

The Asymmetry Engine, Appendix C: Fermi tables

Martin Roark

Appendix C 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#fermi>.

License: [CC BY 4.0](#).

C:1 Every count in the book that starts from watts or dollars is derived here from a few inputs, so a reader can change one input and redo the arithmetic. The formal objects live in their chapters: the count of agents is (10.1), the physical share is (15.1), and the rule that only the binding constraint earns rent is (16.1).

C:2 Two conventions settle most disagreements between sources. Lab and world gigawatts are critical IT, the power that reaches servers, and power drawn from the utility is 20–30% higher ([Patel](#)). Measured throughputs are per utility gigawatt, while the book's experiment, Experiment 10.1, counts per critical-IT gigawatt. Unless a row says otherwise, an agent is a long-context frontier stream at 100 tok/s, as in Definition 10.1.

C.1 Agents per gigawatt

C:3 The count is accelerators per gigawatt, times output tokens a second per accelerator, divided by each agent's speed: the product form of (10.1). The first factor comes from public anchors that agree within about 30%. Each anchor's all-in power is its power divided by its accelerators, and it includes each accelerator's share of host processors, network, power delivery and, at the utility, cooling.

C:4

Anchor	Power	Accelerators	All-in power each
OpenAI, Stargate	more than 5 GW	"over 2 million chips"	at most 2.5 kW
Oracle, Q1 FY27	850 MW	"more than 300,000 GPUs"	at most 2.8 kW
NVIDIA, Rubin	100 MW	40,000, with power oversubscription	2.5 kW
SpaceX S-1, Colossus II	220 MW of IT	110,000 GB300	2.0 kW of IT
NVIDIA, PORTS-Pike	4.25 GW of IT	about 1.5 million a generation	2.8 kW of IT
Adopted, GB300 class (OpenAI)	1 utility GW	3.9×10^5	2.55 kW
Adopted, Rubin, between the two NVIDIA anchors	1 utility GW	$2.8\text{--}4 \times 10^5$	2.5–3.6 kW

C:5 The second factor is the measurement of 10:18, reported as total tokens a second per utility megawatt, cached input included ([InferenceX](#)). The traces' token counts put 135–214 tokens in all behind each output token ([InferenceX](#)). Dividing the measured throughput by that ratio gives output tokens, dividing again by 100 tok/s gives agents, and dividing by accelerators per gigawatt gives agents per accelerator.

C:6

Hardware	Total tok/s per utility MW	Output tok/s per utility GW	Agents per utility GW	Agents per accelerator
B200, 8-GPU servers	6.95 million	32–51 million	0.3–0.5 million	0.75–1.2
GB300 NVL72	28.5 million	133–211 million	1.3–2.1 million	3.4–5.4
Rubin NVL72	59.4 million	278–440 million	2.8–4.4 million	7–16

C:7 Two other workloads bracket these rows. A bigger model, Kimi K3 with 2.8 trillion parameters and 104 billion active, runs 16,762 tokens a second in all per GB300 accelerator at 50 tok/s per agent (InferenceX), which by the same ratio is 0.6–1.0 million agents per utility gigawatt at half the speed. Short-context chat, DeepSeek-R1 with a thousand tokens in and a thousand out on GB200 NVL72, ran about 51 streams per GPU at 91 tok/s and 179 at 43 tok/s (InferenceX). GB200 draws about 1.9 kW of IT power (S-1), about 2.4 kW at the utility, or 4.2×10^5 accelerators a utility gigawatt, so that is about 20–75 million streams per utility gigawatt.

C:8 Money follows from the same rows once both sit on one basis. Build cost, \$50–60B a gigawatt (Patel; Huang), and base rent, \$10–15 million a megawatt-year (Patel), are taken as critical IT, like Patel's gigawatts, and a critical-IT gigawatt of GB300 runs 1.6–2.6 million agents, the measured 1.3–2.1 million per utility gigawatt times 1.25. Each agent therefore carries about \$19,000–37,000 of capital and \$3,800–9,200 a year of rent.

