchases, capital and supply-chain activities accounted for 23 times more carbon than operational activities; and amortising the manufacturing footprint of a Google Pixel 3 requires running MobileNet inference continuously for three years, beyond the typical handset lifetime. A peer-reviewed synthesis over corporate sustainability reports, so the underlying inputs are self-published. The 23x figure is specific to a fleet whose operational emissions had been driven near zero by market-based procurement, which is an accounting result rather than a physical one.
VERIFIED

61. [61] Li, J., Islam, M. A., Ren, S., "A Case Study of Environmental Footprints for Generative AI Inference: Cloud versus Edge", ACM SIGMETRICS Performance Evaluation Review, 2025, measured per-inference energy. A Samsung S24 NPU against an NVIDIA A100: Llama-2-7B 262.26 J against 3,803.71 J (93% lower), Stable Diffusion 1.5 34.27 J against 714.86 J (95%), ControlNet 103.62 J against 1,292.34 J (92%), TrOCR 0.84 J against 3.39 J (75%). Cloud energy includes PUE 1.10 and device energy includes AC-DC charging loss at 0.8 efficiency.
VERIFIED

62. [62] Li, J., Islam, M. A., Ren, S., ACM SIGMETRICS Performance Evaluation Review, 2025, modelled life-cycle outputs and parameters, Tables 2 and 3 and section 4.2. Per 1,000 Llama-2-7B queries: 484.29 gCO2 on an A100, of which 412.07

operational and 72.22 embodied, against 75.45 gCO2 on a Samsung S24, of which 28.41 operational and 47.04 embodied, an 84% reduction, with water down 95%. Stated inputs: carbon intensity 390 gCO2/kWh (a US average and not a marginal factor), PUE 1.10, on-site water 1 L/kWh, grid water 3.142 L/kWh, embodied carbon 518 kgCO2 per cloud GPU and 39.2 kgCO2 per Samsung S24, both amortised over 5 years, cloud utilisation 1.0 and phone utilisation 0.19 (279 minutes per day of non-voice activity).

MODELLED

63. [63] Xue, et al. (University of Edinburgh, Johns Hopkins, Cisco Research), "Towards Decentralized and Sustainable Foundation Model Training with the Edge", HotCarbon 2025, arXiv:2507.01803. Reports that the carbon footprint of edge devices is dominated by embodied carbon, over 80% for mobile devices, that edge devices sit idle at least 75% of the time, and that offloading one H100's worth of compute to 69 smartphones or 15 laptops yields a modelled net 8x or 4x reduction in total carbon, falling to 6x and 3.5x once communication carbon is included.

MODELLED

64. [64] Chen, Goldner, Yildiz, Mandel, Cheng, Hester, Gupta, "From Component to System: Rethinking Edge Computing Design Through a Carbon-Aware Lens", ACM SIGENERGY Energy Informatics Review 5(2), July 2025. Cradle-to-gate embodied carbon for five off-the-shelf microcontroller boards ranges from 0.52 kgCO2e (Raspberry Pi Pico) to 2.59 kgCO2e (Coral Dev Board Micro), and the most energy-efficient board is not the carbon-optimal board: the Coral wins on keyword-spotting energy while its higher embodied carbon makes it worse than an Arduino Nano for short battery-powered deployments. Bill-of-materials estimates, with the authors flagging which components rest on process-specific literature and which on generic factors. Applies to microcontroller-class devices and not to robot or vehicle compute.

MODELLED

65. [65] Pirson, T., Bol, D., "Assessing the embodied carbon footprint of IoT edge devices with a bottom-up life-cycle approach", Journal of Cleaner Production, 2021, arXiv:2105.02082. The production carbon footprint between simple and complex IoT edge devices varies by a factor of more than 150; worldwide production of IoT edge devices is estimated at 22 to 562 MtCO2e per year in 2027 depending on deployment scenario, and the authors state that comparison with two vendors' disclosures suggests their bottom-up estimates should be revised upwards by roughly 2x for truncation error. Cradle-to-gate only, so no use phase and no end of life. The 150x spread is the important figure for this review: an edge device is not a unit of account.

