

The Asymmetry Engine, Chapter 10: The mass of intelligence

Martin Roark

Chapter 10 of *The Asymmetry Engine: How AI Closes the Gaps, and What Stays Scarce*, October 2026. The online edition, with the interactive figures and the links between chapters, is at <https://future-seems-so-good.com/blog/the-asymmetry-engine#gigawatts>. License: CC BY 4.0.

^{10:1} In 2026 the frontier labs plan, finance and are constrained in gigawatts, so the gigawatt is the natural unit of intelligence supply, the I of (0.2). A chain of physical factors turns a gigawatt into agents working at once. The chain runs through bytes of memory more than through arithmetic, and the counts are large: a few million long-context frontier agents per utility gigawatt on current racks, and about nine million for the two largest labs at the end of 2026.

10.1 The unit of account

^{10:2} The industry counts in gigawatts: labs contract compute in them, suppliers quote content per gigawatt, utilities and turbine makers ration interconnection in them, and analysts forecast gigawatts added a year. OpenAI's chief financial officer gives its history, as 0.2 GW in 2023, 0.6 GW in 2024 and about 1.9 GW in 2025. Dylan Patel, who runs SemiAnalysis, gave the 2026 figures on Dwarkesh Patel's podcast (no relation) on 25 August: "At the beginning of this year, OpenAI started at 2 gigawatts and Anthropic at less than 2. End of this year, they're both above 5." In March he had them both at ten gigawatts by the end of 2027.

^{10:3} That is the "3-4x" growth Patel describes, a doubling about every six months, and the two labs took "about 30% of the compute added this year." For the world he gave "30 this year, 50 next year, 70 in '28. '29 should be on the order of 90-100"; the host summed this to "over 200 gigawatts of world compute by the end of 2028," and Patel agreed. Seventy percent of the new watts go to the United States and under ten percent to China. They are better watts, too: GB300, TPUv7 and Trainium3 are "3-5x more performance per watt than the prior-generation chips," so intelligence supply grows faster than the gigawatts.

^{10:4} Three conventions move these numbers by 20–100%. Patel counts critical IT, the power reaching servers; power drawn from the utility is "20-30% higher," and generation higher again, since the PORTS-Pike campus in Ohio needs "at least 10 GW of new energy generation, which results in 8 IT-GW" (filing). Contracted capacity runs ahead of energized capacity: OpenAI "surpassed" its 10 GW Stargate goal in April 2026 while operating about 2 GW. And chip deals and data-center deals describe the same hardware, so adding them counts a gigawatt twice. As of October 2026, every figure here for the end of the year or later is a forecast. The experiment below counts per critical-IT gigawatt and the measurements per utility gigawatt.

10.2 From watts to accelerators to dollars

^{10:5} A utility gigawatt holds about 3.9×10^5 GB300-class accelerators at 2.55 kW each, all in, or roughly $2.8\text{--}4 \times 10^5$ Rubin accelerators, and anchors from OpenAI, Oracle, NVIDIA and SpaceX agree within about 30% (\$C.1). All-in power, two to three times a chip's rating, includes its share of host processors, network, power delivery and cooling. At 288 GB and 8 TB/s an accelerator, a GB300 gigawatt holds about 110 PB of HBM and streams about 3.1 EB/s of it.

^{10:6} A gigawatt costs "roughly \$50 billion" to build in Patel's estimate and "\$60 billion today" in Jensen Huang's, whose company's own content rises from about \$18B a gigawatt for Hopper to \$25B for Blackwell and \$40B for Vera Rubin (NVIDIA call). Renting one costs a base of "\$10 or \$13 or \$15 million per megawatt" a year, while Anthropic's revenue has "gone as high as \$50 million per megawatt" (Patel). Put on Patel's critical-IT basis, the measured counts below give each concurrent frontier agent roughly \$19,000–37,000 of capital and \$4,000–9,000 a year of rent (C:8).

