THE MISSING METER

WHAT EVERY INFRASTRUCTURE BUILDOUT LEAVES BEHIND, AND THE UNIT AI STILL LACKS

A PREPRINT

Sidney Scott
The Ashby Institute
sidney@theashbyinstitute.org

ABSTRACT

Two facts are true simultaneously. The four biggest tech companies are projecting a combined $725 billion in capital expenditures in 2026, the largest private infrastructure program in history. And the price of the thing infrastructure produces is collapsing: querying a model at GPT-3.5 capability fell more than 280-fold in under two years. There is no contradiction in recording physical investment and collapsing unit economics at the same time. The defining characteristic of a tipping point is that every general purpose infrastructure has crossed: railroads in the 1840s, electricity in the 1900s, payments in the 1970s, container shipping in the 1960s, and the internet in the 2000s. On the other side of each turning point: a standardized unit of account, a coordination layer that priced and routed work, and a migration of value away from the owners of physical assets to the operators of that layer. AI has the physical layer and is getting the protocols. It does not yet have the meter, or the unit the meter would read, which the paper refers to as the *served token*. The unit is not without precedent. Since 1988, the Transaction Processing Performance Council has priced useful work under an enforced deadline, with an audited cost basis, and a correctness floor; what has never happened is the instantiation of that pattern for generative inference, or its migration from a benchmark harness to production traffic. This paper describes the historical pattern, states its boundary condition, answers the strongest argument against it, specifies the unit in falsifiable terms, and shows three consequences the specification forces into the open: that the cost of useful work is a function of offered load with an interior optimum well below engineering capacity, that two production services on identical weights and hardware differ fortyfold in that cost while every metric in commercial use reports them as identical, and that the standard telemetry bus cannot presently resolve the deadlines the interactive market is buying against. Results are computed on 63,824 requests of public production inference traffic released by two independent operators and on public provider benchmarks.

Disclosure. Every figure here comes from public filings, published benchmarks, and primary sources. Contested and estimated numbers are marked as such, and every figure carries a provenance label under the taxonomy of Appendix A.6. The author builds and invests in commercial infrastructure informed by this argument, and has drawn on discussions with researchers and investors across the field for feedback on earlier drafts. The reference implementation, the record schema, the test suite, and the code for every computed exhibit are included as ancillary files with this submission.

Keywords Inference economics · Unit of account · Service-level objectives · Goodput · Price-performance benchmarking · AI infrastructure · Standardization

THE ARGUMENT IN SIX CLAIMS

Each claim below is counterintuitive, load-bearing, and stated with the condition under which it is wrong.

1. **The meter is not missing because it is hard. It is missing because AI has not adopted an instrument that has existed since 1988.** The Transaction Processing Performance Council has priced useful work under an enforced deadline, with an audited three-year cost basis, a correctness floor, and a mandatory third-party audit, for thirty-eight years. Two consortia now hold adjacent halves of the unit AI needs and neither has joined them. *Wrong if* a cross-vendor, production-traffic, service-level-conditioned cost unit is already in commercial use and this survey missed it.
2. **The cheapest tokens in the market are, for a large class of buyers, the most expensive placement available at any price.** Fifteen providers serve one open-weight artifact at an 8.6-fold price spread and a 21.7-fold speed spread. Under an interactive deadline the cheapest provider delivers approximately zero compliant output, so its cost per unit of useful work diverges while its advertised price stays lowest in the market. Raw dollars per million tokens ranks all fifteen identically for buyers whose correct rankings are opposite. *Wrong if* latency-conditioned and raw price rankings coincide in practice.
3. **The correct utilization for an inference fleet is not one hundred percent, and every metric in commercial use says otherwise.** Cost per unit of useful work is U-shaped in offered load, and economic capacity sits strictly below engineering capacity. Across 63,824 requests of real production traffic from two independent operators the measured optimum is 15 percent utilization at a ten-millisecond deadline and is strictly below engineering capacity at every deadline tested, against the 83 to 95 percent a queueing model predicts. A paper publishing only the model would have been directionally right and wrong by a factor of five. Dollars per GPU-hour is flat across that entire range; raw dollars per million tokens falls monotonically and therefore instructs the operator to run at saturation. *Wrong if* the interior optimum vanishes on a workload where it has not yet been tested; on the two tested it is present and severe.
4. **A trillion-dollar industry's standard telemetry cannot resolve the latency bounds its most valuable workloads are sold against.** OpenTelemetry's default histogram boundaries for time-per-output-token begin at ten milliseconds. An interactive objective of six milliseconds, tighter than MLPerf's Interactive bound and well inside what voice agents require, falls in the first bucket and cannot be measured at all. Every figure quoted against it is an interpolation across a discarded distribution. *Wrong if* the conventions are revised or per-request latency is standardized on spans; that is Prediction 8, and its cost is one line in a configuration file.
5. **Value does not always migrate to the coordination layer, and two of the five historical cases in this paper contradict the thesis.** Semiconductor fabrication kept its margin with no unit of account ever appearing, because the physical layer is not contestable. Cloud computing's coordination layer was bundled into an asset owner before a neutral standard could set, which is why the internet row of this paper's own table names AWS rather than a clearing house. The pattern holds only where the physical layer is contestable and the coordination function is standardized before it can be captured. *Wrong if* either condition proves unnecessary; Prediction 7 is the test.
6. **Whether any of this is true cannot currently be scored, and that is the paper's central claim rather than an evasion.** No public series of GPU utilization by age cohort exists anywhere, so the depreciation dispute is unresolvable by measurement. Enterprise AI adoption is reported at 20, 50, and 70 percent by three credible instruments, and one survey's measured rate doubled when a question was reworded. The bubble debate compares capital expenditure in dollars to demand in vibes. *Wrong if* the industry transacts in a work-denominated unit by end-2029; if it does not, the central claim of this paper fails on its own kill criterion.

Sections 8 through 10 defend each claim; Appendix A specifies the unit; Predictions 1 through 8 are the paper's full falsification surface, each dated.

1 Two facts that cannot both persist

In the first quarter of 2026, Google, Amazon, Microsoft, and Meta told their investors they intend to spend approximately $725 billion on capital expenditure this year, up roughly 77 percent from a record $410 billion in 2025 [1]. Goldman Sachs projects a combined $5.3 trillion from these four companies between fiscal 2025 and fiscal 2030 [2]. Investment in information processing equipment and software, about 4 percent of US GDP, was responsible for 92 percent of US GDP growth in the first half of 2025 [3]. Strip it out and the American economy grew at a 0.1 percent

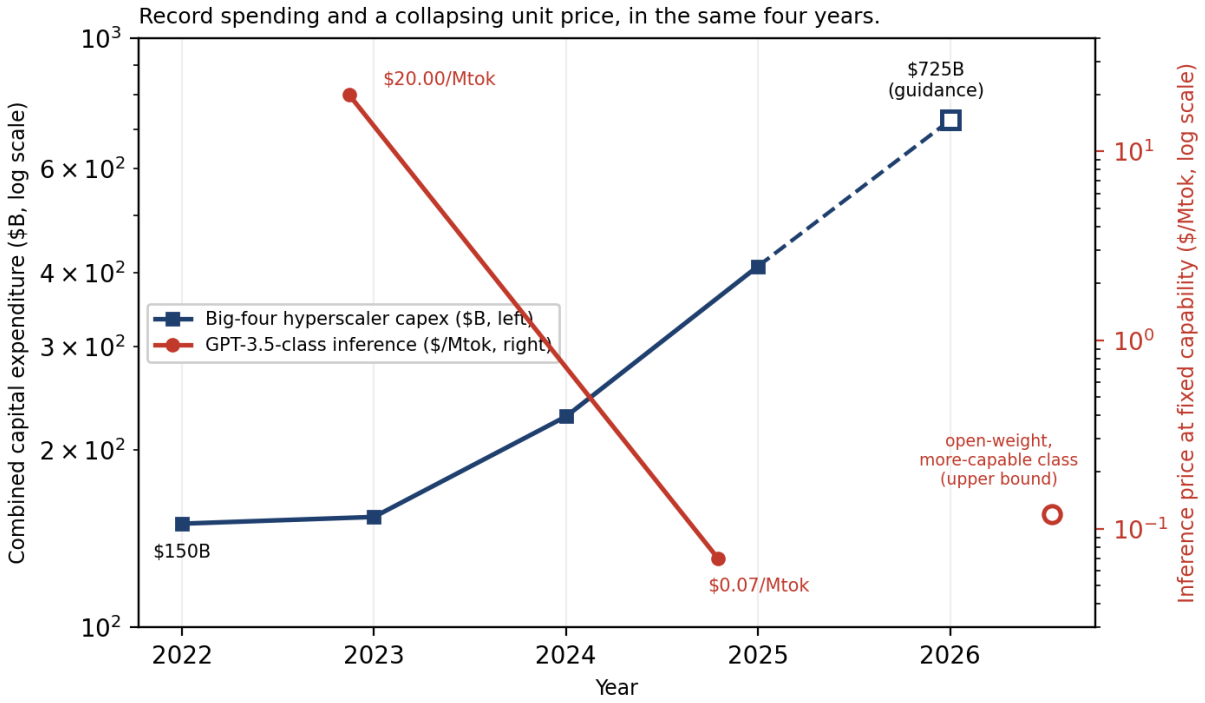

Figure 1: Combined capital expenditure of the four largest hyperscalers (left axis, $B, log scale) against the price of inference at fixed GPT-3.5-equivalent capability (right axis, $/Mtok, log scale), 2022 to 2026. Spending compounds at roughly 48 percent a year while the unit price of the output falls roughly 95 percent a year. Capex totals are company-wide, not AI-only; 2022 to 2024 are compiled from company reports [68], 2025 to 2026 per [1], with 2026 shown as guidance (open marker). Cost points are the Stanford AI Index series [4]; the 2026 open marker is a strictly more capable open-weight model's blended price, an upper bound on the same capability class, never connected to the series [38]. Capex [SPEC], prices [MEASURED, THIRD PARTY], 2026 points [EST].

annual rate. The decomposition is contested in an instructive way: because much of the equipment is imported, the net domestic contribution to GDP is far smaller, on one estimate 0.24 percentage points of the half's 1.6 percent annualized growth [45]. The dispute changes what the spending does to GDP arithmetic. It does not change the spending.

At the same time, the unit price of the output is in freefall. Stanford's 2025 AI Index measured the cost of GPT-3.5-equivalent capability falling from $20.00 per million tokens in November 2022 to $0.07 by October 2024, a 280-fold reduction in twenty-three months, with task-level declines ranging from 9x to 900x per year [4]. H100 rental prices fell from a 2023 peak near $8 per hour to $1 to $2 by late 2024 [5]. In January 2025, a Chinese lab's demonstration that frontier reasoning could be reproduced at a fraction of assumed cost erased $589 billion of Nvidia's market value in a single trading day, the largest one-day loss in stock market history [6].

Markets are treating these two facts as a paradox, and the paradox has a name: bubble. The historical record suggests something more specific. Peak physical spending and collapsing unit costs, at the same moment, is what the crest of an infrastructure installation phase looks like. It looked exactly like this in 1846, in 1929, and in 2000. What follows the crest is not the death of the industry. It is the arrival of the industry's second act, which in every prior case was larger than the first, and which in every prior case was captured by different companies, operating a different layer, measured by a unit that did not exist during the buildout.

This paper makes four claims. First, the pattern is real: five prior infrastructures followed the same four-phase arc, and the mechanism that drives the arc is visible in AI today. The pattern also has a boundary condition, stated in Section 2, and two of the five cases sit on the far side of it. Second, the strongest objection, that AI capital depreciates too fast for the pattern to hold, is partially correct and must be answered rather than dismissed; the answer changes what to build, not whether the transition occurs. Third, the binding constraint on the efficiency era is now measurement. AI is not the first trillion-dollar infrastructure that lacks a unit for its output; it is the first that has one available, in an adjacent domain, and has not adopted it. Fourth, once such a unit is specified with care, it forces two facts into the open that no

current metric can express: the cost of useful work has an interior optimum in offered load, and the industry's standard telemetry cannot resolve the latency bounds its most valuable workloads are sold against.

2 The pattern, and its boundary

Infrastructure revolutions follow a fixed sequence, and it has been mapped. Carlota Perez, studying five technological revolutions since 1771, named the structure [7]: an installation period, financed by speculative capital, builds the physical substrate in a frenzy that always overshoots. A turning point, usually a crash, transfers the substrate from speculators to operators at a discount. A deployment period follows, in which production capital puts the overbuilt substrate to work, and the technology's real economic payoff arrives, decades after the initial breakthrough and years after the bubble that supposedly discredited it.

The frenzy is not a defect in the process. It is the financing mechanism. No rational allocator funds 80 million miles of fiber against demonstrated demand; only a mania does, and the deployment era then inherits the fiber at cents on the dollar. As the venture capitalist Fred Wilson compressed Perez's corollary: nothing important happens without crashes [8]. The economics beneath the pattern are not mysterious: information industries run on increasing returns, in which the leader's advantage compounds rather than dissipates [9], and general-purpose technologies pay off on a J-curve, with measured productivity lagging years behind the intangible investment the buildout requires [10].