VERIFIED

66. [66] Dell Technologies, "Product Carbon Footprint: PowerEdge R6725", produced January 2025 using PAIA v1.4.3 (MIT Materials Systems Laboratory). Total 3,424 kgCO2e with an uncertainty of plus or minus 4,345 kgCO2e, a 5th percentile of 1,976 and a 95th percentile of 23,261; phase breakdown manufacturing 638, transport 177, end of life 25 and use 2,583 kgCO2e, over a 4.0 year life at 4,844.28

kWh per year on an EU grid at an assumed PUE of 1.0. The published uncertainty band is wider than the central estimate and the 95th percentile is 11.8 times the 5th, so any embodied-versus-operational split computed from it inherits that band. The listed component values sum to 3,423 against a stated total of 3,424, which is rounding. This configuration carries no GPU.
SELF-PUBLISHED

67. [67] NVIDIA (2025), approximately 1,312 kgCO2e per HGX H100 unit on a cradle-to-gate basis, cited through "From Cradle to Cloud: A Life Cycle Review of AI's Environmental Footprint", arXiv:2605.05416. The only accelerator-level embodied figure located, published by the party selling the accelerator and retrieved through a secondary review rather than from NVIDIA's own document in this pass. Cradle-to-gate only.
SELF-PUBLISHED

68. [68] UNITAR / ITU, "Global E-waste Monitor 2024", March 2024. Global e-waste reached 62 million tonnes in 2022, up 82% from 2010, with only 22.3% documented as formally collected and recycled, projected to reach 82 million tonnes by 2030 as the documented collection rate falls to 20%; small IT and telecommunication equipment accounts for 4.6 million tonnes at a 22% documented collection rate. Documented collection is the measurable quantity, and undocumented flows are estimated by difference and are the larger share.
VERIFIED

69. [69] Green Software Foundation, Software Carbon Intensity (SCI) Specification v1.0, reported to have achieved ISO standard status as ISO/IEC 21031:2024. It amortises embodied carbon by both time-share and resource-share, M = TE x TS x RS, where TS is the fraction of the device's expected lifespan the software occupies and RS the fraction of the device's resources reserved for it, so idle hardware time is charged to nobody and low-duty-cycle distributed hardware has no accounting home for the embodied carbon it does not amortise.
REPORTED

70. [70] Solar Roadways SR3 pilot, Sandpoint, Idaho: 13.9 m2 across 30 tiles at 1.529 kW rated capacity for $48,734 installed, about $33,000 per installed kW or roughly twenty times a utility solar plant, generating 52.397 kWh in six months, with roughly 75% of panels broken before installation, 25 of 30 malfunctioning in the first week, and shutdown in December 2018. Compiled from The Conversation (2018) and Wikipedia; the operator's primary generation log was not retrieved.
REPORTED

71. [71] Wattway solar road at Tourouvre, France: 1 km of carriageway with about 2,800-3,000 m2 of panels for roughly US$5.2 million, designed for 790 kWh/day and delivering about half of that, with part of the road demolished in 2018 for wear damage and the installation closed in 2019; rotting leaves and panel breakage were the reported causes. Le Monde reporting relayed by Global Construction Review and ScienceAlert, 2019; the operator's metering series was not retrieved and the design figure is the developer's.

REPORTED

72. [72] SolaRoad cycle path, Krommenie, Netherlands: a 70-72 m installation costing EUR 3.5 million produced 9,800 kWh in its first year, with reported first-year specific yields of 73 kWh/m2/yr (2014 version) and 93 kWh/m2/yr (2016 version) declining to about 41 kWh/m2/yr as top-layer light transmission degraded. The coating detached in December 2014, was fully replaced in October 2015, cracked again in February 2017, and the experimental top layer was removed in November 2020 and replaced with ordinary asphalt. A cycle path carries far lighter loading and less soiling than a highway lane, so these figures are optimistic for a trafficked road surface.
REPORTED

73. [73] IEA-PVPS, "Soiling Losses: Impact on the Performance of Photovoltaic Power Plants". Soiling costs photovoltaics 3-5% of annual energy production globally, accumulating at roughly 0.1-0.3% additional power loss per day without rain and reaching 5-20% over extended dry periods, with above-20% losses common in dust-intensive regions and one desert system losing up to 60% after six months uncleaned. These are figures for tilted, cleanable utility and rooftop arrays.
VERIFIED

74. [74] "Development of a Pavement-Embedded Piezoelectric Harvester in a Real Traffic Environment", Sensors, 2023, PMC10181361. A month of monitoring at the BR-290 federal highway toll plaza in Brazil produced a peak of 55.6 microwatts from a single transducer with a 16 g tip mass and about 50 mWh per month from 16 transducers housed in four steel boxes of 400 x 100 x 50 mm, under roughly 1,500 vehicles per day at 40 km/h, with peak voltages of 0.1-0.8 V by vehicle class.
VERIFIED