^{10:7} Counting agents from arithmetic overstates them. A GB300 utility gigawatt does about 6×10^{21} dense FP4 operations a second, enough for tens of billions of tokens a second from a model with 49 billion active parameters, yet measured agentic serving delivers $1.3\text{--}2.1 \times 10^8$ output tokens a second per utility gigawatt (InferenceX). To decode one token, the chip re-reads the agent's whole context, its key–value cache, from memory, and agents carry large contexts: across 393 real Claude Code sessions the median

request held 142,016 input tokens and produced 444 ([AgentX](#)). Long-context decoding is bound by memory bandwidth, memory capacity and the floor of Definition 5.2, so a count from FLOPs runs 10–100 times high.

10.3 The agent equation

10:8 The simplest count multiplies accelerators per gigawatt by tokens a second per accelerator and divides by each agent's speed:

$$N(P) \approx \frac{P}{P_{\text{acc}}} \cdot \frac{X_{\text{acc}}(\nu)}{\nu}$$

Everything hard sits in the throughput $X_{\text{acc}}(\nu)$, which depends on the speed each agent needs and on the context it carries.

10:9

Definition 10.1 (Agent; serving instance). An **agent** is one decode stream that sustains ν output tokens a second while holding a live context of n_{ctx} tokens at b_{kv} bytes of cache each. A **serving instance** is a group of n_{inst} accelerators, each with HBM capacity H_{hbm} , bandwidth BW , FLOP rate $\mathcal{F}_{\text{acc}}$ and all-in power P_{acc} . It holds the weights, N_{tot} parameters at b bytes each, and serves n agents per accelerator at once; each decode step accepts $n_{\text{acc}} \geq 1$ tokens per agent, more than one with multi-token prediction or speculative decoding.

▼ The decode step of a serving instance

After the floor t_0 , a step reads the weights the batch touches and every agent's cache, or computes, whichever takes longer:

$$t_{\text{step}}(n) = t_0 + \max \left(\frac{s_{\text{moe}} N_{\text{tot}} b / n_{\text{inst}} + n n_{\text{ctx}} b_{\text{kv}}}{\text{BW } \eta_{\text{bw}}}, \frac{1.3 n n_{\text{acc}} (1 + n_{\text{in}}) (2N_{\text{act}} + \mathcal{F}_{\text{attn}})}{\mathcal{F}_{\text{acc}} \eta_f} \right)$$

Here $s_{\text{moe}}(n)$, local, is the share of mixture-of-experts weights a batch touches; n_{in} is uncached input tokens per output token; $\mathcal{F}_{\text{attn}}$ is attention FLOPs per token; η_{bw} and η_f are achieved fractions of peak; 1.3 is an allowance for overhead. The floor t_0 is the one Definition 5.2 defines, synchronization plus one read of the active weights, calibrated on measured speeds (§B.8). Latency adds only cache reads to it; here a large batch touches most experts, so the bandwidth term also carries the weights it touches, $s_{\text{moe}} N_{\text{tot}} b$. The active weights are then counted twice, which for DeepSeek V4 Pro costs about 0.06 ms a step on a GB300 rack and 0.5 ms on an 8-GPU B200 server, so the count errs slightly low. A batch is feasible when each agent keeps its speed, $n_{\text{acc}}/t_{\text{step}} \geq \nu$, and everything fits, $N_{\text{tot}} b / n_{\text{inst}} + n n_{\text{ctx}} b_{\text{kv}} \leq 0.92 H_{\text{hbm}}$.

10:10 Solving both conditions for n , choosing the instance size n_{inst} that maximizes it, and multiplying by accelerators per gigawatt and by serving utilization η_{srv} gives the count, with $t_{\text{left}} = n_{\text{acc}}/\nu - t_0$ the time left in each step after the floor and $t_W = s_{\text{moe}} N_{\text{tot}} b / (n_{\text{inst}} \text{BW } \eta_{\text{bw}})$ the part of it spent streaming weights:

10:11

$$N(P) = P \frac{\eta_{\text{srv}}}{P_{\text{acc}}} \min \left\{ \underbrace{\frac{0.92 H_{\text{hbm}} - N_{\text{tot}} b / n_{\text{inst}}}{n_{\text{ctx}} b_{\text{kv}}}}_{\text{capacity}}, \underbrace{\frac{\text{BW } \eta_{\text{bw}} (t_{\text{left}} - t_W)}{n_{\text{ctx}} b_{\text{kv}}}}_{\text{bandwidth}}, \underbrace{\frac{\mathcal{F}_{\text{acc}} \eta_f t_{\text{left}}}{1.3 n_{\text{acc}} (1 + n_{\text{in}}) (2N_{\text{act}} + \mathcal{F}_{\text{attn}})}}_{\text{compute}} \right\} \quad (10.1)$$