Infrastructure	**Installation frenzy**	**Turning point**	**Unit of account**	**Coordination layer**	**Deployment-era winner**
Railways (UK)	263 railway acts in 1846; capital tripled in a decade	1847 crash; shares fell $\sim$66%	ton-mile	timetables, clearing houses	industrial economy on cheap freight
Electricity	competing plants, per-lamp pricing, $\sim$6% utilization	consolidation, 1900s	kilowatt-hour	metering, load balancing	the utility model; 20¢ to 2.5¢/kWh by 1909
Payments	1958 Fresno card drop; 22% delinquency	1970: banks surrender the network	the cleared transaction	authorization, settlement	Visa: $>$50% net margins owning nothing
Container shipping	incompatible boxes, port-by-port chaos	ISO standardization, 1968	TEU	intermodal scheduling	freight \$5.86 to \$0.16/ton
Internet	$\sim$\$500B–\$1T telecom capex; 80M miles of fiber	2000–02: NASDAQ -78%; $\sim$95% of fiber dark	metered Mbps, then the API call	TCP/IP, DNS, virtualization	AWS, Google, streaming, on stranded fiber

Table 1: Five installments of the pattern, compressed. Sources: railways [11]; electricity [12]; payments [13][14][15]; containers [16]; internet [17][18][19].

Two regularities in this table do the argument's work.

The bust never killed the industry. Britain's railway shareholders were ruined in 1847; Britain kept the 8,000 miles of track, and the track kept lowering freight costs for a century [11]. Telecom equity holders were wiped out in 2001; the fiber stayed in the ground, bandwidth prices fell up to 90 percent, and YouTube, founded in 2005 on that glut, generated roughly \$31 billion for Alphabet in 2023 and passed \$60 billion across advertising and subscriptions in 2025 [20]. Capital is destroyed at the turning point. Capacity is not. The write-down is the mechanism by which the next era acquires its inputs below cost.

Value migrated to the layer that measured and routed, not the layer that owned. Visa issues no cards, extends no credit, and owns no banks; it operates authorization, clearing, and settlement, and in fiscal 2024 it carried 234 billion transactions and roughly \$16 trillion of volume at net margins above 50 percent [15]. Samuel Insull did not win Chicago by generating more electricity than his competitors; he won by metering demand, discovering that diverse customers have diverse peak loads, and using the meter to sell the same fixed capital many times over. His unit, the kilowatt-hour, plus his metric, load factor, converted electricity from a per-lamp luxury into a utility, cutting its price 87 percent and growing his customer base 40-fold across two decades [12]. The box that Malcolm McLean standardized cut cargo handling from \$5.86 to \$0.16 per ton, a 97 percent collapse, and the TEU became the unit in which the entire logistics industry denominates itself [16]. In each case the physical layer was necessary, competitive, and margin-poor. The measurement-and-coordination layer was singular, standard, and margin-rich.

2.1 The boundary condition

The second regularity has a boundary, and the table itself marks it. Value migrates to the coordination layer when two conditions hold together: the physical layer is *contestable*, meaning several substitutable owners compete to supply it, and the coordination function is *standardizable and non-excludable*, meaning no single participant can own the rules and still have them adopted. Payments satisfied both, with thousands of banks and a consortium constituted so that none could capture it. Containers satisfied both, with many carriers and an ISO standard. Railways and electricity satisfied both after consolidation forced the issue.

Where either condition fails, value stays with the owner, and this paper's own table contains two such cases. Semiconductor fabrication fails the first: process capability is not substitutable across suppliers at the leading edge, and TrendForce puts a single foundry at roughly 70 percent of the global pure-play market for 2025 on revenue of $122.5 billion, with no competitor above eight percent [71].[1] No metering layer above fabrication can extract, so the foundry holds the margin and no unit of account ever appears. Cloud computing failed the second, in a specific and instructive way: virtualization and the API call *were* the coordination layer, and they were bundled into an asset owner before any neutral standard could set, which is why the internet row of the table names AWS rather than a clearing house. The lesson is not that the pattern failed there. It is that the coordination layer was captured by an owner rather than constituted as a consortium, and the window in which that capture is possible closes once a standard exists.

This is what makes the present moment specific rather than merely analogous. AI inference satisfies the first condition unusually well: the same open-weight artifact is served by fifteen substitutable providers at an 8.6-fold price spread, which is the definition of a contestable physical layer, and Appendix A.9 computes with those numbers. Whether it satisfies the second is undecided and is being decided now, in the donation of MCP, A2A, and AP2 to neutral foundations on one side and in vertically integrated serving stacks on the other. Prediction 7 is the test, and the AWS case is the outcome in which this paper is directionally right about the layer and wrong about who owns it.

The comparison is no longer a private analytic choice. The Bank for International Settlements, in its June 2026 Annual Economic Report, plots the AI boom against canal mania, railway mania, the 1920s, and the dot-com buildout on a single chart, and observes that every prior episode ended in an investment reversal and recession [50]. The central bankers' chart and this paper's exhibit contain the same history and serve opposite purposes. Theirs is a warning about the fall. This paper accepts the warning and maps what, in every prior case, was standing on the far side.

3 Act I: the bust that builds

The crash destroys capital, never capacity. It is worth watching this happen twice, because the two runs bracket the industrial era and agree on every particular.

Britain, 1845 to 1850. Parliament approved 263 railway acts in 1846 alone, authorizing roughly 9,500 miles of new route [11]. Paid-up railway capital more than tripled inside the decade, from about £30 million to over £100 million by 1849 [11]. George Hudson, the Railway King, held the mania's mirror up early: he was paying dividends out of new subscribers' capital, a structure that acquired its modern name only decades later. Monetary tightening and the commercial crisis of 1847 called the loans; shares fell about two thirds by 1850, and only about two thirds of the authorized mileage was ever built [11]. The investors' losses were permanent. So was the track's usefulness.

The United States reran the experiment with fiber, 1996 to 2006. The Telecommunications Act of 1996 opened the field; carriers laid roughly 80 million miles of fiber against demand that did not yet exist; annual capex peaked near $120 billion in 2000, with cumulative investment estimated well beyond $500 billion [17]. The NASDAQ closed at 5,048.62 on March 10, 2000, and fell roughly 78 percent over the following thirty-one months [18]. WorldCom filed the then-largest bankruptcy in American history in July 2002; the sector produced more than sixty bankruptcies [18]. Then the second act: with the overwhelming majority of the fiber dark and bandwidth prices down as much as 90 percent [19], the crash's stranded capacity became the deployment era's raw material. YouTube was founded in 2005 because the bandwidth glut made consumer video streaming economically thinkable; Google bought it in 2006 for $1.65 billion; YouTube's revenue across advertising and subscriptions passed $60 billion in 2025, returning the purchase price roughly every ten days [20].

The founding dates tell the same story from the other side. Google incorporated in 1998 and went public in 2004, across the crash. Salesforce was founded in 1999, Facebook in 2004, YouTube in 2005; AWS launched in 2006 on

[1] Definitions differ materially. Counterpoint's broader "Foundry 2.0" measure, which folds in packaging and adjacent segments, yields roughly 39 percent for the same firm. Both are analyst estimates, not audited filings. The pure-play measure is used here and stated as such.

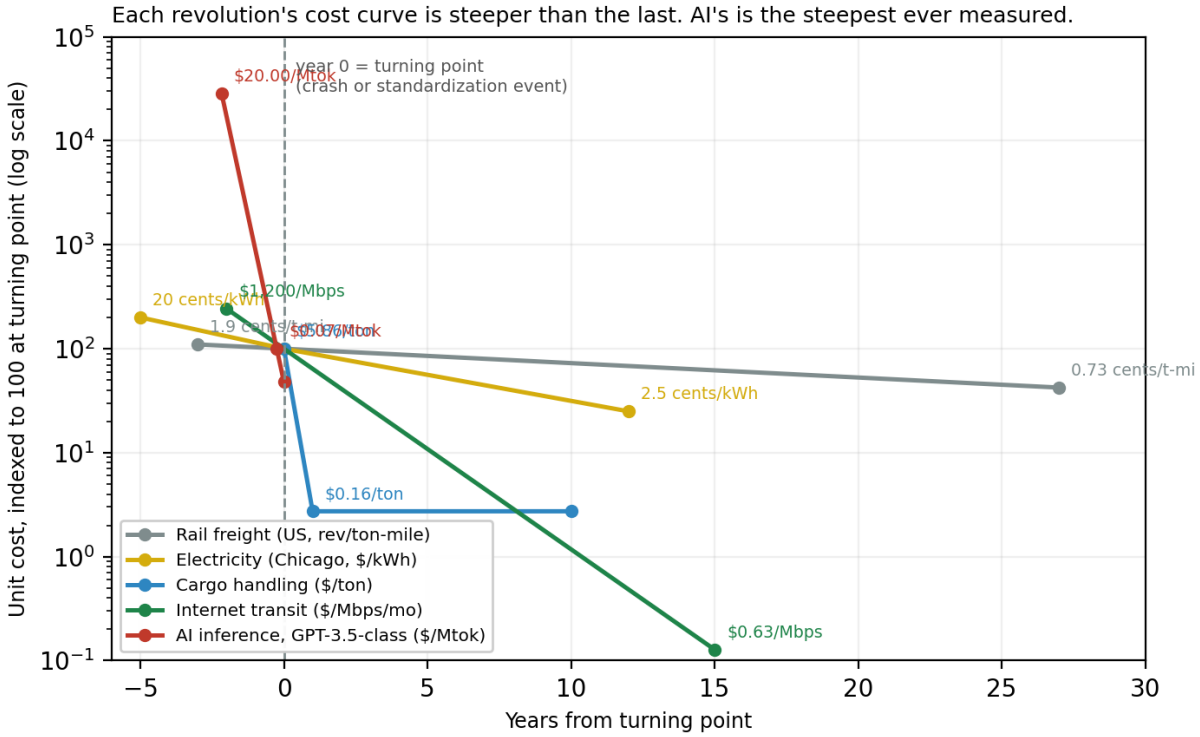

Figure 2: Unit cost, indexed to 100 at each turning point (rail: Panic of 1873; electricity: demand metering, 1897; cargo handling: Ideal X, 1956; internet transit: the 2000 crash; AI inference at GPT-3.5-equivalent capability: January 2025), log scale, against years from turning point. Measured points only; lines are log-linear connections between them, not data. Year-0 anchors are measured where a measured point falls at year 0 and otherwise log-interpolated from the nearest measured points; AI's anchor extrapolates 2.5 months beyond its last measured point. Cargo handling is a one-time step change rather than a sustained decline. Sustained annualized declines: rail 3.1 percent over 30 years, electricity 11.5 percent over 17, internet transit 35.9 percent over 17, AI inference 94.8 percent over its first two measured years. Sources: [4] [12] [16] [39] [40]. Endpoints [MEASURED], interpolations [SIM].

hardware and bandwidth economics the bust created. The companies that defined the deployment era were assembled in the wreckage of the installation era, buying its outputs at liquidation prices.

One discrimination inside these stories carries the weight of Section 6 and should be named now: what survived was the long-lived layer. The rail bed outlived the locomotive makers; the fiber outlived the carriers; the right-of-way outlived the operators. Busts do not spare assets uniformly. They spare the assets whose useful life exceeds the mania's attention span.

4 Act II: the meter

Electricity's installation period built generators. Its deployment period was built by a measuring instrument.

When Samuel Insull took over Chicago Edison in 1892, the company served about 5,000 customers and was one of roughly two dozen electric operators competing in a city of a million [12]. The industry's structural problem was utilization, and nobody could see it, because electricity was sold the way gas had been, per lamp and at flat rates, so a customer's bill had no relationship to the capital his usage pattern consumed. On an 1894 visit to Brighton, England, Insull found the instrument that made consumption visible, Arthur Wright's maximum-demand meter, licensed it, and by 1897 had introduced Wright's two-part demand tariff in Chicago, the first in America [12].

Visibility revealed the structural fact the flat rate had hidden: different customers peak at different hours, the factory by day, the streetcar at dusk, the household at night. Diversity of peaks meant the same turbine could serve all three in rotation, so Insull bought load in every form he could find, pursuing the streetcar companies from 1902 until, by 1909,

a handful of transit operators consumed almost twice the energy of every other light and power customer combined; he measured everything in one unit, the kilowatt-hour, and managed to one metric, load factor, the ratio of average to peak demand [12]. The numbers were the argument: by 1914 Commonwealth Edison's customers held 26,640 kilowatts of combined potential demand against an actual system peak of 9,770, a diversity factor near 2.7, the same capital sold nearly three times over [12]. Average revenue per kilowatt-hour fell from twenty cents in 1892 to a dime by 1897 and 2.5 cents by 1909; customers grew from 5,000 to 200,000 by 1913 and roughly half a million by 1920 [12]. The meter did not record that business. The meter created it.

One more honesty the record requires: the demand meter was not a neutral cost-allocation device. Economic historians read the two-part tariff as institutionalized price discrimination, engineered to undercut the customer's alternative of generating his own power [12]. The meter grew the market by pricing each buyer at what the load was worth, and the reader will recognize the same mechanism operating, vendor by vendor, in the service tiers of Section 8.1. And there is a coda, instructive rather than decorative: Insull's holding-company pyramid collapsed in 1932, he was prosecuted three times and acquitted three times, and he died in a Paris Metro station in 1938 with thirty francs in his pocket against fourteen million dollars of debts [12]. The financier died broke. The unit, the meter, and the utility model did not notice.

Cargo ran the same play with a box. Malcolm McLean's converted tanker *Ideal X* sailed from Newark on April 26, 1956, carrying 58 containers of standard dimension [16]. Hand-loading break-bulk cargo cost $5.86 per ton; loading the standardized box cost $0.16, a 97 percent collapse achieved not by better ships or stronger cranes but by agreeing on the shape of the thing being measured [16]. ISO standardized the container in 1968, the twenty-foot equivalent unit became the industry's denominator, and world trade reorganized around a unit of account that fits on a flatbed.

Between them, the meter and the box state this chapter's law: a standardized unit, made measurable, converts a luxury into a utility, and the price collapse that follows is not a threat to the industry but the mechanism of its expansion.

5 Act III: the interchange

Payments is the cleanest precedent, because its turning point has a date, a room, and a governance document.