C:9 Most of a gigawatt is spent re-reading memories, and a pass through the bandwidth term shows how much. One critical-IT gigawatt of GB300 is about 490,000 accelerators at about 2.05 kW of IT power each (S-1), each with 288 GB of HBM read at 8 TB/s (NVIDIA): 140 PB streaming 3.9 EB/s. An agent at 100 tok/s with a 100,000-token context at 100 KB a token holds a 10 GB cache. At 1.9 tokens accepted per step, its rack re-reads that cache about 53 times a second, 0.53 TB/s. The efficiency, floor and utilization are the experiment's reference values (§B.8).

C:10

Step	Factor	Agents per critical-IT GW
Bandwidth over cache traffic	$3.9 \text{ EB/s} \div 0.53 \text{ TB/s}$	7.4 million
Achieved fraction of peak bandwidth, η_{bw}	$\times 0.72$	5.3 million
Time left after a 5 ms floor in a 19 ms step	$\times 0.74$	3.9 million
Weights of a 1.6-trillion-parameter FP4 model, 11 GB per accelerator	$\times 0.86$	3.4 million
Serving utilization, η_{srv}	$\times 0.68$	2.3 million

C:11 Per utility gigawatt, at 2.55 kW all-in, the same pass gives about 1.8 million, the default reading of Figure 10.1. The experiment's reference configuration, the same model at 70 KB a token and 1.8 tokens a step, gives 2.8 million per critical-IT gigawatt, and its median over 1.5–10 trillion-parameter models gives 1.31 million. Capacity binds less at this speed, since 140 PB holds about 14 million such caches. The roofline is optimistic: run on the configuration that was measured, it gives 1.6–6.1 times the measured streams per accelerator. Its medians land near the measured rows only because larger models and utilization offset that optimism.

C:12

In the model	B200, 8-GPU	GB300 NVL72	Rubin NVL72
Median agents per critical-IT GW, 100 tok/s	0.37 million	1.31 million	4.9 million

In the model	B200, 8-GPU	GB300 NVL72	Rubin NVL72
The same per utility GW, $\div 1.25$	0.29 million	1.05 million	3.9 million
Median at 50 tok/s	2.3 million	5.3 million	6.8 million
Median streams per accelerator, 100 tok/s	1.06	3.9	19.3
Streams per accelerator, measured configuration	4.59	8.53	27.9

C:13 Speed is the input to change with most care. On one chip, cost and energy per token rise steeply as an agent's speed nears its ceiling (Proposition 5.1), so a count rescaled from 100 tok/s to a faster target by $1/\nu$ alone overstates the agents a gigawatt can run, and the power it implies for a given count is a floor.

C.2 Fleets

C:14 A fleet's count is its gigawatts times the rows above. Patel's lab figures are critical IT and the measured rows are per utility gigawatt, so multiplying the two, as these rows do, understates each count by roughly a fifth.

C:15

Fleet	Gigawatts	Long-context agents	Arithmetic
OpenAI, start of 2026	about 2	about 1 million	0.7–1 million B200-class accelerators at 0.75–1.2 each
OpenAI, end of 2026	about 6, 40% for inference	3–5 million; 8–13 million if all served agents	2.4 or 6 GW $\times$ 1.3–2.1 million
OpenAI and Anthropic, end of 2026	about 11, 40% for inference	about 6–9 million	4.4 GW $\times$ 1.3–2.1 million
The same, in the model	10–12.5	8.8 million median, 2.1–27.6 million	inference share 35–60%, GB300
OpenAI and Anthropic, end of 2027, in the model	18–26	34 million median, 14–75 million	half GB300, half Rubin
OpenAI and Anthropic, end of 2028, in the model	35–70	82 million median, 15–236 million	Rubin, models of 3–20 trillion parameters
World additions, 2026	30	39–132 million if all served agents	30 $\times$ 1.3–4.4 million; about 12 million accelerators

C:16 The gigawatts come from Patel: OpenAI at about 2 GW at the start of 2026 and both labs "above 5" at its end (August), OpenAI at "six plus" and both labs at ten by the end of 2027 (March), and inference at "40%" of the labs' compute. In the model, the chance that the two labs run at least ten million agents is 45% at the end of 2026 and 96% at the end of 2027.