10:12 Speed decides between capacity and bandwidth, and context length between them and compute. Cache bytes divide both memory terms, so a slower agent fills memory first: capacity binds below about 87 tok/s on Rubin and 37 on GB300 at the reference values of Figure 10.1, and compute only for contexts of a few thousand tokens. Both hyperbolas of (5.1) sit inside: with one agent and no floor the bandwidth term caps speed times active bytes, and t_{left} is the second hyperbola, so as ν approaches n_{acc}/t_0 the count falls to zero however many gigawatts there are.

10:13 Most of a gigawatt is spent re-reading memories. An agent at 100 tok/s with a 100,000-token context drags a 10 GB cache that its rack re-reads about 53 times a second, so on GB300 bandwidth alone would admit about 7.4 million such agents per critical-IT gigawatt, and the floor, the weights, achieved efficiency and utilization cut that to about 2.3 million (§C.1).

10:14

Proposition 10.1 (What moves the count). In the bandwidth-bound regime of (10.1), let $s_W = t_W/t_{\text{left}}$ be the share of each step's budget spent streaming weights. The count is proportional to power and to serving utilization, and inversely proportional to power per accelerator and to cache bytes per agent, $n_{\text{ctx}}b_{\text{kv}}$. Its elasticity is $1/(1 - s_W)$ in achieved bandwidth $\text{BW } \eta_{\text{bw}}$ and $-s_W/(1 - s_W)$ in weight bytes; it is $(n_{\text{acc}}/\nu)/(t_{\text{left}} - t_W)$ in tokens accepted per step, the negative of that in speed, and $-t_0/(t_{\text{left}} - t_W)$ in the floor.

▼ *Proof*

The bandwidth term is $N = P(\eta_{\text{srv}}/P_{\text{acc}})(\text{BW } \eta_{\text{bw}} t_{\text{left}} - s_{\text{moe}} N_{\text{tot}} b/n_{\text{inst}})/(n_{\text{ctx}} b_{\text{kv}})$ with $t_{\text{left}} = n_{\text{acc}}/\nu - t_0$. Differentiate its logarithm term by term, holding s_{moe} fixed; it tends to 1 at large batch. ■

10:15 Two elasticities are structural, +1 in gigawatts and -1 in cache bytes per agent; the others grow as the step budget fills, with model size dominating when bandwidth is scarce and speed near the floor. Each hardware generation raises bandwidth and lowers floors, pushing s_W toward zero and leaving cache bytes as the main source of variation. In the experiment's model, weights carry the largest share of the variance in the count on 8-GPU B200 servers and cache bytes about two-thirds of it on Rubin, where the fitted elasticities in context length and bytes per token, -0.92 and -0.91, approach -1.

10:16

Experiment 10.1 (Agents per gigawatt).

Setup: (10.1) solved for 20,000 draws of frontier models per configuration on 8-GPU B200 servers, GB300 NVL72 and Rubin NVL72 racks.

Parameters: 1.5–10 trillion parameters, 1.5–5% active; 40,000–160,000 tokens of context at 40–250 KB a token; 100 tok/s per agent; 55–80% serving utilization.

Result: medians of 0.37, 1.31 and 4.9 million agents per critical-IT gigawatt, or 0.29, 1.05 and 3.9 million per utility gigawatt; at the measured configuration the roofline gives 1.6–6.1 times the measured streams per GPU; the two largest labs' inference shares run a median 8.8 million agents at the end of 2026 and 34 million at the end of 2027.

Shows: in the model, the counts are an upper envelope that agrees with measurement only after offsets, so measured rows win where they exist. Method and tables: §B.8, Figure B.5.

10:17

Interactive in the online edition: <https://future-seems-so-good.com/blog/the-asymmetry-engine#fig:fermi>

Figure 10.1 (interactive). The chain from one utility gigawatt to accelerators, to tokens a second under the tightest of three limits, to agents. Lower the speed ν and watch the binding limit pass from bandwidth to capacity; lengthen the context n_{ctx} and watch the count fall in proportion; then raise ν toward a few hundred tokens a second and watch the count collapse into the floor as the gigawatts needed for ten million agents climb.