Bank of America dropped 65,000 unsolicited BankAmericards on Fresno, California in September 1958 and within a year had saturated the state with more than 2 million cards and 20,000 merchants [13]. The installation chaos was total: delinquency ran to 22 percent against the roughly 4 percent planned, fraud and credit losses approached $20 million, and the program's architect resigned before its second Christmas [13]. The product was real and the losses were real at the same time, which is the signature of every installation period. The industry's response was the signature of every turning point: not retreat, but rules.

The breaking layer was settlement between banks. Interchange was bilateral, paper-based, and adversarial, and by 1968 it was failing at scale. Dee Hock's insight was structural: the network could be trusted only if it was owned by its members and controlled by none of them, and in 1970 Bank of America surrendered the program to National BankAmericard Inc., a consortium constituted so that no member could capture it [14]. What NBI built next was not a better card. It was the coordination stack: BASE I in 1973 replaced telephone-chain authorization with electronic approval in seconds; BASE II in 1974 replaced weeks of paper clearing with overnight electronic settlement, saving an estimated $15 million in its first year; the renamed Visa put magnetic stripes and terminals at the point of sale by decade's end [14].

The economics on the far side are this paper's exhibit for where value settles. Visa issues no cards, extends no credit, holds no consumer relationships, and owns none of the banks; in fiscal 2024 it reported $35.9 billion of net revenue on 234 billion processed transactions and roughly $16 trillion of volume, at net margins above 50 percent [15]. The banks that own the physical layer of the relationship compete. The layer that owns the rules does not. Every agent-payment protocol announced in the last eighteen months, and there have been three, is a bet that this history is about to run again with machines as the cardholders.

6 The strongest objection, taken seriously

Every analogy to fiber and rails runs into an immediate and substantially correct objection: GPUs are not rails. A rail bed laid in 1846 moved freight in 1946. Dark fiber lit in 2005 was, physically, the same fiber buried in 1999. But an H100 purchased in 2023 is economically obsolete on a three-to-five-year schedule, and the models trained on it obsolete in months. If the assets scrap instead of stranding, the central mechanism of the pattern, cheap inherited substrate seeding the deployment era, fails. Michael Burry made the aggressive version of this argument in November 2025, estimating that hyperscalers will understate depreciation by $176 billion between 2026 and 2028 by stretching

useful-life assumptions [21]. His figure is his own estimate, is disputed, and could not be independently confirmed by the outlets that reported it; it is carried here [EST] because the paper argues about the dispute rather than from the number. The accounting question underneath it is not disputed.

The honest answer decomposes the buildout, because "AI capex" is not one asset class.

A data center is a stack of depreciation schedules. Land, shells, substations, transmission interconnects, cooling plant, and fiber depreciate over roughly 14 to 40 years; these are the rails, and they are on the order of 40 percent of the up-front cost of a modern AI campus [23]. Power is the extreme case: grid interconnection rights now carry multi-year queues [22] and are arguably the scarcest asset in the entire stack; they do not depreciate at all. The constraint has grown severe enough that the industry now expects tens of gigawatts of new datacenter capacity to be powered behind the meter by 2028, bypassing the interconnection queue entirely [48]. Silicon depreciates over 3 to 6 years, but cascades rather than scraps: frontier training silicon becomes inference silicon becomes batch and embedding silicon, each stage serving a lower-value workload at a lower implied price, which is itself a mechanism for the efficiency era to inherit capacity below cost. Models depreciate in months, and this is the objection's strongest ground: model weights are the first infrastructure asset in this table with negative shelf life, since a stale frontier model is worth less than the electricity to serve it.

So the objection is right that this cycle's write-down will be harsher and faster than fiber's, concentrated in silicon and weights. It is wrong about what survives. The pattern does not require every asset to strand gracefully; it requires the deployment era to inherit capacity cheaply, and the durable half of the stack (power, shells, interconnection, and the cascading tail of silicon) is precisely the capacity the efficiency era runs on. The 1840s investor who owned locomotives lost differently than the one who owned the right-of-way. The discrimination between them is the work of the next sections' unit, and the refusal to discriminate, one blended "AI capex" number amortized on one blended optimistic schedule, is exactly the accounting fog Burry is pointing at. The sensitivity is not hypothetical: an independent cost model of a one-gigawatt campus finds that moving the assumed IT lifespan from five years to three raises the annualized cost of the same physical facility from $8.5 billion to $12 billion, a 41 percent swing produced by nothing but the depreciation assumption [23].

This is also the thesis's primary falsification condition, stated now rather than buried: if GPU useful lives prove to be two to three years and cascading fails to preserve resale utilization, the inheritance mechanism breaks, and the deployment era must be financed a second time at full price. Watch the disclosed depreciation schedules and the secondary GPU market. If second-life silicon clears at utilization above roughly 60 percent, the pattern holds. If it scraps, this paper is wrong.

Four further objections deserve full strength rather than a footnote.

First, scaling may not be finished, and value may stay concentrated at the frontier. The frontier labs are still betting that larger models and test-time compute deliver returns, and the live version of the objection was stated by a prominent public investor the week this paper was first drafted: cheap, mostly open-source tokens are likely already the majority of volume, while the majority of economic value still accrues to the most intelligent models [41]. If that concentration persists, margin stays at the model layer and the routing thesis stalls. This paper does not require scaling to stop; it requires the frontier's lead to be rentable rather than owned, which is what the token-share prediction in Section 10 tests.

Second, this bubble's deflation could destroy more than the last one did. The 2000 crash burned equity; this cycle finances itself with debt and with itself, and the warning is now institutional. The BIS's June 2026 report names the AI capex boom a systemic pressure point: the five largest hyperscalers are spending more than $1 trillion across 2025 and 2026 while outpacing their earnings and free cash flow, circular financing is defined by name, and deal terms are disclosed so poorly that the same asset being pledged multiple times is a live risk [50]. Its companion bulletin quantifies the channel with the least oversight: private-credit loans to AI-related companies grew from roughly $3 billion in 2010 to more than $40 billion in 2025 [51]. Nvidia has committed up to $100 billion to OpenAI, which has committed roughly $300 billion to Oracle, $90 billion to AMD, and $38 billion to AWS against reported revenue near $13 billion [36]; and Bain estimates the buildout implies roughly $2 trillion of annual revenue by 2030 and comes up $800 billion short [37]. If the loop unwinds through credit rather than equity, contagion could run faster and wider than 2002. The pattern survived 1847 and 2001. It is not entitled to survive everything.

Third, demand may simply lag, and the strongest version of this objection is measured, not asserted. An MIT Media Lab study of enterprise deployments, published in mid-2025 against an estimated $30 to $40 billion of enterprise generative-AI spending, found 95 percent of pilots produced no measurable profit-and-loss impact, a condition its authors called high adoption but low transformation while flagging their own work as a directional snapshot [65]. This objection also carries a sting in its tail that belongs to Section 8: the demand it worries about cannot currently be measured. As of mid-2026 the Census Bureau's business survey puts AI adoption near 20 percent of firms, spend-

based measurement from corporate card data puts it near 50 percent, and executive surveys put it near 70 percent [66]; when the Census reworded its survey question in November 2025, measured adoption roughly doubled without a single firm changing its behavior [66]. A firm-level study of 21,559 companies states the condition plainly: AI adoption does not have a single empirical rate [67]. The demand-lag objection is real, and it is also unscorable with current instruments, which is this paper's central claim restated as a survey problem.

Fourth, the framework itself is soft. Perez's phases are a map drawn mostly in hindsight, and critics correctly note that the boundaries are unfalsifiable as stated. That is precisely why this paper stakes itself on the dated predictions of Section 10 and on the specification of Appendix A rather than on the pattern's authority: the history motivates the thesis; only the predictions and the unit can carry it.

7 The turn is already measurable

The claim that AI is crossing its turning point does not rest on forecast. It rests on prices.

The output price collapsed before the input spend peaked. The 280-fold decline in GPT-3.5-class inference cost, the 9x-to-900x annual per-task declines, hardware cost falling roughly 30 percent annually and energy efficiency improving roughly 40 percent, and the open-versus-closed model performance gap narrowing from 8 percent to 1.7 percent in a single year [4]: these all predate the 2026 capex guidance. In prior cycles, unit price collapse followed the crash. In this one, efficiency arrived while the frenzy was still accelerating, because the efficiency is architectural (mixture-of-experts, distillation, quantization, speculative decoding, test-time compute) rather than merely scale-driven. The installation and deployment phases are overlapping, which compresses the timeline but does not change the sequence of what wins. If that overlap seems to blur the turning point, note what the term actually marks in every prior cycle: not the day spending stops, but the day the marginal dollar starts earning more in the coordination layer than in the physical one. By that definition the point is crossed when the meter beats the machine, and Sections 8 through 10 argue that crossing is now.

The efficiency shock has already had its market event. DeepSeek's January 2025 release, matching OpenAI's o1 on reasoning benchmarks at $0.55 per million input tokens against $15, triggered the largest single-day market value loss in history [6][24]. The reported $5.6 million training figure was misleading, since it excludes cumulative R&D and a GPU fleet estimated above $500 million [25], but the market's inference was correct even where the number was not: capability had decoupled from capex. Satya Nadella's same-day response, invoking Jevons's 1865 paradox, was the correct macroeconomic reading and the correct microeconomic warning at once [26][27]. Cheaper intelligence means more intelligence consumed; it also means the margin moves to whoever decides, per task, how little intelligence is enough. The rental market then priced the paradox directly: H100 rates that had collapsed to $1.70 per hour by October 2025 rose about 40 percent within five months as reasoning and agentic workloads absorbed the surplus [5].

The buyers have started measuring. By mid-2026 the strategic vocabulary had flipped in public. Perplexity's chief executive, whose company arbitrages every frontier model against every other, put it as: the model alone is no longer the product, it is the harness, the orchestration system, and the answer is always to use whatever is best for the task [28]. Benchmark's Peter Fenton, in the same piece, predicted 90-plus percent of tokens will come from open-weight models within 18 to 24 months. Cisco disclosed token spending approaching $10,000 per employee per year and stood up an internal routing discipline to control it; Palo Alto Networks' chief executive stated AI pricing needs to fall 90 percent [29]. And the largest buyer of all has conceded the era its critics named: Microsoft's chief executive, urging employees away from frontier models for non-frontier problems, acknowledged the tokenmaxxing habit while instructing his company out of it [44].

The protocols are arriving in the historically correct order. Payments needed interchange, then authorization (BASE I, 1973), then settlement (BASE II, 1974). Between late 2024 and early 2026, AI acquired the same three layers in the same order: the Model Context Protocol standardizing how agents reach tools and data, open-sourced in November 2024 and donated to a neutral foundation in December 2025 [30]; Agent2Agent standardizing how agents reach each other, donated to the Linux Foundation within three months of launch [31]; and three competing agent-payment rails standardizing how agents settle, with AP2 passing onward to the FIDO Alliance [32][33][34]. The donation of these protocols to foundations is not altruism; it is the participants' recognition, learned from TCP/IP's victory over proprietary networking, that a coordination layer is only valuable if it is universal, and only universal if no one owns it. The rails are being laid. What runs on rails is traffic, and traffic requires a tariff, and a tariff requires a unit.

8 The missing meter

Here is the strangest fact about a trillion-dollar industry in 2026: it cannot answer the question "what does your product cost?" — and it is not because the instrument is unknown. The fused unit has existed, in a governed and audited form, since 1988.

The Transaction Processing Performance Council pairs a throughput denominator with an audited cost basis under enforced response-time bounds and a correctness floor, and publishes the result as price-per-tpmC under multi-party governance with mandatory third-party audit. TPC-C requires that 90 percent of each interactive transaction type complete within five seconds, and Stock-Level within twenty; it mandates ACID properties as an audited requirement rather than a vendor claim; it fixes the cost basis as a three-year total cost of ownership including maintenance, under a published Pricing Specification rather than vendor discretion; and it obliges a Full Disclosure Report that a competitor can attack [72][73]. Databases have known how to price useful work under a deadline since before the web existed.

The claim of this section is therefore narrower and harder than "no such unit exists." It is that the pattern has never been instantiated for generative inference, and has never migrated from a benchmark harness to production traffic. Stated precisely: as of mid-2026 there exists no unit of account for machine intelligence that (a) fuses a dollar denominator, a service-level condition, and an output-quality floor; (b) is computed over the buyer's own production workload rather than a fixed benchmark corpus; and (c) is governed by a multi-party body. Every word is load-bearing, and the survey in Section 8.1 is organized by which leg each near-miss amputates and which domain it misses.

Compute is bought in GPU-hours. It is consumed as answered queries, completed tasks, resolved tickets, and generated tokens, each under a latency constraint that determines whether the output was worth anything at all: a voice agent's reply that arrives in 200 milliseconds is a product, and the same reply in four seconds is a refund. Between the GPU-hour on the invoice and the useful-work-under-a-deadline on the income statement, there is no standard unit. Dollars per million tokens is the leading candidate and the industry's de facto price sheet, but a raw token is a kilowatt-hour with no voltage standard: tokens vary in quality, in urgency, and in usefulness, and a token of correct answer and a token of hallucination invoice identically. Palantir's Alex Karp channels the buyer's version of the objection: enterprises, he reports, are paying for tokens that create no value [35]. The complaint is right about the numerator and wrong to stop there. The kilowatt-hour did not succeed because electrons are homogeneous. It succeeded because the meter, the unit, and the tariff made them commensurable: firm power and interruptible power, peak and off-peak, all denominated in one unit with quality and reliability priced as explicit terms.

The historical sequence is exact enough to be a specification. Before Insull's meter, electricity was priced per lamp: a flat fee per connected bulb, whatever it consumed, exactly as AI subscriptions today price per seat whatever the seat consumes. Per-lamp pricing made load invisible, so plants ran at 6 percent utilization, so capital costs stayed brutal, so electricity stayed a luxury. The demand meter made consumption visible; visibility revealed that customers peak at different hours; diversity of peaks meant the same turbine could serve the factory by day and the streetcar by night; load factor became the metric that turned utilization into strategy; and the price fell 87 percent while the market grew 40-fold.