C:17 World additions depend on whose model is used. Patel's is bound by the supply chain, and Dario Amodei's extrapolates demand.

C:18

World AI power added, GW	2026	2027	2028	2029
<u>Patel</u> , August 2026	30	50	70	90–100

World AI power added, GW	2026	2027	2028	2029
<u>Amodei</u> , February 2026	10–15	30–40	about 100	about 300

C:19 Patel's additions grow 1.4–1.7 times a year and Amodei's about three times. Sam Altman's goal for one company, "a gigawatt of new AI infrastructure every week," is 52 GW a year. Epoch AI put the installed base at about 31 GW all-in at the end of 2025, roughly 25 GW of critical IT. Patel's 150 GW of additions over 2026–28 then bring critical IT to about 175 GW at the end of 2028, against about 165 GW for 8.2 billion brains at 20 W. That margin of 6%, or 10% if all 31 GW count as critical IT, is the one behind Forecast 10.1.

C.3 Ten million researchers

C:20 Ten million automated researchers, each running at 10,000 tokens a second around the clock, the picture that 11:55 puts below 1%, is $10^7 \times 10^4 = 10^{11}$ output tokens a second. Three methods turn that into power, and they differ in what they assume about context and utilization.

C:21

Method	Arithmetic	Power
Output tokens only, at 50% of peak	a 10-trillion-parameter model with 400 billion active needs 8×10^{11} FLOP a token, so 8×10^{22} FLOP/s; at half of a <u>GB300</u> 's 15 PFLOP/s, 1.1×10^7 accelerators at 2.5 kW	25–30 GW
The same at production decode utilization	6.9% of peak instead of 50%	about 190 GW
Long-context agentic rates	10^{11} tok/s is 10^9 agents at 100 tok/s, $\div 1.3$ –4.4 million per utility GW	230–770 utility GW

C:22 The first method counts arithmetic for output tokens alone, the floor for short contexts. Its 50% is generous. DeepSeek's published production decode ran about 1,850 output tokens a second per H800 on a 37-billion-active model (DeepSeek); at $2 \times 37 \times 10^9$ FLOP a token that is 1.4×10^{14} FLOP/s, 6.9% of the 1,979 TFLOP/s of FP8 the H800 shares with the H100. The third method uses the measured agentic rates, in which 135–214 tokens in all stand behind each output token, so prefill and attention dominate. Automated researchers rewriting training code are long-context agents, so the third figure is the realistic one, and it is still a floor, because a stream at 10,000 tok/s forfeits the batching those rates assume.

C:23 The agentic figure runs from all of the world's expected AI compute at the end of 2028, over 200 GW of critical IT or over 250 GW at the utility, to three times it. In the model it is about 30 times the decode output of the two labs' median fleet at the end of 2027. If each of the two labs ran ten million, every figure doubles. At \$10–15B a gigawatt-year of rent, the 180–620 critical-IT gigawatts it needs cost about \$2–9T a year.

C:24 The buildable versions share an aggregate of 10^9 tokens a second. Ten million agents at 100 tok/s need $10^7 \div 1.3$ –4.4 million per utility gigawatt, 2.3–7.7 utility GW, one lab's fleet in 2026–27. A hundred thousand at 10,000 tok/s need single streams faster than any GPU runs, which means SRAM, wafer-scale or hard-wired designs (§5.4). §11.6 argues what each version would buy.

C.4 Macro arithmetic

C:25 The money, the build-out and the speed limits come down to thirteen lines of arithmetic, each from public numbers. The last column names where each result is used.