10.4 What a gigawatt buys

10:18 Measurements on real agentic-coding traffic set the count. SemiAnalysis served DeepSeek V4 Pro, 1.6 trillion parameters with 49 billion active, on replayed Claude Code sessions at 100 tok/s per agent. Per utility gigawatt that comes to 0.3–0.5 million agents on 8-GPU B200 servers, 1.3–2.1 million on GB300 NVL72 and 2.8–4.4 million on Vera Rubin NVL72, against tens of millions of short-context chat streams.

10:19 The intuitive picture is a million agents with one B200 each. The count is accurate for the hardware it names: 8-GPU B200 servers run 0.75–1.2 long-context agents per GPU, and Kimi K3 at 50 tok/s on the same servers gives 0.6–1.0 (InferenceX). A million B200-class accelerators draw 2–3 GW, OpenAI's fleet at the start of 2026. The mechanism is batching. A 1.6-trillion-parameter model needs about 0.8 TB of FP4 weights and a B200 holds 180 GB, so serving instances span 8–72 accelerators and interleave many agents' decode steps: each chip serves fragments of many agents, and each agent runs on fragments of many chips. For the 2026 fleet one agent per accelerator is a floor, since GB300 racks run 3.4–5.4 per GPU and Rubin more.

10:20 At the end of 2026, OpenAI's roughly 6 GW, about 40% of it serving inference, supports about 3–5 million concurrent frontier agents at GB300 efficiency, and 8–13 million if the whole fleet served agents. The two labs' inference shares run about 6–9 million on the measured rows. The experiment's median, an upper envelope, is 8.8 million at the end of 2026 and 34 million at the end of 2027, with a 96% chance of at least ten million by then (§C.2). The world's 2026 additions alone, 30 GW or about 12 million accelerators, would run 40–130 million long-context agents, each available 168 hours a week (§8.3). The measured counts multiply critical-IT gigawatts by per-utility rates, so run about a fifth low (C:14).

10:21 Four omissions push these counts up. They cover inference only, and in Patel's split "50% of the compute is research, 10% ... development, and then 40% is inference" (episode). An agent waiting on a tool holds no decode slot, so live sessions outnumber streams: one B300 server held about 48 resident sessions per GPU (InferenceX). Lighter work fits more streams, and at 50 tok/s the model's GB300 count quadruples.