AI's equivalent unit must denominate what the buyer actually buys, and what the buyer buys has three legs: cost, useful output, and a service-level condition. Something of the form dollars per unit of verified useful work, conditional on a stated latency and quality floor, is the shape the unit must take; every simpler candidate amputates a leg. This paper names that unit the **served token**, abbreviated SVT, and specifies it in Appendix A: one output token delivered inside its service-level objective and above its quality floor, within a fixed model artifact, with the compliant task as the primary and cross-artifact denominator. Dollars per GPU-hour drops the output. Dollars per raw token drops the quality and the deadline. Benchmarks drop the price. Uptime SLAs drop everything but the deadline. The industry's current metrics are the per-lamp tariffs of 1895: each defensible, none commensurable, all jointly guaranteeing that capital cannot compare placements and that depreciation schedules can blur because no external unit prices what the depreciating asset actually produces.

The consequences of the missing unit are the industry's three most-discussed pathologies, usually discussed as if they were unrelated. Circular financing is only sustainable, and only dangerous, because no standard unit prices the useful work at the end of the loop; revenue commitments denominated in dollars of future compute purchases are unfalsifiable precisely because compute's output has no meter [36]. The depreciation dispute is unresolvable without a unit: whether a four-year-old GPU is an asset or scrap depends entirely on the cost of useful work it can still deliver under some objective, a number no current disclosure states. And the "is it a bubble" debate itself is a measurement failure: Bain's estimate that the buildout requires $2 trillion of revenue against a $800 billion shortfall [37] compares capex in dollars to demand in vibes, since the demand side has no unit either. The meter is not a bookkeeping nicety. It's the tool that allows the turning point to work its way through to deployment, not paralysis.

8.1 The prior art, and which leg each one amputates

The industry and the standards bodies have built every leg of the meter separately, which makes the survey of near-misses the strongest evidence that the fused unit is missing rather than impossible. Each candidate below is real, useful, and incomplete in a specific, documentable way.

The whole unit exists, for a different domain, on a fixed corpus. TPC-C (Revision 5.11, February 2010) and TPC-H (Revision 3.0.1, April 2022) report price-performance — price-per-tpmC, and dollars per thousand queries-per-hour at a stated scale, $/kQphH@Size — under enforced response-time constraints and a correctness floor, audited by a TPC-certified auditor under consortium governance, with the cost basis fixed by a published Pricing Specification rather than by the vendor [72][73][74]. What they do not do, and were never intended to do, is measure the buyer's own traffic: the workload is a fixed synthetic corpus, the system under test is a vendor-assembled configuration, and the result describes a purchasable system rather than a consumed service. The migration from harness to production traffic is precisely the open problem, and TPC's pricing rules, audit function, and Full Disclosure Report are the strongest available template for solving it. The specification in Appendix A is an attempt to port that template across a domain boundary, not to invent one.

TPC has already crossed into AI, and stopped short of the deadline. TPCx-AI (Revision 2.0.0) benchmarks an end-to-end data-science pipeline and reports both AIUCpm@SF, AI use cases per minute at a scale factor, and its price-performance companion $/AIUCpm@SF, with a Full Disclosure Report and a priced configuration [75]. It is the closest existing artifact to a priced AI unit and it is instructive precisely where it stops: it imposes no per-request tail-latency constraint, so a submission may aggregate stage runtimes by geometric mean without ever stating what fraction of individual requests arrived in time. TPCx-AI has the cost leg and the work leg and not the service-level leg; MLPerf Inference has the service-level leg and the quality leg and not the cost leg. The two consortia have between them assembled every component of the unit and have not joined them.

The cost leg is standardized. The FinOps Foundation's FOCUS specification (v1.4, ratified 4 June 2026, under the Linux Foundation) defines a vendor-neutral schema for cost and usage data, and since v1.2 in May 2025 has normalized billing in non-monetary units explicitly, defining a virtual currency as a provider-defined unit of account such as a credit, a token, or a DBU [52]. Exports are offered by the major clouds. FOCUS answers what was spent and on what, across vendors. It contains no useful-work denominator and no service-level or quality dimension anywhere in its schema. The Foundation's own AI working group states the goal of linking cost-per-unit-of-work to business value while conceding the industry lacks generally accepted frameworks for it [53].

The counting leg is standardized, experimentally, and cannot yet resolve the deadlines the market buys against. OpenTelemetry's generative-AI semantic conventions standardize token accounting as span attributes, `gen_ai.usage.input_tokens` and `gen_ai.usage.output_tokens`, with an aggregate metric `gen_ai.client.token.usage` split by token type. Latency is standardized separately and differently: `gen_ai.server.time_to_first_token` and `gen_ai.server.time_per_output_token` are histogram *metric instruments* in seconds, not span attributes, and the whole generative-AI convention set remains at Development status as of SemConv v1.40.0 in April 2026 [54]. Two consequences follow, and the second is the more serious.

First, a histogram is an aggregate. It can report what share of requests fell below a bound, which is exactly a compliance rate, but it cannot report *which* request failed, so it cannot be joined to a per-request quality result. The counting bus as standardized will therefore support the service-level leg of this unit and not the quality leg, and per-request latency remains a vendor extension.

Second, and concretely: the default explicit bucket boundaries for time-per-output-token begin at ten milliseconds and step next to twenty-five. An interactive service-level objective of six milliseconds per output token — comfortably inside what a voice agent requires, and considerably tighter than MLPerf Inference's forty-millisecond Interactive bound — falls inside the first bucket and cannot be resolved from default telemetry at all. Any figure quoted against it is an interpolation across a distribution the histogram has already discarded, and under this paper's provenance rules is [SIM] rather than [MEASURED]. The standard measurement bus cannot presently see the deadlines the interactive market is buying against. Practitioners running the open-source stack report the same condition from the other side: the serving engine, the router, the gateway, and the cache tier each expose metrics, and the fleet-level quantities that determine whether the system is economic, cache hit rate across replicas and where a slow request spent its time, are assembled by nobody [89]. The stack is observable in pieces and opaque as a whole. This is a narrow, fixable defect, and fixing it is the smallest concrete contribution this paper can ask of an existing body: add finer boundaries below ten milliseconds to the time-per-output-token histogram, or standardize per-request latency as span attributes so that the join to quality becomes possible. It is registered as Prediction 8.

The service-level and quality legs are standardized, without cost. MLPerf Inference's Server scenario is the strongest prior art in this survey and the paper credits it as such. Its LLM benchmarks enforce 99th-percentile bounds on the two quantities Appendix A specifies — for Llama 2 70B Interactive, time-to-first-token no worse than 450 milliseconds and time-per-output-token no worse than 40 milliseconds, with the standard variant at 2,000 and 200 milliseconds respectively, and Llama 3.1 405B at 6,000 and 175 — at enforced accuracy floors of 99 or 99.9 percent of a reference model [55]. This is the service-level and quality half of the meter, operating since 2024, governed by a consortium, at v5.1 as of results published in September 2025. It has no dollar denominator in any scenario, runs vendor-submitted hardware against fixed benchmark datasets rather than production traffic, and yields a throughput figure. If MLCommons adds a cost denominator to the Server scenario, the unit this paper calls for will be most of the way to existing, and that possibility is registered as one of two nearest live falsifiers.

The service-level-conditioned denominator is standardized in the research literature, without price or quality. *Goodput* — throughput counted only over requests meeting declared latency targets — has a long prior life in networking and was carried into machine-learning systems through cluster scheduling before arriving at inference [79]. DistServe gives the definition this paper builds on: the number of completed requests per second adhering to service-level objectives, where the objectives are attainment targets on p90 time-to-first-token and p90 time-per-output-token, reported as per-GPU goodput [76]. The surrounding work on chunked prefill and phase disaggregation optimizes against the same denominator [77][78], and the disaggregated architecture that separates the compute-bound prefill phase from the bandwidth-bound decode phase onto distinct pools is now the reference design for serious serving stacks [88]. Goodput is the served token's service-level leg, already rigorous, already load-swept, and already the working vocabulary of every serious serving system. What it has never carried is a price or an acceptance test: a goodput-optimal placement serving fluent nonsense scores identically to one serving correct answers, and goodput can rank two schedulers but cannot rank two purchases. The served token is goodput with the invoice attached and the quality floor declared, which is the difference between a systems metric and a unit of account.

The rate-per-useful-work specification is standardized, for carbon. The Green Software Foundation's Software Carbon Intensity specification, ratified as ISO/IEC 21031:2024, defines $\mathrm{SCI} = ((E \times I) + M)/R$, where R is a functional unit declared by the practitioner — per user, per API call, per transaction — and the specification is explicit that SCI is a rate and never a total [80]. It is structurally the same object as the unit proposed here with carbon in the numerator, it is multi-party governed, and its route from consortium draft to ISO standard is the closest available precedent for how a rate-per-useful-work specification gets ratified. Read it as the process template.

The energy denominator is standardized, without dollars or quality. SPECpower_ssj2008 reports overall ssj_ops/watt, measured across a graduated ladder of ten target load levels in ten-percent increments plus active idle, with the metric formed as total operations over total average power across all eleven points [81]. It establishes two things this specification needs: that a consortium can standardize a rate per unit of useful work, and that load-swept rather than single-point reporting is a viable practice. Section 8.2 argues that load-swept reporting is not merely viable here but mandatory.

The cost and verified-output legs are combined, without a service level. The ARC Prize reports model performance as a two-axis matrix of verified score against measured dollars per task, and recent academic work has formalized the same idea as cost-of-pass, the expected monetary cost of a correct solution [56]. Two legs, rigorously joined, on a single benchmark, unconditioned on latency: a batch unit, not a service unit.

The cost and service-level legs are combined, vendor by vendor. Azure's provisioned throughput units, Google's generative AI scale units, and OpenAI's priority, standard, flex, and batch tiers all price differentiated service levels, which proves service-level-conditioned pricing is commercially viable today [57]. None is portable across vendors.

And the token itself is not a unit. Different tokenizers produce different token counts for identical text, so a token is not comparable across providers even before quality and latency enter; when one lab changed tokenizers in 2026, measured costs on identical prompts moved by double-digit percentages [58]. Appendix A.3 makes this a rule rather than a caveat.

The pattern across all ten is exact, and it has a precedent. In 1881 electricity had a dozen competing units of electromotive force and fifteen of resistance; it took three International Electrical Congresses, from Paris 1881 to Chicago 1893, to resolve them into the standard units the meter could then price [59]. AI inference in 2026 is not quite electricity in 1881, because the units are not competing — they are disjoint. TPC has price and correctness without the deadline. TPCx-AI has price without the deadline. MLPerf has the deadline and the quality floor without price. Goodput has the deadline without either. FOCUS has the money without the work. OpenTelemetry has the counting without the price, and cannot resolve the tightest deadlines. SPECpower and SCI have the rate form with a different numerator. Every leg is measured somewhere by someone competent, priced nowhere in common, and no congress has been convened. A 2025 survey of the field reaches the same verdict in its own words: there is still no widely adopted, rigorous,

end-to-end framework for measuring and comparing these dimensions across heterogeneous hardware and software [60].

8.2 The unit is load-dependent, and that is the whole of load factor

A value of C is undefined without an offered load, for the same reason a kilowatt-hour price is undefined without a time of day. Latency rises with utilization; compliance is a threshold function of latency; W is compliance times billed output. The unit therefore inherits the queueing behavior of the system it measures, and the inheritance is not a nuisance term. It is the mechanism by which the meter tells an operator something no current metric can.

Consider a serving node rented at a fixed hourly rate, offered Poisson traffic, under a stated bound on p99 time-per-output-token. At low utilization the rental amortizes over too little delivered work and C is high. As load rises, C falls. Past a knee, queueing pushes the latency distribution across the bound, compliance collapses, and W falls faster than throughput rises, so C turns and climbs. C is U-shaped in offered load and has an interior minimum.

Name that minimum the **economic capacity** of a placement, as distinct from its **engineering capacity**, the load at which the system saturates. Exhibit 3 computes both under an M/M/1 approximation at five service-level tightnesses, with λ the offered load, μ the service rate, D the per-request deadline implied by the TPOT bound, and compliance $P(T \leq D) = 1 - e^{-\mu(1-\rho)D}$ for $\rho = \lambda/\mu$ [82].