C:26

Quantity	Arithmetic	Result	Used in
\$10T as a share of 2027 world output	$\$10T \div \$131.9T$ (<u>IMF</u>)	7.6%	16:3
Capex in gigawatts	\$1T, \$4T and $\$10T \div \$50\text{--}60B$ a gigawatt	17–20, 67–80 and 170–200 GW	16:5
\$10T against one year's build	$170\text{--}200\text{ GW} \div \text{the } 50\text{ GW expected in } 2027$	3.3–4×	16:5
Growth of the physical base	$(120/30)^{1/3} - 1$, from 30 GW in 2025 to 120 GW in 2028 (<u>Morgan Stanley</u>)	about 59% a year	§11.3
Software against watts	$\ln 13 \approx 2.56$ for the price of a fixed capability, falling $13\times$ a year (<u>Epoch AI</u>), against $\ln 1.6 \approx 0.47$	about 5.5×	§11.3
Gas-turbine bookings	GE Vernova 53 GW firm and 63 GW reserved (<u>GE Vernova</u>); Siemens Energy 69 and 26 (<u>Siemens Energy</u>)	122 GW firm, 211 GW in all; GE Vernova alone holds 2.7 years of its 20 GW a year firm, 5.8 in all	§15.3
EUV tools a gigawatt of Rubin	about 55,000 wafers of 3 nm, 6,000 of 5 nm and 170,000 of DRAM, about two million exposures (<u>Patel</u>)	about 3.5 tools	§10.6
The lithography ceiling	about 700 tools by 2030 $\div 3.5$ tools a gigawatt	about 200 GW a year	§10.6
Anthropic's multiple	$\$2T$ of IPO talk (<u>Mint</u>) $\div$ more than $\$65B$ of July run-rate (<u>Reuters</u>)	about 31×	§16.6
GDPval's review share	$\$86$ to review $\div \$361$ to produce; $109 \div 404$ minutes (<u>GDPval</u>)	0.24 of the cost, a ceiling of 4.2× on savings; 0.27 of the time, a ceiling of 3.7× on speed	(15.1)
Amdahl ceilings	$1/s_{\text{phys}}$ at 0.5, 0.2, 0.1 and 0.01	2×, 5×, 10×, 100×	(15.1)
Cognition ten times faster	$1/(s_{\text{phys}} + (1 - s_{\text{phys}})/k_{\text{cog}})$ at $k_{\text{cog}} = 10$ and the same shares	1.8×, 3.6×, 5.3×, 9.2×	(15.1)
A capped loop's share of its cap	$1 - g_C/g_J$ at returns to research $r = 1$, with $g_J = 12 \ln 2 \approx 8.3$ a year, a one-month first doubling, and $g_C = \ln 3$, experiment compute growing $3\times$ a year	about 0.87	Proposition 11.2

C:27 Three rows need a note. Turbine reservations are slots held by down payments and firm backlog is contracted orders, so the firm figure is the one to set against SemiAnalysis's three-to-four-year lead times for turbines (SemiAnalysis). The $13\times$ a year is Epoch's measured cost per question at fixed capability, and 1.6 is the build-out's 59% a year. The capped loop is the reduced research loop $\dot{S} = g_J S^q$ at $r = 1$ under a ceiling that grows with experiment compute at g_C , the case of clause (iv) of Proposition 11.2 that Figure 11.2 draws at these rates: its share x of the ceiling obeys $\dot{x} = g_J x(1 - x) - g_C x$, whose stable point is $x = 1 - g_C/g_J$ (x local).

C:28 The toy chain of (16.1) has five inputs in equal proportions, one unit of each per unit of output, sold at a price of 1. Its inputs are the book's illustration, chosen so that cognition starts scarce.

C:29

Input	Unit cost	Capacity	With cognition 100× cheaper
Cognition	0.30	50	cost 0.003, capacity 5,000
Memory	0.10	80	unchanged
Power	0.10	90	unchanged
Fabs	0.10	120	unchanged
Hands	0.10	100	unchanged

C:30 Output is the smallest capacity, 50, and the unit margin is $1 - 0.70 = 0.30$, so the surplus is 15.0 and all of it goes to cognition, the binding constraint. With cognition a hundred times cheaper and abundant, the unit margin is $1 - 0.403 = 0.597$, memory binds at 80, and the surplus more than triples to 47.8, all of it to memory.