10:22

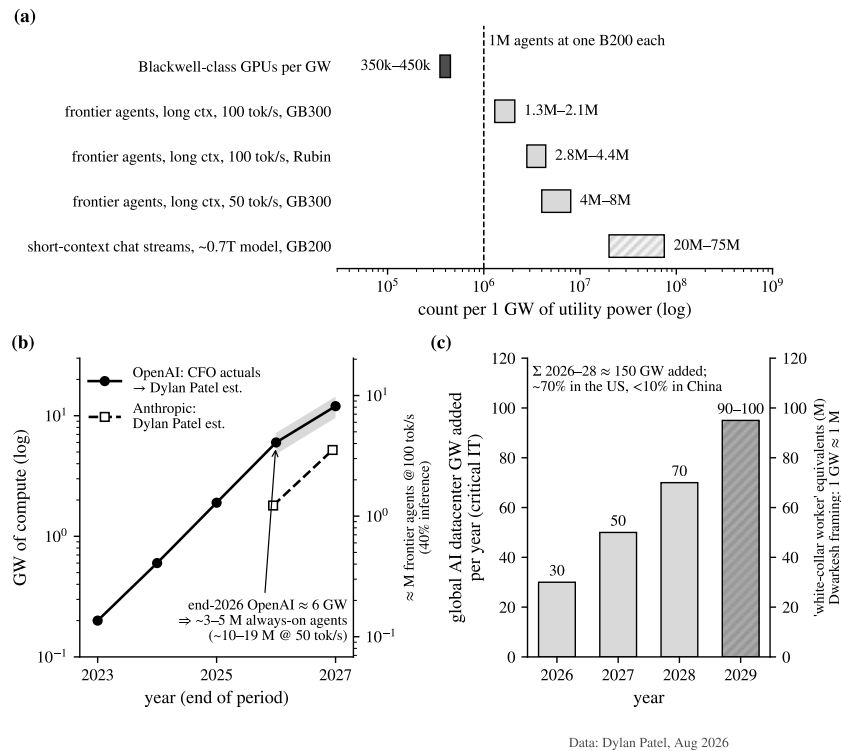


Figure 10.2. What one utility gigawatt buys by workload, from accelerators to long-context frontier agents at 100 and 50 tok/s and short-context DeepSeek-R1 chat streams measured on GB200 at 43–91 tok/s (a); OpenAI's and Anthropic's compute in gigawatts, actual and forecast, with the agents a 40% inference share would run (b); and the world's yearly additions of AI data-center power in critical IT (c). A utility gigawatt runs 1.3–4.4 million long-context frontier agents or 20–75 million chat streams, and each year adds tens of gigawatts; Patel's forecasts in (b) are critical IT, so its agent axis runs up to a fifth low.

10.5 Four ways to see the numbers

10:23 The human brain runs on about 20 W, so a gigawatt is the power of fifty million brains. Most of those watts are not computation: cortical computation takes about 0.1 W of them (6:2), and an energy audit puts communication at 3.5 W (Levy and Calvert). The brain is a fifth of the body's metabolic load, so the body runs its brain at a power usage effectiveness near 5, against a data center's 1.2–1.3. In the experiment's model a median GB300 agent draws about 760 W of IT power, thirty-eight brains, and a Rubin agent about 200 W, ten brains.

10:24 At the experiment's medians a GB300 IT gigawatt emits about 1.3×10^8 output tokens a second, 7.6 J per token counting idle capacity; Rubin spends 2.0 J and 8-GPU B200 servers 27 J. People read aloud at about 183 words a minute (Brybaert), about four tokens a second at three-quarters of a word per token, so a brain that spoke without pause would spend about 4.9 J per token. Per

emitted token, silicon and flesh are within a factor of two. The asymmetry is duty cycle: people speak about 16,000 words a day ([Mehl et al.](#)), roughly ninety minutes, so per token actually spoken the brain spends about 80 J, ten times GB300's cost. The two labs' median fleet at the end of 2027 would emit about 1.1×10^{17} tokens a year, against about 6×10^{16} for everything [8.2 billion people](#) say, if Mehl's student sample generalizes, which is generous. At [Shannon's](#) bit per letter of English, an agent at 100 tok/s emits about 400 bits a second around the clock, forty times what Zheng and Meister measure as "the information throughput of a human being" ([Neuron](#)).

10:25 Patel's host offered the same order as an intuition pump: "a gigawatt that can sustain a population of, say, roughly a million white-collar workers ... That would be \$100 billion," or \$100,000 a head ([episode](#)). The measured count on GB300 is 1.3–2.1 million agents per utility gigawatt. The money is half the pump's: Anthropic's best revenue, \$50 million per megawatt-year, is \$50B per gigawatt-year, against a base rent of \$10–15B.

10:26 About 8.2 billion people ([UN](#)) at 20 W draw about 165 GW of brain power. [Epoch AI](#) put AI data centers at about 31 GW at the end of 2025, "comparable to peak power usage in New York State." Patel's additions take the world [past 200 GW](#) by the end of 2028, so during 2028 its AI data centers would draw more power than every human brain combined, counting critical IT alone. The crossing is one of mass. Power is a poor proxy for thought, since estimates of the brain's computation span four orders of magnitude, 10^{13} – 10^{17} FLOP/s ([Carlsmith](#)), and \$18.3 sets the crossing against the sunlight Earth intercepts.

10:27 I give the first crossing even odds, since Patel's 150 GW of additions over 2026–28, on an installed base of 25–31 GW, clear 165 GW of critical IT by a tenth or less. I give the second about three in ten, because the growth in model size that Forecast 10.4 expects spends much of each hardware gain.

10:28

Forecast 10.1 (More watts than brains). By the end of 2028 the world's AI data centers have more critical-IT capacity than all human brains draw combined, about 165–170 GW.