75. [75] "A high density piezoelectric energy harvesting device from highway traffic: system design and road test", Applied Energy, 2021. Reports 2,381 mW maximum output and a power density of 23.81 W/m2 in road tests at 50 km/h; other designs in the same literature report 3.4 mW and 2.6 mW under simulated wheel loads and 736 microwatts from 48 rectified cantilevers in parallel. The headline is a peak instantaneous figure under a passing axle and not a time average.
VERIFIED

76. [76] Road thermoelectric generator literature: Applied Energy (2022), Renewable Energy (2022), Journal of Electronic Materials (2022) and MATEC TranSET (2019). A 900 cm2 road thermoelectric generator system reached a maximum 6.207 V and 700.49 mW; a field system peaked at 162.33 mW in summer at a 37.3 degrees C gradient and 34.63 mW in winter at 13.7 degrees C; a South Texas installation of two modules at a 34 degrees C gradient produced 34 mW; and a prototype sustained about 10 mW continuously for 8 hours from a 6.4 x 6.4 cm module. Every one of these systems needs a hot side in the pavement and a cold side in air or flowing water. The summer-to-winter ratio of 4.7 is the only seasonal measurement located anywhere in this review and it is thermal rather than solar.
VERIFIED

77. [77] Ambient RF energy harvesting surveys: Khemar, et al., IET Microwaves Antennas and Propagation (2018), finding average RF power in the 0.9-3 GHz range around -12 dBm near Paris with the highest average outdoor power density -7 dBm/m2 on UMTS-2100 and most cases between -35 and -10 dBm/m2; and Mimis, et al., IET MAP (2015), finding indoor averages below 1 nW/cm2 across urban and suburban sites near Bristol.
VERIFIED

78. [78] Hygroelectric generator literature: Nature Communications 15:47652 (2024), reporting a maximum 60.4 microwatts/cm2 and an average 3.0 microwatts/cm2 sustained over three months outdoors; Yao and Lovley, Nature 578 (2020), the original Air-gen protein-nanowire device at about 0.5 V and 17 microamps/cm2; and Advanced Energy and Sustainability Research (2025) on graphene-oxide films at 0.42 microwatts/cm2 transient improving to 27 microwatts/cm2 with asymmetric GO. These are laboratory-fabricated films in open outdoor exposure.
VERIFIED

79. [79] "Feasibility of Highway Energy Harvesting Using a Vertical Axis Wind Turbine", Energy Engineering 115(2), 2018, and related prototype studies: up to 48 W from a highway-induced average wind speed of 4.4 m/s on a roadside vertical-axis turbine, and 0.5 W at 2.78 m/s rising to a maximum 7.5 W at 11.11 m/s for a small composite-bladed unit. These outputs come from swept areas on the order of a square metre on a shoulder-mounted turbine.
VERIFIED

80. [80] MLCommons, MLPerf Tiny v1.4 results, July 2026. ASYGN's ColibriNPU completed a Visual Wake Words inference in 22.2 microjoules, which MLCommons notes would let a CR2032 coin cell run one inference per second for more than three years; Syntiant's NDP120 completed a keyword-spotting inference in about 4.3 ms consuming 35 microjoules at 30 MHz. MLPerf Tiny energy results are submitter-run on the submitter's own hardware under the EEMBC EnergyRunner harness with cross-submitter review. All four tasks are very small: 96x96 grayscale person detection, ten-keyword spotting, CIFAR-scale classification and anomaly detection, and that limit governs the reading of section 8.
REPORTED

81. [81] Syntiant NDP120 power figures from the vendor's own release, stating that the part can run an always-on hotword detector at under 280 microwatts with a 3.3% duty cycle. The vendor's figure for the vendor's part, without an independent harness.
SELF-PUBLISHED

82. [82] Federal Communications Commission, 5.9 GHz Second Report and Order and subsequent orders, FCC 24-123 and DA 25-125, 2024-2025. After designating DSRC as the standard for intelligent transportation services more than twenty years earlier, the Commission found that DSRC has not been meaningfully deployed and that the mid-band spectrum has largely been unused for decades, reallocated the lower 45

MHz of 5.9 GHz to unlicensed use, stopped issuing new DSRC licences on 11 February 2025 and bars DSRC operation after 14 December 2026. The stated reasons are spectrum use, technology transition to C-V2X and deployment economics.