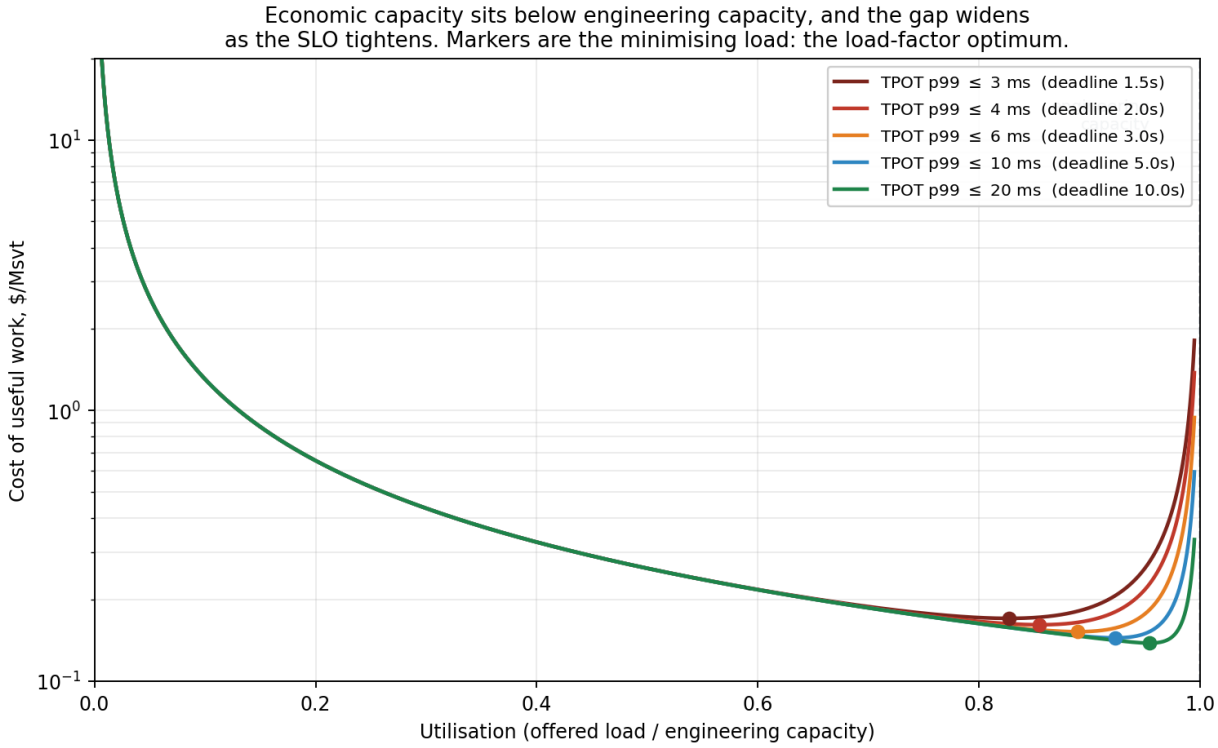

Figure 3: Cost of useful work against utilization, at five service-level tightnesses. Markers are the minimizing load: the load-factor optimum. Engineering capacity is 10 requests per second in all curves. [SIM]: M/M/1 FCFS, 500 output tokens per request, node rented at $2.35 per hour. Code in the ancillary files.

Three readings, in ascending order of importance.

First, **economic capacity is strictly below engineering capacity, and the gap widens as the deadline tightens.** Under a twenty-millisecond bound the two nearly coincide and an operator loses nothing by running flat out. Under a three-millisecond bound the optimum sits at 83 percent utilization, and running at 95 percent costs 53 percent more per unit of useful work than running at the optimum. Same hardware, same weights, same traffic; only the deadline differs, and the correct operating point moves.

Second, **no current metric can see this table.** Dollars per GPU-hour is flat across the entire horizontal axis, so it returns the same number at the optimum and at collapse. Raw dollars per million tokens falls monotonically in

p99 TPOT bound	economic capacity	load factor	C^* (\$/MSVT)	C at 95% util.	penalty
3 ms	8.27 req/s	82.7%	0.171	0.261	1.53x
4 ms	8.55 req/s	85.5%	0.162	0.217	1.34x
6 ms	8.89 req/s	88.9%	0.152	0.177	1.16x
10 ms	9.23 req/s	92.3%	0.144	0.150	1.04x
20 ms	9.54 req/s	95.4%	0.138	0.138	1.00x

Table 2: Economic capacity, load factor, and the penalty for running at engineering capacity. [SIM].

utilization, because it counts emitted tokens whether or not they arrived in time, so it instructs the operator to run at saturation, which is wrong under every row of the table and most wrong under the tightest. Only a work-denominated, service-level-conditioned unit has an interior optimum, and an interior optimum is the thing an operator needs to be told.

Third, and this is the section's point, **the third column is Insull's load factor.** He measured the ratio of average to peak demand, discovered that diverse customers peak at different hours, and sold the same turbine several times over by mixing them; Commonwealth Edison's diversity factor of 2.7 in 1914 was that discovery in one number. The analogue here is exact and it is computable: interactive traffic under a tight deadline has a low economic capacity and leaves headroom, batch traffic under a loose one has a high economic capacity and can consume it, and the operator who mixes deadline classes against the same fixed capital raises the blended load factor and lowers blended C. That arbitrage is the deployment era's margin, it is captured above the silicon rather than by it, and Section 9's claim about routing is this table read across its rows.

The M/M/1 approximation is deliberate and its direction of error is stated. A continuously batched inference server has a batch-size-dependent service rate, prefill and decode contend asymmetrically, and production arrivals are burstier than Poisson. All three move the knee left. The curve above is therefore the optimistic bound. Section 8.4 tests that claim against production traces and confirms it: the measured optimum falls as low as 15 percent utilisation where this model predicts 82.7, so the synthetic curve is not merely optimistic but optimistic by a factor of five.

8.3 The discontinuity is the point, and it must be published

Compliance is binary. A placement delivering 99 tokens per second against a 100-token floor produces zero served tokens and a divergent C; one delivering 101 produces a finite one. The cliff is not an artifact to be smoothed. It is the economics: a voice agent's reply arriving in 200 milliseconds is a product and the same reply in four seconds is a refund, and a unit that averaged across that boundary would report a quantity nobody buys. The same choice is made, for the same reason, by every prior art in Section 8.1 that has a deadline at all: MLPerf's Server scenario admits or rejects a submission against its p99 bounds, and DistServe's goodput counts a request or does not.

But a cliff makes the point estimate of C fragile in a way the point estimate cannot express, so the specification requires the curve alongside it. Exhibit 4 sweeps the throughput floor across the fifteen providers of the worked example and reports the cheapest compliant placement at each.

Across two thousand reconstructions of the twelve providers whose figures are not public, the identity of the cheapest compliant placement changes a median of four times as the deadline sweeps, in every reconstruction without exception, and at its worst single step a one-unit change in the deadline moves the cost of useful work by a median of 178 percent. Raw dollars per million tokens ranks these fifteen placements identically at every floor. That invariance is the defect, stated as a measurement rather than an argument.

Two disclosures follow, and Appendix A.7 requires both. A conforming record publishes the sensitivity curve, not only the point. And it publishes the **margin to the objective**: the ratio of measured percentile to bound. A placement compliant at 0.99 of its bound and one compliant at 0.40 report the same C and are not the same asset, and the depreciation question of Section 6 is exactly a question about which of the two a four-year-old accelerator has become.

8.4 The unit computed on production traces

Sections 8.2 and 8.3 are simulated and labelled so. This section computes the same quantities on real production workloads, which changes one of the paper's own claims and confirms another.

The traces are four production workloads released by two independent operators on two continents. From Microsoft, *AzureLLMInferenceDataset2023* under CC-BY: 19,366 requests to a production conversation service and 8,819 to a

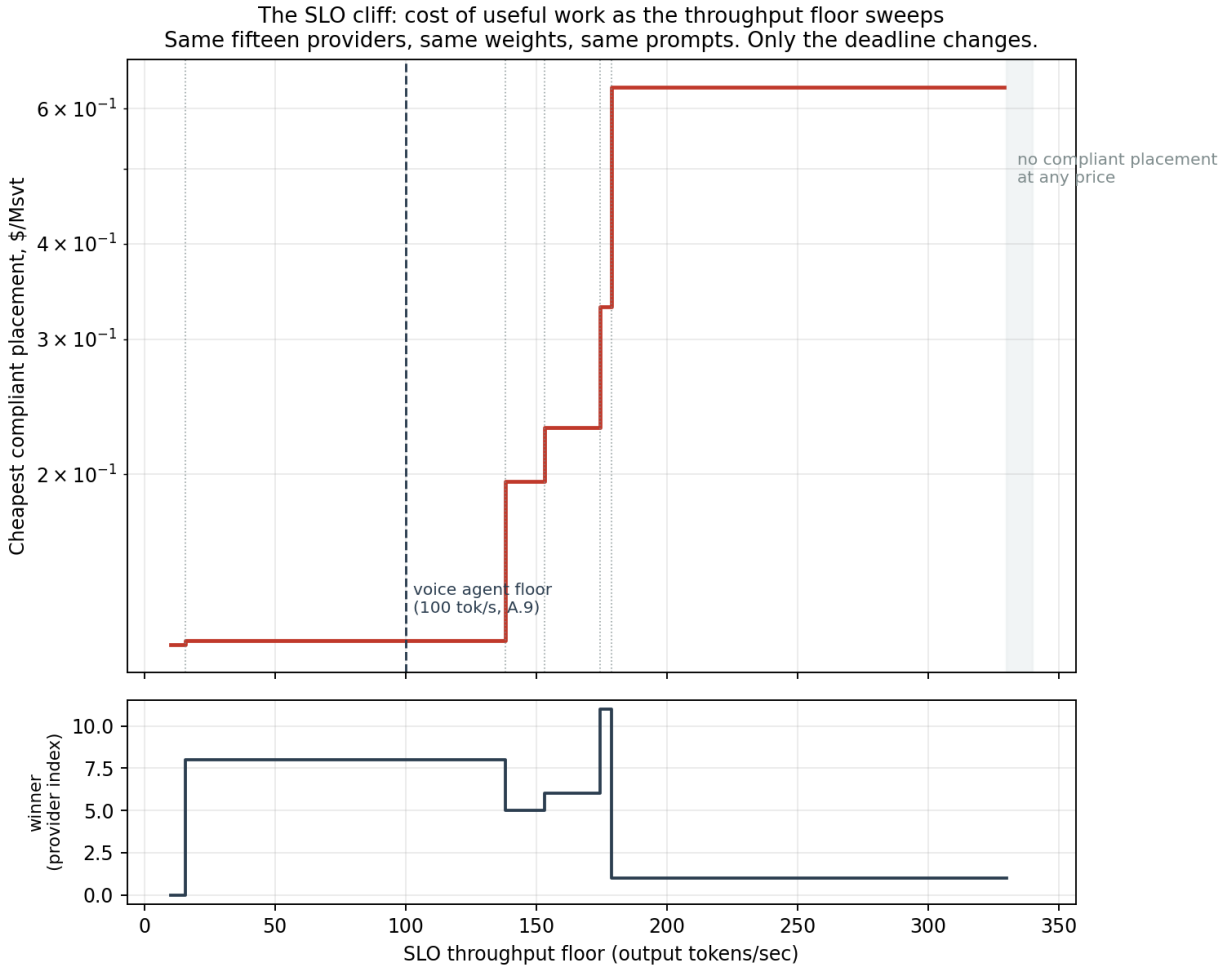

Figure 4: Cheapest compliant placement as the service-level throughput floor sweeps, across fifteen providers serving one open-weight artifact. Dotted vertical lines mark changes in the winning placement. Three anchor points [MEASURED, THIRD PARTY] from [38]; the twelve undisclosed providers [SIM] from a documented generator; derived statistics inherit the weakest label.

production code service [87]. From Moonshot AI, the traces released with the Mooncake serving platform that runs Kimi: 23,608 requests from a long-context service and 12,031 from a conversation service, sanitised in the same way and for the same stated reason, that a trace of lengths and arrival times preserves the workload while disclosing no customer content [88]. Each record carries an arrival timestamp, a context-token count, and a generated-token count; the Mooncake records additionally carry prefix-block hashes. What they do not carry is latency. No public LLM serving trace does, as of mid-2026, because arrival and size are what a billing system records and delivered-under-a-deadline is what no standard unit has made worth recording. That absence is this paper's argument appearing as a data limitation, and it dictates the method: the measured arrivals and measured token counts are replayed through a continuously batched serving model calibrated to a 70B-class node, and latency is produced by the replay. Under A.6 inheritance the derived cost figures are therefore [SIM]; what the traces contribute, and what the synthetic analysis had to assume, is the true arrival process and the true joint distribution of context and generation lengths.

Three findings, and the first is the one that matters most for how this paper should be read.

The synthetic model was optimistic, in the direction claimed and by a wider margin than claimed. Section 8.2 stated that batching, prefill-decode contention, and bursty arrivals all move the knee left, and that the M/M/1 curve was therefore an optimistic bound. It is. Against measured arrivals the economic optimum falls from 82.7 to 15 percent utilisation at a ten-millisecond deadline on the conversation workload, and to 10 percent on the code workload: the synthetic model overstates the correct operating point by 5.5-fold and 8.3-fold respectively. At a loose fifty-millisecond

Operator	Workload	Requests	Rate	Dispersion	Ctx p50	Gen p50	C^*
Microsoft	conversation	19,366	5.53/s	1.39	1,020	129	$1.06
Microsoft	code	8,819	2.57/s	13.18	1,469	13	$41.86
Moonshot	long-context	23,608	6.56/s	15.02	6,345	30	$2.80
Moonshot	conversation	12,031	3.40/s	7.89	6,909	350	$2.10

Table 3: Four production services, two operators, 63,824 requests. Counts, rates, index of dispersion, and token distributions are [MEASURED] from [87][88]; C^* is the minimum cost of useful work in dollars per million served tokens over the utilisation sweep, [SIM]. Index of dispersion is the variance-to-mean ratio of per-second arrival counts; a Poisson process has 1. Microsoft workloads are scored at a 20 ms time-per-output-token bound and a one-second time-to-first-token bound; Moonshot workloads at the 100 ms and 30 s bounds their operator declares for these services, because a long-context service is not sold against a chat deadline and scoring it against one would measure the wrong contract.