Horizon 2029-06-30

Probability 50%

Check Read the first consolidated tally of global AI critical-IT capacity at the end of 2028 published by the horizon, such as Epoch AI's data-center database or SemiAnalysis's models; a figure below 165 GW falsifies the claim.

10:29

Forecast 10.2 (The joule crossing). By the end of 2029, on hardware shipping in 2028 or 2029, a long-context frontier agent at 100 tok/s uses 1 J or less of critical-IT energy per output token at full load, against about 4–6 J measured on GB300.

Horizon 2029-12-31

Probability 30%

Check Read independent agentic benchmarks of 2028–29 hardware, such as SemiAnalysis's InferenceX, at 100 tok/s per agent and about 100,000 tokens of context on the leading open-weight frontier model, and convert their throughput per megawatt into joules of critical-IT energy per output token, a utility megawatt being 0.8 MW of critical IT. The forecast holds at 1 J or less and fails above it.

10.6 Ceilings

10:30 The three terms of (10.1) take turns. In the experiment's model at 100 tok/s, bandwidth binds in 97–98% of draws on B200 and GB300 and in 65% on Rubin. Halving the speed quadruples GB300's count, as Proposition 10.1 predicts when the floor and the weights fill most of each step, but raises Rubin's by only 40%, because at 50 tok/s Rubin's memory fills and capacity binds in 97% of its draws. Either way the cache is the lever. Latent, compressed and sparse attention and quantized caches cut bytes per token; offloading a cache while its agent waits on a tool raises live agents above concurrent streams; multi-token prediction and speculative decoding raise the tokens accepted per step. Architecture matters to first order: on the same GB300 racks, Kimi K3, with 2.8 trillion parameters and 104 billion active, runs about half as many agents as DeepSeek V4 Pro even at half the speed (§C.1).

10:31

Forecast 10.3 (The cache wall). By the end of 2027 cache bytes per agent are the main determinant of agents per gigawatt on Rubin-class racks: at 100 tok/s, quadrupling each agent's context cuts the agents a megawatt serves by at least a third.

Horizon 2027-12-31

Probability 75%

Check Read the agentic benchmarks of Rubin-class racks published in 2027, such as SemiAnalysis's AgentX, that report output throughput per megawatt at 100 tok/s per agent for 25,000- and 100,000-token contexts. The forecast holds if the 25,000-token figure is at least 1.5 times the 100,000-token figure; it fails if the ratio is lower or if no such pair is published.

10:32 Bytes also trade against model size. The experiment's fleet projection lets frontier models grow from 1.5–6 trillion parameters in 2026 to 3–20 trillion in 2028. Agents per inference gigawatt then roughly double from 2026 to 2027 as Rubin arrives and stay flat into 2028, because larger models absorb Rubin's bandwidth. This is the Jevons effect inside a gigawatt: cheaper bytes are spent on bigger minds, so the fleet's count grows about as fast as its gigawatts and its quality as fast as its bandwidth.

10:33

Forecast 10.4 (Minds grow to fill the bandwidth). Through 2028, labs spend most of each bandwidth doubling on larger or longer-thinking models, so concurrent frontier agents per gigawatt at a fixed speed grow far more slowly than bandwidth per gigawatt, and the fleet's count tracks its gigawatts.

Horizon 2028-12-31

Probability 65%

Check Read independent agentic benchmarks, such as SemiAnalysis's InferenceX, that report concurrent agents per megawatt at 100 tok/s for each year's leading open-weight frontier model on the newest rack-scale hardware at the end of 2026, 2027 and 2028. The forecast fails if that count grew faster than $2\times$ a year in both 2027 and 2028, or if no such benchmarks are published; it holds otherwise.

10:34 Before lithography, equipment binds. Gas turbines and generator step-up transformers take "three to four years, versus a historical norm of roughly 18 months," and behind-the-meter generation will power "well over half of new US datacenters in 2028+" (SemiAnalysis). Patel still expects US power to scale, with "over 16 different manufacturers of power-generating things just from gas alone"; at Davos Elon Musk expected "more chips than we can turn on."