VERIFIED

83. [83] US Department of Transportation / Federal Highway Administration, Connected Vehicle Pilot Deployment Program results and findings, ROSA P 68128, and ITS Deployment Evaluation lesson 2025-L01267. Three sites, New York City DOT, Wyoming DOT and the Tampa Hillsborough Expressway Authority, with a combined $42 million from September 2015; the published findings record lessons on roadside unit procurement, design, installation and testing and a recommendation to augment roadside units with more traditional ITS technologies.

VERIFIED

84. [84] Vaire Computing, Ice River test chip results, measured March 2025 and published September 2025, reported by EE Times and eeNews Europe: an energy-recovery factor of 1.77 for a capacitor array and 1.41 for an adder circuit at a 500 MHz data rate in 22 nm CMOS, above the proof-of-concept threshold of 1.0, with commercially published adiabatic systems reporting about 50% recovery on real workloads. The ratios originate with the company that built the chip and reach this review through trade press, so they are the vendor's own measurement. Adiabatic dissipation falls roughly as RC over clock period, so recovery improves only by slowing the clock.

REPORTED

85. [85] Loihi 2 benchmark literature, 2024-2025, including Meyer, et al. (2024) on streaming state-space models and Abreu, et al. (February 2025) on matmul-free spiking language models, collected via arXiv:2503.18002v2. Reported measured advantages span roughly 2x to 145x and are strongly workload-dependent: about 20x on energy-delay product for a converted robotics control pipeline against GPU; 145x energy and 7x latency against an embedded GPU module in streaming state-space modelling; 2-4x lower energy on small-batch audio and video; and about half the joules per token for a matmul-free spiking language model against edge GPU transformer inference. Most results are produced by or with the hardware vendor's research group and few have independent replication. All are core-level: none includes the sensor, analog-to-digital conversion, radio or storage a deployed unit must also power. The advantage is largest for temporally sparse event-driven input and smallest for dense batched work.

REPORTED

86. [86] "A comparative review of deep and spiking neural networks for edge AI neuromorphic circuits", Frontiers in Neuroscience, 2025. In a reported gesture-recognition comparison a conventional deep network reached 92.1% accuracy at 320 mJ while a spiking network reached 89.7% at 105 mJ, a 67% energy reduction for a 2.4 percentage point accuracy loss; across spiking convolutional networks the reported reduction in energy per inference is 50-80% relative to the ANN counterpart, and the accuracy gap against structurally equivalent ANNs is described

as a persistent impediment. A single task comparison inside a review article whose ANN baseline optimisation level is not established. Accuracy loss matters differently for a wake-up trigger than for a safety-relevant detection.

VERIFIED

87. [87] Innowattech piezoelectric highway pilot, run from 2009 with the Technion and the Israel National Roads Company on stretches of the Ayalon, Coastal and Trans-Israel highways with devices 5 cm below road level, reported by the company to generate 200 kWh for a single lane and 1 MWh across four lanes, relayed by ISRAEL21c, IEEE Spectrum and NoCamels, 2009-2012. The figures are the vendor's and the units are stated without a time base or road length in the sources located, so they cannot be converted to a power density.

SELF-PUBLISHED

88. [88] Savings claims for decentralized and distributed GPU compute networks, aggregated from vendor and trade marketing for Akash, Aethir, io.net, Render and Clore, 2025-2026, ranging from 45% to 90% cheaper than AWS with AWS on-demand list pricing as the explicit comparison baseline. Claims by parties selling the alternative. None publishes the workload, batch size, utilisation or availability terms under which the comparison holds, and none of the providers' actual transacted prices is audited in this review.

SELF-PUBLISHED

89. [89] Institute for Physical AI @ Bailey Military Institute, Charlot Lab, Technical Report TR-2026-25, "The Energy-First Turn", v0.3, 2 August 2026. Cited for substrate and model-class coverage, which this review cross-references rather than duplicates. Graded self-published because it is a prior report of the author's own institution, under the same tie-break applied to every other party in this review.

SELF-PUBLISHED

90. [90] Institute for Physical AI @ Bailey Military Institute, Charlot Lab, Technical Report TR-2026-26, "Building the Energy Compute Future", v0.1, 4 August 2026. Source of the resource-claim standard applied in sections 8 and 9, of the composed-path caution in its section 5, of the Landauer floor of about 2.9 zeptojoules per erased bit and the roughly eight orders of magnitude between it and picojoule-scale operations, and of the name-the-baseline and weakest-provenance conventions in its section 7. Graded self-published because it is a prior report of the author's own institution.