References

1. March. <https://www.dwarkesh.com/p/dylan-patel>
2. OpenAI, Stargate. <https://openai.com/index/stargate-advances-with-partnership-with-oracle/>
3. Oracle, Q1 FY27. <https://stockanalysis.com/stocks/orcl/transcripts/689425-q1-2027/>
4. NVIDIA. <https://www.nvidia.com/en-us/data-center/vera-rubin-nv172/>
5. SpaceX S-1. <https://www.sec.gov/Archives/edgar/data/1181412/000162828026036936/spaceexplorationtechnologi.htm>
6. NVIDIA, PORTS-Pike. <https://investor.nvidia.com/news/press-release-details/2026/NVIDIA-Guarantees-SB-Energys-PORTS-Pike-Technology-Campus-in-Ohio-to-Exclusively-Host-NVIDIA-AI-Compute/default.aspx>
7. OpenAI. <https://openai.com/index/jalapeno-first-results/>
8. InferenceX. <https://inference.semianalysis.com/blog/vera-rubin-nv172-agentic-inference>
9. InferenceX. <https://inference.semianalysis.com/blog/agentx-inferencexv3-does-cuda-moat>
10. InferenceX. <https://inference.semianalysis.com/rankings/fastest-gpu-for-kimi-k3>
11. InferenceX. <https://inference.semianalysis.com/blog/gb200-nv172-vs-b200-disagg-deepseek-r1-fp4-dynamo-trt>
12. NVIDIA call. <https://www.theglobeandmail.com/investing/markets/stocks/NVDA/pressreleases/4355994/nvidia-nvda-q2-2027-earnings-call-transcript/>
13. Dwarkesh Patel's podcast. <https://www.dwarkesh.com/p/dylan-patel-3>
14. 288 GB and 8 TB/s an accelerator. <https://www.nvidia.com/en-us/data-center/gb300-nv172/>
15. Amodei. <https://www.dwarkesh.com/p/dario-amodei-2>
16. goal. <https://blog.samaltman.com/abundant-intelligence>
17. Epoch AI. <https://epoch.ai/data-insights/ai-datacenter-power>
18. DeepSeek 2025. https://github.com/deepseek-ai/open-infra-index/blob/main/202502OpenSourceWeek/day_6_one_more_thing_deepseekV3R1_inference_system_overview.md
19. H100. <https://www.nvidia.com/en-us/data-center/h100/>
20. IMF projects. <https://www.imf.org/-/media/files/publications/weo/2026/april/english/tablea.pdf>
21. Morgan Stanley. <https://www.listennotes.com/podcasts/thoughts-on-the/can-the-ai-spending-boom-pay-9FwTmYCIEd1/>
22. Epoch. <https://epoch.ai/publications/the-plunging-price-of-thought>
23. GE Vernova. https://www.governova.com/sites/default/files/gev_webcast_pressrelease_07222026.pdf

24. Siemens Energy. https://assets.siemens-energy.com/dam/3e846440-66dd-4a56-be75-b49d004e6741/2026-08-05_Q3_Analyst_presentation-pdf_Original%20file.pdf
25. Mint. <https://www.livemint.com/companies/news/anthropic-ipo-may-come-in-november-why-its-2-trillion-valuation-is-raising-eyebrows-11790902031997.html>
26. Reuters. <https://www.reuters.com/technology/anthropic-revenue-run-rate-tops-65-billion-source-says-2026-08-17/>
27. Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang et al. (2025). GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. *arXiv*. <https://arxiv.org/html/2510.04374v1>
28. SemiAnalysis. <https://newsletter.semianalysis.com/p/us-grid-constraints-towards-40gw>