10:35 Lithography binds next. In Patel's arithmetic, "three and a half EUV tools satisfies a gigawatt," so about \$1.2B of tools holds up about \$50B of data center, and ASML's output, about 70 tools a year rising to "a little bit over 100 by the end of the decade," puts about 700 in service by 2030: a ceiling near 200 GW of AI chips a year if every tool served AI (\$C.4). Sam Altman's goal of "a gigawatt of new AI infrastructure every week," 52 GW a year, would take a quarter of it, and a Taiwan disruption would cut world additions to "maybe 10 gigawatts across Intel and Samsung, or 20." Past about 2028 the mass of intelligence grows at the rate of lithography.

10:36 In (0.2), intelligence supply is gigawatts times efficiency, and the last line says that closing the research gap raises efficiency, so a software gain multiplies every gigawatt at once: "If you make an improvement in AI software, it has the potential to be immediately applied to all of the GPUs that you already have" (Carl Shulman). This chapter's gigawatts enter (11.2) twice. The inference fleet sets the automated research labor L , and the research share sets the experiment compute C , whose growth the capex cycle sets. A swarm draws its N from the same count (§9.4), and §11.3 asks how fast the loop can accelerate this mass.

References

1. gives its history. <https://openai.com/index/a-business-that-scales-with-the-value-of-intelligence/>
2. Dwarkesh Patel's podcast. <https://www.dwarkesh.com/p/dylan-patel-3>
3. March. <https://www.dwarkesh.com/p/dylan-patel>
4. filing. <https://www.sec.gov/Archives/edgar/data/1045810/000104581026000069/sbeoainvidia-portsrelease.htm>
5. "surpassed". <https://openai.com/index/building-the-compute-infrastructure-for-the-intelligence-age/>

6. OpenAI. <https://openai.com/index/jalapeno-first-results/>
7. 288 GB and 8 TB/s an accelerator. <https://www.nvidia.com/en-us/data-center/gb300-nvl72/>
8. NVIDIA call. <https://www.theglobeandmail.com/investing/markets/stocks/NVDA/pressreleases/4355994/nvidia-nvda-q2-2027-earnings-call-transcript/>
9. InferenceX. <https://inference.semianalysis.com/blog/vera-rubin-nvl72-agentic-inference>
10. AgentX. <https://inference.semianalysis.com/agentx/methodology>
11. InferenceX. <https://inference.semianalysis.com/rankings/fastest-gpu-for-kimi-k3>
12. 180 GB. <https://www.nvidia.com/en-us/data-center/hgx/>
13. InferenceX. <https://inference.semianalysis.com/blog/agentx-inferencev3-does-cuda-moat>
14. William B. Levy, Victoria G. Calvert (2021). Communication consumes 35 times more energy than computation in the human cortex, but both costs are needed to predict synapse number. *Proceedings of the National Academy of Sciences*. <https://www.pnas.org/doi/10.1073/pnas.2008173118>
15. Marc Brysbaert (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*. <https://doi.org/10.1016/j.jml.2019.104047>
16. three-quarters of a word per token. <https://help.openai.com/en/articles/4936856>
17. Matthias R. Mehl, Simine Vazire, Nairán Ramírez-Esparza, Richard Bennett Slatcher, James W. Pennebaker (2007). Are Women Really More Talkative Than Men? *Science*. <https://www.science.org/doi/10.1126/science.1139940>
18. 8.2 billion people. <https://population.un.org/wpp/>
19. Claude E. Shannon (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*. <https://doi.org/10.1002/j.1538-7305.1951.tb01366.x>
20. Neuron. [https://www.cell.com/neuron/fulltext/S0896-6273\(24\)00808-0](https://www.cell.com/neuron/fulltext/S0896-6273(24)00808-0)
21. Epoch AI. <https://epoch.ai/data-insights/ai-datacenter-power>
22. Carlsmith. <https://coefficientgiving.org/research/how-much-computational-power-does-it-take-to-match-the-human-brain/>
23. SemiAnalysis. <https://newsletter.semianalysis.com/p/us-grid-constraints-towards-40gw>
24. Davos. <https://www.weforum.org/podcasts/meet-the-leader/episodes/conversation-with-elon-musk-davos-2026/>
25. goal. <https://blog.samaltman.com/abundant-intelligence>
26. Carl Shulman. <https://www.dwarkesh.com/p/carl-shulman>