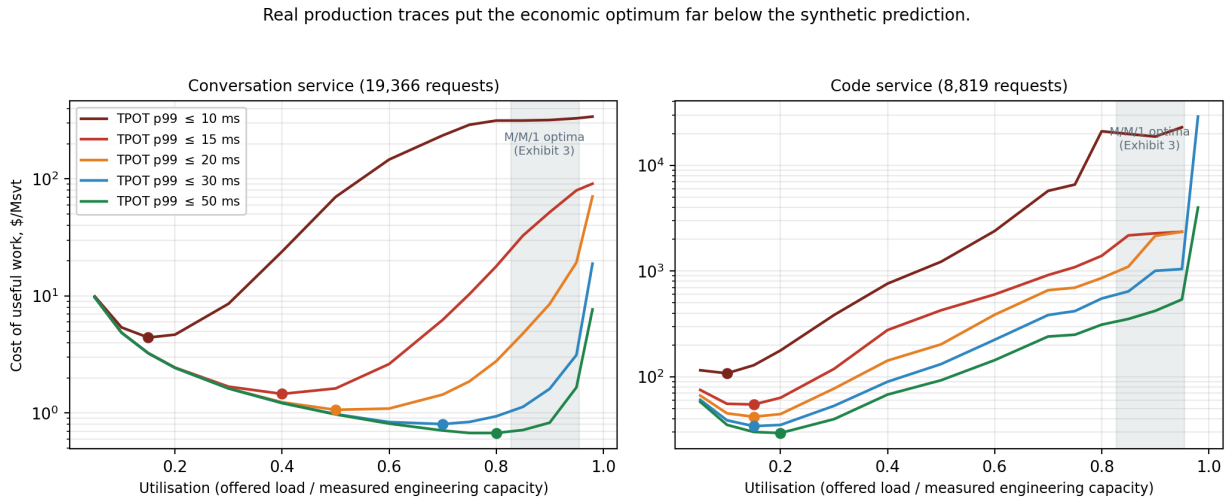

Figure 5: Cost of useful work against utilisation on the two production traces, at five deadlines. Markers are the measured economic optimum. The shaded band is the range of optima the M/M/1 model of Exhibit 3 predicted. Arrivals and token counts [MEASURED]; latency and cost [SIM].

deadline the gap narrows to 1.2-fold on conversation and remains 4.8-fold on code. A paper that had published only Exhibit 3 would have been directionally right and quantitatively wrong by most of an order of magnitude, which is the argument for the instrumentation in a sentence.

The relationship itself survives the change of operator, which is what makes it a regularity rather than an artefact of one company's traffic. Scored against the service levels their own operator declares, the Moonshot long-context workload optimises at 65 percent utilisation under a fifty-millisecond bound, 80 percent under a hundred, and 90 percent under two hundred; the Moonshot conversation workload gives the same three figures. Across four workloads, two operators, and two continents, economic capacity is monotone increasing in the deadline and strictly below engineering capacity everywhere it was measured. The level is workload-specific and the direction is not.

The direction is independently corroborated by measurement this paper did not perform. A practitioner benchmarking a single H100 on a 12B-class open model reports that raising offered load from 8 to 16 requests per second raised throughput from 1,894 to 2,069 tokens per second, a gain of nine percent, while p99 time-to-first-token rose from 474 milliseconds to 24.5 seconds, a degradation of roughly fiftyfold [89]. That is one workload on one accelerator and is cited as a practitioner benchmark rather than a controlled study, but its provenance is the complement of this section's: latency there is [MEASURED] on real hardware and the workload is synthetic, where here the workload is [MEASURED] from production and the latency is [SIM]. The two disagree about nothing. A throughput-denominated metric records a nine percent improvement across that transition and a work-denominated one records near-total collapse, which is the same fact Exhibit 5 draws as a curve.

Two production workloads on identical weights and identical hardware differ in cost of useful work by a factor of nearly forty. At a twenty-millisecond deadline the conversation service costs $1.06 per million served tokens and the code service $41.86. The reason is visible in the token distributions and invisible on any invoice: the code service has a median context of 1,469 tokens and a median generation of 13, so the node spends its capacity on prefill and emits almost nothing to be sold. Dollars per GPU-hour prices these two workloads identically because it cannot see the workload. Raw dollars per million tokens prices them identically because it counts emitted tokens without reference to what they cost to produce or whether they arrived in time. The served token separates them by a factor of 39.5 ($41.86 versus $1.06), and the separation is the entire basis on which a router or a capacity planner would act.

Production arrivals are not uniformly burstier than Poisson, and where they are, the load factor collapses. The conversation trace has an index of dispersion of 1.39, close to the Poisson value of 1. The code trace has 13.18. The paper's blanket assertion that production traffic is burstier than Poisson is therefore true of one of these two services and roughly false of the other, and the correction is more useful than the assertion: burstiness predicts the load-factor penalty. The near-Poisson workload reaches 80 percent utilisation at a loose deadline, while the thirteen-fold-dispersed workload never exceeds 20 percent at any deadline tested. An operator cannot know which regime a workload is in without measuring it, and the index of dispersion is one line of arithmetic over data every serving system already emits.

A production platform independently adopted this paper's accounting rule, and reports what it is worth. Mooncake, the serving platform behind a commercial model service, states the rule in its own words: only requests that fully complete under the service-level objective are counted, and otherwise all previously consumed and generated tokens are not counted and the corresponding resources are wasted [88]. That is Appendix A.2 arrived at independently by an operator under production pressure rather than proposed by a specification. The platform goes further in the direction this paper argues: it uses service-level attainment itself as the measure of system load, and rejects requests before prefill when the objective is unachievable, precisely because computation spent on work that will miss its deadline is spend with no served token to show for it. The reported value of getting this right is large and is measured on identical hardware. Replaying real traces across twenty nodes in each configuration, roughly 100 percent of requests met the time-between-tokens objective under the disaggregated placement against 57 percent under the coupled baseline, and the platform served approximately 75 percent more requests within the same objectives [88]. Identical spend, identical silicon, a 1.75-fold difference in compliant work, produced entirely by placement. Dollars per GPU-hour is exactly equal across that comparison. The served token is the quantity in which the difference is denominated, and Section 9's routing margin is that number.

Two consequences for Section 9's routing argument. The demand-diversity arbitrage is now quantified rather than asserted: a conversation workload under a fifty-millisecond deadline leaves 80 percent of a node consumed and a code workload under ten milliseconds leaves 10 percent, so mixing deadline classes against the same fixed capital is worth a large multiple, not a marginal improvement. And the mixing gain is largest exactly where the current metrics are most misleading, because the two workloads that differ fortyfold in C are the two that dollars per GPU-hour reports as identical.

The replay harness, the fitted workload statistics, and the sweep outputs are included as ancillary files, so a reader with the four public traces can reproduce every number in this section.

9 What the efficiency era requires

The efficiency era needs five things built, and the history dictates their order.

A unit of account with three legs. Cost, verified output, service level: dollars per unit of useful work at a stated latency percentile and quality floor. Defined by an open specification, measurable by independent parties with published methodology, and stated with provenance, since a measured number, a configured number, and a simulated number are different species and must be labeled as such. The unit that wins will be boring, auditable, and slightly too conservative, because the kilowatt-hour, the TEU, and price-per-tpmC all were. Appendix A drafts the specification and computes a worked example from public data.

Meters before markets. An instrumentation layer that measures delivered work under real workloads, as distinct from benchmark performance under ideal ones, in the same way the demand meter was distinct from the nameplate rating of the dynamo. The engineering community has already built the denominator and does not call it a unit of account. Goodput counts only work delivered inside its service-level objective, and since 2024 every serious serving system has been optimized against it. What goodput has never carried is a price or an acceptance test, so it can rank two schedulers and cannot rank two purchases. The instrumentation layer this section asks for is therefore not a greenfield build. It is goodput, already deployed, joined to a billing export that already exists in the FOCUS schema, under a quality

floor the buyer already runs as an eval. The unit is a join, not a sensor, and Appendix A.9 computes it from nothing but public numbers to prove the point. Load factor's AI analogue, delivered-useful-work over provisioned-capacity, becomes the operator's core metric, and Section 8.2 makes it computable rather than analogical. The silicon layer has already begun competing on it: new inference architectures marketed in 2026 claim model FLOPs utilization above 80 percent against the 20 to 50 percent GPUs typically deliver, and their designers' framing, that the unit of analysis is the cluster rather than the chip, is the load-factor argument restated in hardware [43].

Routing as load balancing. Once work is metered in a common unit, placement becomes an optimization: which model, which silicon, which region, which batch window, subject to the task's deadline. Fenton's 90 percent open-weight prediction is really a routing prediction, since open weights are what make placement contestable. Insull's demand diversity has an exact analogue, computed in Section 8.2: interactive traffic peaks when batch traffic can wait, reasoning-heavy tasks tolerate latency that voice cannot, and the operator who mixes them against the same fixed capital sells the same GPU many times over. That arbitrage is the deployment era's margin, and it is captured above the silicon, not by it.

Honest depreciation, denominated in the unit. Assets carried at the value of the useful work they can still deliver, disclosed by asset class rather than blended. This is the answer to Burry that the hyperscalers cannot currently give, because giving it requires the meter.

Protocols owned by no one. The MCP-to-foundation move is the correct template and should be treated as precedent: interchange succeeded because Dee Hock structured NBI so that no member bank could own it [14], and TCP/IP beat superior proprietary alternatives because it had no owner to fear. There is a formal literature on when this works: open standards win where network effects and expectations can be coordinated and lose where an incumbent can bundle before expectations converge, which is the same boundary condition Section 2.1 derives from the AWS case [83][84]. Any unit of account or measurement standard that one vendor controls will fail on arrival for the same reason.

An energy axis, measured now, priced later. The unit's natural second denominator is joules per compliant token, and unlike the quality leg it is already being measured at every scale. United States data centers drew 176 terawatt-hours in 2023, 4.4 percent of national electricity, and the national laboratory that tracks them projects 325 to 580 terawatt-hours by 2028 [69]; the IEA puts global data-centre demand near 415 terawatt-hours in 2024 and expects roughly 945 by 2030 [69]. Per-query measurement has begun from the top down: Google disclosed a median energy cost of roughly a quarter of a watt-hour per text prompt in 2025, and European law obliges general-purpose model providers to document energy consumption [70]. The form that axis should take is already standardized in an adjacent domain. ISO/IEC 21031:2024 defines software carbon intensity as a rate over a practitioner-declared functional unit and is explicit that it is a rate and never a total [80]. Joules per compliant token is that specification with this paper's unit substituted for the functional unit, which is the correct relationship between the two and the reason Appendix A.10 defers the energy term rather than inventing one.

10 Predictions

A thesis that cannot lose is not a thesis. Dated, checkable, with the conditions under which this paper is wrong.

Prediction 1. Open-weight token share. By mid-2028, the majority of inference tokens by volume are served from open-weight models. Baseline, stated with its denominator because the denominator is where this debate hides: on OpenRouter, open-weight models reached roughly one third of routed token volume by late 2025 and rose through mid-2026 [61]. But routers exclude first-party traffic, where closed models dominate: Google alone disclosed processing 3.2 quadrillion tokens per month by May 2026, roughly twenty-five times OpenRouter's entire throughput [62]. Corrected for first-party volume, the defensible whole-market open-weight share as of mid-2026 is in the low double digits. The prediction is scored against the whole-market share. It is currently false, which is what makes it a prediction. If frontier closed models still carry the whole-market majority by mid-2028, the routing thesis is overstated and the model layer, not the coordination layer, is where value is consolidating.

Prediction 2. The unit is born, or a close call is consummated. Two consortia own half the unit each and either could shut it down. The unit resolves through prior art rather than new construction, provided MLCommons adds a cost denominator to the MLPerf Inference Server scenario already subject to p99 latency bounds and reference-accuracy floors. It works out from the other direction if TPC adds a per-request tail-latency constraint to TPCx-AI, which already reports $/AIUCpm@SF under audited pricing rules. This paper counts both as the nearest live falsifiers, and as confirmation rather than refutation.

Prediction 2b. The original unit as written. End-2028 – At least one open multi-party specification for cost-of-useful-work-under-service-level published with measurement methodology and adopted by at least two major clouds or serving frameworks in pricing or disclosure. If at the end of 2029 compute is still overwhelmingly transacted in raw

GPU-hours, with no work-denominated unit in commercial use, then the central claim of this paper fails on its own kill criterion.

Prediction 3 . Downward convergence of depreciation. By the end of 2028, two of the five largest AI spenders will file abbreviated or asset-class segmented depreciation schedules for AI silicon. Baseline, from property and equipment notes of latest 10-Ks [63]: Amazon cut servers from six years to five, effective January 2025, citing the accelerated pace of technology development, especially in artificial intelligence, at a cost of about $0.7 billion of 2025 operating income; Alphabet, Microsoft, Meta and Oracle most recently extended. The prediction is therefore half fulfilled at publication: one of the two movers exists, no company yet reports accelerators as a segmented asset class, and four of five most recently moved in the direction Burry disputes.

Prediction 4. The correction, and the survival. Total tokens served continues to grow quarter over quarter without interruption through a material repricing of AI-exposed equities and credit before the end of 2028. If the correction comes and token volume also shrinks for back-to-back quarters demand was all speculative and the Jevons read was off.

Prediction 5. Second-life silicon clears. Through 2029, cascaded GPUs one generation behind sustain secondary-market utilization above roughly 60 percent for inference workloads. Baseline: none exists, and the paper reports that plainly, because no public, systematic measurement of GPU utilization by age cohort is published anywhere as of mid-2026, which is itself Section 8's argument wearing different clothes. Directional evidence points both ways: AWS's chief executive stated in January 2026 that AWS has never retired an A100 server and remains sold out of them, and Google reports full utilization on seven- and eight-year-old TPUs; against that, secondary H100 prices fall to a fraction of peak within two to three years and brokers describe the market as structurally opaque [64]. Prices and sold-out claims are not utilization rates. Scoring this prediction requires the meter it presupposes, so the prediction carries its own instrumentation clause: if no cohort-level utilization series exists by end-2027, the prediction is unscorable and is counted against the thesis, not for it.

Prediction 6. Routing is a line item. By end-2027, model-routing or inference-optimization is a named budget category or named vendor line in mainstream enterprise IT surveys, like cloud FinOps did circa 2021. Weak form check; it's not a falsification, it's a timing miss.