SELF-PUBLISHED

91. [91] Institute for Physical AI @ Bailey Military Institute, Charlot Lab, Technical Report TR-2026-30, "Physical AI and Logistics Management for Telecommunications", v0.1, 5 August 2026. Source of the report format, the grading convention and the tie-break used here, and of the CTIA count of 447,605 US cell sites at year-end 2024 used as a denominator in section 5. The CTIA figure is a trade-association survey of member-supplied data and covers all carriers, so the zones-per-cell-site ratio computed from it is an upper bound. Graded self-published because it is a prior report of the author's own institution.

SELF-PUBLISHED

92. [92] Author's calculations. Every figure carrying this index is computed rather than observed and is graded modelled; where an input is an assumption rather than a source, the sentence carrying the figure says so. Derivations include: average continuous power from annual energy (20.1 GW); reproduction of the 74-132 GW range at the stated 50% utilisation and its re-run at 70% (53.0-94.5 GW); the 18.3% compound growth rate; the 92.4% indirect water share and the 0.375 L/kWh consistency check; the 41.7% non-accelerator and 8.3% idle shares of per-prompt energy; the 5.21x ERCOT queue-to-peak ratio; closed-loop ceilings of 142.4, 57.9, 26.2, 22.7, 22.4, 21.0, 6.25 and 1.5 Hz and addressable transport shares of 59.4% and 60.8%; the 5.3 ms fibre propagation floor; separation distances of 26, 95, 157 and 1,320 mm at K = 2,000 mm/s; the policy network share of 15.1%, cadence share of 17.2% and 327 ms of headroom; 19.2 MW against 1.2 GW at 1.6%; about one zone per 22,380 cell sites; EUR 19.4 per MWh of heat value, the 20.4% heat fraction implied by a 20% reuse quota, and 35% of the Stockholm target; per-accelerator-hour rates of $6.8800 and $2.5869, discounts of 62.40%, 60.00% and 56.80%, the 6.38% gap left by a 60%-off claim, premium multipliers of 1.2500, 1.3497, 1.3462, 1.7513, 1.1995 and 1.3738, egress-to-transit ratios of 324x and 583x, the 0.616 token break-even, capital penalties of 2.5x and 1.43x, and a 14.2% owned-device break-even at an assumed $3,000 price; server embodied shares of 24.5%, 10.4%, 7.7% and 46.4%, the implied 133 gCO2e/kWh, and the 43 gCO2e/kWh crossover; embodied shares of 14.9% and 62.3% in the single published 7B row, their 4.2x ratio, the 1.96% break-even duty cycle, and the 93.6% water check; utilisation ratios of 2.1x, 4.2x, 1.8x and 1.05x; marginal-to-average factor ratios of 1.62 and 1.99 and the 373 gCO2/kWh conversion; road-surface photovoltaic specific power of 0.861, 4.68 and 10.62 W/m2 and the 0.78% capacity factor of the weakest installation; per-100-cm2 outputs of 8.61-106.2 mW, 8.14 mW, 300 uW, 4.28 uW and 0.53 uW and photovoltaic shares of 96.6-99.7% and 50.5-92.6%; 388-4,784 inferences per second at raw harvest and 39-478 after the assumed 90% derate; continuous areas of 0.66-8.13 m2 and 1.41-17.4 m2 with shortfalls of 66x-813x and 141x-1,740x; duty fractions of 0.12-1.52% and 0.057-0.71% and burst intervals of 66 s to 13.6 min; the assumed one-third winter floor giving 2.87-35.4 mW, 13-160 inferences per second and 0.041-0.51% duty; the 4.7 seasonal thermoelectric ratio; 43.5% and 29.1% saved at reversible-logic recovery factors of 1.77 and 1.41; and three-company first-half 2026 property, plant and equipment of $245.69bn. Working was executed and checked programmatically on 5 August 2026.
MODELLED

Market and economic review, not investment advice. Figures carry the verification grade under each reference: **verified** (peer-reviewed or independently replicated), **reported** (a named source stated it, not independently confirmed), **self-published** (the organisation's own figure) and **modelled** (computed here, not measured). A

modelled figure is never presented as a measurement. Corrections are welcome and will be recorded.