Prediction 7. The value migration itself. By 2030, the market value created by companies operating measurement, routing, and agent-settlement layers founded after 2023 exceeds the value created by companies whose primary asset is owned accelerators, on capital invested. This is the Visa test, and it is also the test of Section 2.1's boundary condition: if the coordination layer is instead bundled into an existing asset owner before a neutral standard sets, the outcome is the AWS case rather than the Visa case, and this paper is right about the layer and wrong about the owner.

Prediction 8. The measurement bus learns to see interactive deadlines. By end-2027, OpenTelemetry's generative-AI semantic conventions either reach stable status with per-request latency expressible on spans, or revise the default explicit bucket boundaries for the `gen_ai.server.time_per_output_token` histogram to resolve bounds below ten milliseconds. Failure indicates that the standard telemetry bus remains unable to measure the service class the interactive market is actually buying, which would slow every prediction above it. This is the smallest concrete change any existing body could make in the direction of this paper's argument, and its cost is a line in a configuration file.

11 The implication, stated in full

If the pattern holds and its boundary condition is satisfied, the most valuable layer in artificial intelligence will not be the models and will not be the chips. It will be the layer that measures, prices, and routes intelligence as work: the interchange, the meter, and the dispatcher, fused. That layer does not exist yet. Its unit does not exist yet for this domain, though it has existed for four decades in another, which is the difference between an invention and a port. The protocols beneath it arrived in the last eighteen months, the buyers began demanding it this year, and every prior infrastructure cycle produced exactly one such layer within a decade of its turning point.

The boundary condition of Section 2.1 says what would have to be true for that to happen here, and it is worth restating as a condition rather than a prophecy. The physical layer must stay contestable, which it currently is: fifteen providers serve one open-weight artifact at an 8.6-fold price spread. And the coordination function must be standardized before it can be bundled, which is undecided. Semiconductor fabrication shows what happens when the first condition fails. Cloud computing shows what happens when the second does. AI inference in 2026 has not yet failed either, and the window in which that remains true is the window this paper is about.

Railways got a clearing house. Power got a meter. Freight got a box. Money got interchange. Packets got a protocol. Databases, quietly and without anyone outside the field noticing, got price-per-tpmC in 1988, with an auditor and a disclosure report and a three-year cost basis, and have been able to answer what a transaction costs ever since.

Intelligence gets a meter next. The only open questions are whose, and whether it is neutral before it is bundled.

Acknowledgments

This draft is circulated for comment. The author thanks those who reviewed earlier versions; named acknowledgments will be added as reviewers agree to be listed. Remaining errors are the author's own.

A A draft specification for the unit

Status: draft for public comment. This appendix is deliberately boring. The kilowatt-hour is boring. Price-per-tpmC is boring. That is what made them work. A reference implementation, a JSON Schema for the conforming record, a test suite, and the code for every computed exhibit are included as ancillary files with this submission.

A.1 Scope and design goals

This appendix specifies a unit of account for machine intelligence delivered as a service, and the measurement rules without which the unit is meaningless. Design goals, in priority order: the unit must denominate what the buyer buys rather than what the seller owns; two parties with the same inputs must compute the same number; every number must carry its provenance; and no single vendor may control the definition. The unit is designed to be computed today from data that already exists, and the worked example in A.9 does so.

A.2 Definitions

A **task** is a request with a defined acceptable output. A **quality floor** is an acceptance test for the output, declared before measurement and stated operationally: an eval score threshold, a task-completion check, a programmatic assertion, a human-rating floor, or a reference-accuracy bound against a declared reference artifact in the manner of MLPerf Inference's 99 and 99.9 percent floors. Identification of the model artifact is *not* a quality floor; it is field 2 of the conditioning tuple, and a record supplying it in place of an acceptance test is non-conforming. Where testing every output is impractical, the record declares the evaluated fraction, and the resulting sampling error is propagated into the interval on C.

A **service-level objective** is the latency contract under which output has value, stated as percentile bounds on time-to-first-token and time-per-output-token or their task-level equivalents, over a declared window. **Compliant work** is output that passes the quality floor and meets every term of the objective. Output that fails either is not discounted work; it is zero work that was paid for. Retried requests count their full spend and only their final compliant output.

A.3 The unit

The cost of useful work over a window is

$$C = \frac{S}{W}$$

where S is total spend attributable to the workload in the window, including failed and retried requests, and W is compliant work delivered.

W is denominated in compliant tasks. A compliant task is one request-with-defined-acceptable-output that passed its quality floor and met every term of its service-level objective. Units: dollars per compliant task. This is the primary form and the only form comparable across model artifacts, because the number of output tokens a model emits to complete a task is chosen by the seller.

The served token is the sub-case. Where the model artifact is held fixed — same weights, same version, same quantization, therefore same tokenizer — output token counts become a stable measure of delivered quantity, and W may be denominated in millions of compliant output tokens. One served token (SVT) is one output token delivered inside its service-level objective and above its quality floor. Units: dollars per million served tokens ($/MSVT). This is the right form for the question the routing layer actually asks, which is where to place a fixed artifact, and it is the question of the worked example in A.9.

The commensurability rule. Token-denominated cost is defined within one artifact class and undefined across artifact classes. Any comparison spanning artifacts must be denominated in compliant tasks. A ranking of dollars per million served tokens across different models is not a conservative approximation; it is the error this specification exists to correct, committed inside the specification. The rule also closes an incentive defect: under a token denominator a model that pads its answer lowers its own apparent cost, while under a task denominator padding raises S and leaves W unchanged. The reference implementation raises an exception rather than returning a cross-artifact token ranking.

The definition's force is in what it refuses. Raw \$/Mtok is C computed with the objective and the quality floor deleted. \$/GPU-hour is C computed with W deleted entirely. Benchmark scores are W with S deleted. Goodput is C with S and the quality floor deleted. Each existing metric is this unit with a leg amputated, which is why none of them can price the industry.

A.4 The conditioning tuple

A value of C is non-conforming and undefined unless stated with all six of:

1. **Workload**: the trace or trace-class measured (arrival pattern, input and output length distributions).
2. **Model artifact**: weights, version, quantization, and tokenizer, explicitly. This field is also the artifact-class key that governs A.3's commensurability rule.
3. **Placement**: the serving configuration measured — provider or self-host, silicon, region, batching policy, and offered load.
4. **Service-level objective**: the percentile bounds, stated numerically.
5. **Quality floor**: the acceptance test, stated operationally, with the evaluated fraction.
6. **Window**: start, end, and sustained load, with figures reported at declared percentiles. The window must be long relative to the system's own response time and stationary within itself. A conforming record reports window length divided by measured p99 request latency, and a split-half comparison of C over the first and second halves. Where the halves differ by more than ten percent, the window is non-stationary, C describes a transient rather than a placement, and the record is marked accordingly. This rule exists because the second design goal is reproducibility by an independent observer, and a number computed over a burst is not reproducible by anyone, including the party that computed it.

The tuple is the lesson of the demand meter, of interchange, and of the Full Disclosure Report: the kilowatt-hour prices nothing without voltage, phase, and time-of-day; a cleared transaction means nothing without the operating rules that define cleared; and price-per-tpmC means nothing without the priced configuration.

A.5 Pre-registration rule

The workload scope and objective are declared before measurement, and results are reported for the declared scope and for the aggregate, both. Scope declared after the fact is selection, not measurement. This rule exists because it is the one every party will be tempted to break, and because the difference between a scoped number and an aggregate number, honestly co-reported, is itself diagnostic information about where a placement fails.

A.6 Provenance taxonomy

Every figure in a conforming record carries exactly one label: [MEASURED], observed on the declared workload in the declared window; [CONFIG], read from configuration rather than observed; [SPEC], taken from a datasheet or published price list; [SIM], produced by a model of the system; or [EST], judgment.

Two rules complete the taxonomy. *Inheritance*: a derived figure carries the weakest label among its inputs. *The red line*: a comparison between placements may be labeled measured only if every side of it is measured; a measured number divided by a simulated one is a simulation, and presenting it otherwise is the accounting fog of Section 8 reproduced at the level of a single line item.

The taxonomy is a coarse, auditable simplification of the framework in JCGM 100:2008, the Guide to the Expression of Uncertainty in Measurement, which remains the current metrological standard [85]. That guide distinguishes Type A evaluation, from statistical analysis of observations, from Type B, from any other means including datasheets and judgment; the labels here are a five-level ordinal version of the same distinction, chosen because a billing record is not a laboratory and an ordinal label is auditable by a non-metrologist. Inheritance and the red line are the coarse form of the guide's propagation rules.

A.7 Conforming record

A conforming measurement is publishable as one row: the six-field tuple, S, W, C, the provenance label of each, and the measuring party. Three further fields are required.

An interval on C, derived from the Wilson score interval on the compliance rate and propagated through the reciprocal — the Wilson form rather than the normal approximation, because compliance legitimately sits near zero and near one, exactly where the normal interval returns bounds outside the unit interval [86]. A record reporting C as a scalar is non-conforming.

A sensitivity curve, sweeping the objective across a declared range and reporting compliance and C at each point, for the reason given in Section 8.3.

A margin to the objective, the ratio of measured percentile to bound, for time-to-first-token and time-per-output-token. A placement compliant at 0.99 of its bound and one compliant at 0.40 report the same C and are not the same asset.

Anything less is marketing.

A.8 Governance

The specification succeeds only if no one owns it. The historical instruction is uniform: interchange worked because Hock's consortium prevented any member from owning the rules; TCP/IP beat better-funded proprietary stacks because it had no owner to distrust; TPC's price-performance metrics are trusted across four decades of adversarial vendor competition because the pricing rules and the audit are the consortium's and not the submitter's; and MCP's donation to a neutral foundation followed the same logic. This draft is offered accordingly, for adoption, amendment, or replacement by any multi-party body that preserves A.3 through A.7. The two bodies best positioned are named in Prediction 2.

A.9 Worked example, entirely from public data

All figures in this section: [SPEC] for prices, [MEASURED, THIRD PARTY] for latency and throughput, from public benchmarks of API providers serving one open-weight artifact, accessed July 2026 [38]. Those figures are trailing-72-hour medians; a fully conforming record requires the declared percentile, typically p99, so this example is illustrative of the method at p50, and its own residual, the missing tail, is part of what it demonstrates.

The same open-weight model is served by 15 providers. The public price spread is 8.6x: \$0.12 per million blended tokens at the cheapest, serving an fp8-quantized artifact, to \$1.05 at the highest. The public speed spread is 21.7x: 329.6 output tokens per second at the fastest against 15.2 at the cheapest. Time-to-first-token at the fastest-responding providers is 0.65 to 0.95 seconds. Because the artifact is held fixed across all fifteen, the token denominator of A.3 is defined here; this is not incidental, it is the condition that makes the example legitimate.

Buyer 1: interactive voice agent. Quality floor: the fp8 artifact passes the buyer's task acceptance, declared per A.2. Objective: TTFT $\leq$ 1.0s and sustained output $\geq$ 100 tokens per second, the floor below which generated speech falls behind playback. Under raw \$/Mtok, the \$0.12 provider wins by more than 5x over the fastest provider's roughly \$0.64 blended. Under the objective, the \$0.12 provider's 15.2 tokens per second delivers approximately zero compliant tokens: $W \approx 0$, and C diverges. The cheapest token in the market is, for this buyer, the most expensive placement available at any price. The fastest provider, at 329.6 tokens per second and 0.95s TTFT, is the cheapest compliant placement at $C \approx \$0.64$ per million served tokens.

Buyer 2: overnight batch summarization. Same model, same prompts. Objective: complete within 12 hours, no per-token latency floor. Now every provider is compliant, W equals billed output everywhere, and C reduces to raw price: the \$0.12 provider wins by 5.3x, and Buyer 1's premium placement is paying 21.7x the speed for a deadline that cannot use it.

Same weights, same input, opposite rankings, and the inversion is invisible to both metrics the industry currently prices with: \$/GPU-hour cannot see the workload at all, and raw \$/Mtok ranks the placements identically for both buyers, wrongly for one of them. The spread between those two correct answers, a factor of five on identical work, is the routing margin of Section 9, computed from nothing but public numbers. One further detail the tuple forces into the open: the cheapest artifact is fp8-quantized, so the two buyers are not entitled to assume the same quality floor until it is declared and tested, which under A.4 they must state, and under current market convention they are never asked to.

A.10 What this specification excludes, knowingly

Multi-tenant attribution of S is no longer excluded. Two bases are defensible and they bracket the honest answer. Under *reserved share*, spend is attributed in proportion to reserved capacity, which charges idle reserved capacity to the tenant that reserved it; this is the correct basis under a reservation and the conservative one. Under *token share*, spend is attributed in proportion to realized compliant output, which ignores idle capacity entirely and is a strict lower bound. A conforming record reports C on both bases and names which it uses. The gap between them is not noise to be eliminated. It is the tenant's own load factor restated in dollars, and it is diagnostic in exactly the way Section 4's diversity factor was.

Genuinely open problems, stated so their omission is not mistaken for resolution: quality floors richer than binary acceptance, since graded usefulness resists standardization and the specification chooses auditability over resolution; the price basis, list, committed, or spot, which the tuple discloses but does not normalize; an energy term, since joules per compliant token is measurable today and is the natural second axis, deferred to keep the unit's first version one-dimensional and because ISO/IEC 21031:2024 already supplies its form; and reference service-level classes, the analogue of standard voltages, which only a multi-party body can legitimately set. Each is a reason this appendix is a draft. None is a reason the unit cannot be computed today, because A.9 just computed it.

References

[1] Financial Times compilation of Q1 2026 hyperscaler earnings, as reported in "Big Tech's AI spending plans reach $725 billion," *Tom's Hardware*, April 30, 2026.

[2] Goldman Sachs Global Investment Research, hyperscaler capex outlook, June 2026 ($5.3T FY2025–FY2030).

[3] J. Furman, public commentary, September 27, 2025 (4% of GDP; 92% of H1 2025 GDP growth). [EST]: social-media source, retained because the decomposition is disputed in text.

[4] Stanford Institute for Human-Centered AI, *Artificial Intelligence Index Report 2025*.

[5] SemiAnalysis GPU rental pricing index and H100 rental pricing coverage, 2024–2026. [EST]: article bodies paywalled.

[6] "Nvidia's $589 Billion DeepSeek Plunge Is Largest in Market History," Bloomberg, January 27, 2025.

[7] C. Perez, *Technological Revolutions and Financial Capital*, Edward Elgar, 2002.

[8] F. Wilson, "The Carlota Perez Framework," AVC, February 2015.

[9] W. B. Arthur, "Increasing Returns and the New World of Business," *Harvard Business Review* 74(4), 1996, pp. 100–109.

[10] E. Brynjolfsson, D. Rock, C. Syverson, "The Productivity J-Curve," NBER WP 25148 (2018); *AEJ: Macroeconomics* 13(1), 2021.

[11] A. Odlyzko, "Collective hallucinations and inefficient markets: The British Railway Mania of the 1840s," 2010.

[12] F. McDonald, *Insull*, Univ. of Chicago Press, 1962; T. P. Hughes, *Networks of Power*, Johns Hopkins, 1983, pp. 208, 217; H. L. Platt, *The Electric City*, Univ. of Chicago Press, 1991; J. L. Neufeld, "Price Discrimination and the Adoption of the Electricity Demand Charge," *Journal of Economic History* 47(3), 1987, pp. 693–709.

[13] J. Nocera, *A Piece of the Action*, 1994.

[14] D. Hock, *One from Many: VISA and the Rise of Chaordic Organization*, 2005.

[15] Visa Inc., Form 10-K, fiscal year 2024, SEC EDGAR.

[16] M. Levinson, *The Box*, Princeton Univ. Press, 2006.

[17] Fabricated Knowledge, "Lessons from History: The Rise and Fall of the Telecom Bubble." Ranges contested.

[18] NASDAQ Composite closing high 5,048.62 on March 10, 2000; closing low 1,139.90 on October 4, 2002; contemporaneous reporting on WorldCom and Global Crossing.

[19] "Dark Fiber Is Lighting Up," *Communications of the ACM*. Utilization estimates are contested ranges.

[20] Alphabet Inc., Form 10-K FY2023 (YouTube ads $31,510M); Alphabet Q4 2025 earnings release (YouTube revenue above $60B for 2025).

[21] M. Burry, public commentary, November 10, 2025; CNBC coverage, November 11, 2025. [EST]: disputed, not independently confirmed by the reporting outlet.

[22] Lawrence Berkeley National Laboratory, "Queued Up" interconnection queue reports.

[23] Epoch AI, "Servers account for 60% of the total cost of ownership of a one-gigawatt AI data center," May 14, 2026; corroborating split in McKinsey, "The cost of compute," April 2025.

[24] DeepSeek-V3 technical report and DeepSeek-R1 API pricing against OpenAI o1 list pricing.

[25] SemiAnalysis, DeepSeek total-cost-of-ownership analysis, January 2025. [EST].

[26] S. Nadella, public commentary, January 27, 2025.

[27] W. S. Jevons, *The Coal Question*, 1865.

[28] D. Bosa, "The AI race is shifting from bigger models to cheaper, smarter systems," CNBC, July 10, 2026.

[29] D. Bosa, "Model routing is a fix for AI overspending," CNBC, June 5, 2026; N. Arora on CNBC, July 9, 2026.

[30] Anthropic, Model Context Protocol announcement, November 2024; donated to a Linux Foundation directed fund, December 9, 2025.

[31] Google Developers Blog, Agent2Agent announcement, April 2025; donated to the Linux Foundation, June 2025.

[32] Google Cloud, Agent Payments Protocol announcement, September 2025; transferred to the FIDO Alliance, 2026.

[33] OpenAI and Stripe, Agentic Commerce Protocol, September 2025.

[34] Google and Shopify, Universal Commerce Protocol, January 2026.

[35] A. Karp, CNBC *Squawk Box*, July 1, 2026.

[36] Nvidia–OpenAI letter of intent (September 2025); OpenAI commitments to Oracle, AMD, and AWS, per contemporaneous reporting.

[37] Bain & Company, 6th annual Global Technology Report, September 23, 2025.

[38] Artificial Analysis, Llama 3.3 70B API provider benchmarks, accessed July 12, 2026. Trailing-72-hour medians; blended price at 7:2:1 cache:input:output.

[39] US railroad revenue per ton-mile, *Historical Statistics of the United States*.

[40] W. B. Norton / DrPeering International, "Internet Transit Prices, Historical and Projections." Survey data.

[41] G. Baker, public commentary, July 12, 2026. [EST]: social-media source.

[42] Indeed Hiring Lab software-engineer job-posting series, 2024–2026.

[43] Etched inference-ASIC claims, June 2026. Vendor claims, not independently verified. [EST].

[44] S. Nadella, remarks at a live taping of *Hard Fork*, June 2026.

[45] E. Tedeschi, public commentary, September 28, 2025. [EST]: social-media source.

[46] SemiAnalysis, "Nvidia GPU Debt Backstop Unleashes the AI Project Trinity," July 6, 2026. [EST]: body paywalled.

[47] SemiAnalysis, "TokenBudgeting: Our Conversations with Enterprises on Token Spend," June 30, 2026. [EST]: body paywalled.

[48] SemiAnalysis, "US Grid Constraints: Towards 40GW+ of Behind-The-Meter Datacenter by 2028," June 25, 2026. [EST]: body paywalled.

[49] SemiAnalysis, "Anthropic 3Q26 Profit Over $1B," July 8, 2026. Analyst estimate, not a filing. [EST].

[50] Bank for International Settlements, *Annual Economic Report 2026*, Chapter I, June 28, 2026.

[51] I. Aldasoro, S. Doerr, D. Rees, "Financing the AI boom: from cash flows to debt," BIS Bulletin No. 120, January 7, 2026.

[52] FinOps Foundation, *FinOps Open Cost and Usage Specification (FOCUS)*, v1.2 (ratified May 29, 2025; virtual-currency and token columns), v1.3 (December 2025), v1.4 (ratified June 4, 2026). A Linux Foundation project. `https://focus.finops.org`

[53] FinOps Foundation, *State of FinOps 2026*; Linux Foundation press release, February 19, 2026.

[54] OpenTelemetry, *Semantic Conventions for Generative AI*, SemConv v1.40.0, April 2026. Status: Development. `https://opentelemetry.io`

[55] MLCommons, *MLPerf Inference*, v5.0–v5.1, 2025. Server-scenario p99 TTFT/TPOT constraints at 99–99.9% reference-accuracy floors. `https://mlcommons.org`

[56] ARC Prize leaderboard methodology; K. Erol et al., "Cost-of-Pass: An Economic Framework for Evaluating Language Models," arXiv:2504.13359, 2025.

[57] Azure AI Foundry provisioned throughput; Google Cloud Vertex AI generative AI scale units; OpenAI service tiers. Each single-vendor and non-portable.

[58] OpenRouter documentation on cross-model token accounting; 2026 analysis of a tokenizer change measuring 12–27% cost movement on identical prompts.

[59] International Electrical Congresses, Paris 1881 through Chicago 1893.

[60] "Metrics and evaluations for computational and sustainable AI efficiency," arXiv:2510.17885, October 2025.

[61] OpenRouter, "State of AI" report (with a16z), arXiv:2601.10088; OpenRouter blog, June 30, 2026. Known biases: router traffic only, excludes first-party volume.

[62] S. Pichai, Google I/O keynote, May 20, 2026 (3.2 quadrillion tokens per month across Google surfaces).

[63] Property and equipment notes, most recent Form 10-K filings, SEC EDGAR: Amazon, Alphabet, Microsoft, Meta, Oracle.

[64] M. Garman (AWS), Cisco AI Summit, February 2026; A. Vahdat (Google), a16z Runtime, October 2025; Silicon Data H100 secondary-market analysis, 2024–2025. No cohort-level utilization series exists as of mid-2026.

[65] MIT Media Lab, Project NANDA, "The GenAI Divide: State of AI in Business 2025," July 2025. Authors describe the finding as a directional snapshot; criticized as methodologically preliminary.

[66] U.S. Census Bureau, Business Trends and Outlook Survey, 2025–2026; Ramp AI Index; Yotzov et al. (2026) executive survey.

[67] Ramp Economics Lab and Revelio Labs, "A New Look at AI's Impact on Jobs," June 30, 2026 (21,559 US firms). Correlation, not causation, per authors.

[68] Combined Microsoft, Alphabet, Amazon, and Meta capital expenditure 2022–2024, from company 10-Ks and the Belfer Center report "AI, Data Centers, and the U.S. Electric Grid," February 6, 2026.

[69] A. Shehabi et al., *2024 United States Data Center Energy Usage Report*, LBNL-2001637, December 2024; IEA, *Energy and AI*, April 2025.

[70] Google technical report on the environmental impact of Gemini inference, August 2025; EU AI Act general-purpose model documentation obligations, Annex XI.

[71] TrendForce, global foundry revenue rankings, FY2025 and 4Q25 (TSMC 69.9% of the global pure-play foundry market for FY2025 on revenue of $122.54B; Samsung 7.2%). Analyst estimate, not an audited filing. Counterpoint's broader "Foundry 2.0" definition yields approximately 39% for the same firm.

[72] Transaction Processing Performance Council, *TPC Benchmark C (TPC-C), Standard Specification*, Revision 5.11, February 2010. Metrics tpmC and price-per-tpmC; 90% of interactive transaction types within 5 seconds and Stock-Level within 20; ACID required (Clause 3); Full Disclosure Report required (Clause 8); independent audit by a TPC-certified auditor required (Clause 9). `https://www.tpc.org`

[73] Transaction Processing Performance Council, *TPC Pricing Standard Specification*, Revision 1.1.0. Establishes the three-year total cost of ownership basis including three years of maintenance pricing.

[74] Transaction Processing Performance Council, *TPC Benchmark H (TPC-H), Standard Specification*, Revision 3.0.1, April 28, 2022. Composite metric QphH@Size; price-performance $/kQphH@Size.

[75] Transaction Processing Performance Council, *TPC Express AI (TPCx-AI), Standard Specification*, Revision 2.0.0. Metrics AIUCpm@SF and $/AIUCpm@SF; no per-request tail-latency constraint.

[76] Y. Zhong, S. Liu, J. Chen, J. Hu, Y. Zhu, X. Liu, X. Jin, H. Zhang, "DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving," *OSDI 2024*, pp. 193–210. arXiv:2401.09670.

[77] A. Agrawal, N. Kedia, A. Panwar, J. Mohan, N. Kwatra, B. S. Gulavani, A. Tumanov, R. Ramjee, "Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve," *OSDI 2024*, pp. 117–134. arXiv:2403.02310.

[78] P. Patel, E. Choukse, C. Zhang, A. Shah, I. Goiri, S. Maleki, R. Bianchini, "Splitwise: Efficient generative LLM inference using phase splitting," *ISCA 2024*, pp. 118–132. arXiv:2311.18677.

[79] "Goodput" originates in networking as useful throughput net of overhead, and entered machine-learning systems through deep-learning cluster scheduling before reaching inference; see e.g. "Sia: Heterogeneity-aware, goodput-optimized ML-cluster scheduling," *SOSP 2023*.

[80] ISO/IEC 21031:2024, *Information technology — Software Carbon Intensity (SCI) specification*, Edition 1, March 2024 (Green Software Foundation).

[81] Standard Performance Evaluation Corporation, *SPECpower_ssj2008*. Metric: overall ssj_ops/watt across ten target load levels in 10% increments plus active idle. `https://www.spec.org`

[82] L. Kleinrock, *Queueing Systems, Volume 1: Theory*, Wiley-Interscience, 1975.

[83] J. Farrell and G. Saloner, "Standardization, Compatibility, and Innovation," *RAND Journal of Economics* 16(1), Spring 1985, pp. 70–83.

[84] P. A. David and S. Greenstein, "The Economics of Compatibility Standards: An Introduction to Recent Research," *Economics of Innovation and New Technology* 1(1–2), 1990, pp. 3–41.

[85] JCGM 100:2008, *Evaluation of measurement data — Guide to the Expression of Uncertainty in Measurement (GUM 1995 with minor corrections)*, Joint Committee for Guides in Metrology.

[86] E. B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association* 22(158), June 1927, pp. 209–212.

[87] Microsoft Azure, *AzurePublicDataset*: AzureLLMInferenceDataset2023, released under CC-BY. Two production Azure OpenAI inference services (conversation and code), collected November 2023; fields are arrival timestamp, context tokens, and generated tokens. `https://github.com/Azure/AzurePublicDataset`

[88] R. Qin, Z. Li, W. He, M. Zhang, Y. Wu, W. Zheng, X. Xu, "Mooncake: A KVCache-centric Disaggregated Architecture for LLM Serving," arXiv:2407.00079v4, 2025. Production traces (23,608 long-context and 12,031 conversation requests, with timestamp, input length, output length, and prefix-block hashes) released at `https://github.com/kvcache-ai/Mooncake`.

[89] S. Krishnamurthy, "The State of Open-Source LLM Inference," July 15, 2026. `https://shwethakrishnamurthy.substack.com/p/the-state-of-open-source-inference` [MEASURED, PRACTITIONER]: single-GPU benchmark of one open model on rented hardware, methodology and funding disclosed by the author; cited for the reported knee behaviour and for the observability gap, not for generalisation.

All forward figures (2026 capex, announced commitments) are company guidance or reporting, not measurements. Predictions 1 through 8 are the falsification surface of this paper; if they fail, so does it.