

The Asymmetry Engine, Chapter 16: Capital and the binding constraint

Martin Roark

Chapter 16 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#capital>. License: [CC BY 4.0](#).

16:1 About a trillion dollars of AI capital spending in 2026 is real. Four trillion a year is a 2030 figure found only at the top of the forecasts, and ten trillion a year is in no mainstream forecast. The money buys gigawatts, and a gigawatt is made of turbines, transformers, wafers and memory, each with its own build time. Down the stack from electrons to agents, rent pools where supply is least elastic: memory, packaging, lithography and power equipment earn the widest margins, while the capital-heavy clouds in the middle carry the debt. Under fixed proportions only a binding constraint earns rent, so when cognition becomes abundant the surplus passes to the owner of the next one. Rents summon capacity with a lag, and the stiffest layers boom and bust hardest.

16.1 The capex path

16:2 The trillion for 2026 holds once its scope is stated. As of October 2026, the big four (Amazon, Alphabet, Meta, Microsoft) guide \$720–750B of company-wide capex for calendar 2026, about 76% above the \$416B they spent in 2025 and 14% above their own January guides ([Platformonomics](#); [TMT Finance](#)); Microsoft's often-quoted \$255–260B is its fiscal 2027, against about \$175B for calendar 2026 ([lowdown](#)). Counting neoclouds and the supply chain, total AI capex is about \$870B for [Morgan Stanley](#) and "a little bit over a trillion dollars" for [Dylan Patel](#), and [Dell'Oro](#) expects global data-center capex to "approach \$1 Trillion in 2026", "almost 1% of the gross world product" for [Epoch](#). Analysts' estimates have run low, "too conservative during each of the past three years by an average of 45 pp" in Goldman Sachs's count ([note](#)).

16:3 For 2027 the forecasts run from \$1.2T for the five largest hyperscalers ([Goldman Sachs](#), above a \$1.1T consensus) to "more than \$1.4 trillion ... next year alone" for the build-out ([Morgan Stanley](#)) and about \$2T for the whole supply chain ([Patel](#)). Four trillion a year appears as Patel's 2028 figure in a scenario that reaches 100 GW by 2030, and at the top of the 2030 range: NVIDIA, an interested party, forecasts "\$3 trillion to \$4 trillion" annually ([Yahoo Finance](#)), and [Dell'Oro](#) more than \$3T. Ten trillion a year appears only in the interviewer Dwarakesh Patel's conditional "close to \$10 trillion by the end of 2030", or as a cumulative total, such as [McKinsey's](#) \$5.2T of AI data-center capex to 2030 or SemiAnalysis's \$11T for 2024–29, \$5T of it debt-funded ([Patel](#)). In 2027, \$10T would be about 7.6% of the \$131.9T of world output the [IMF projects](#).

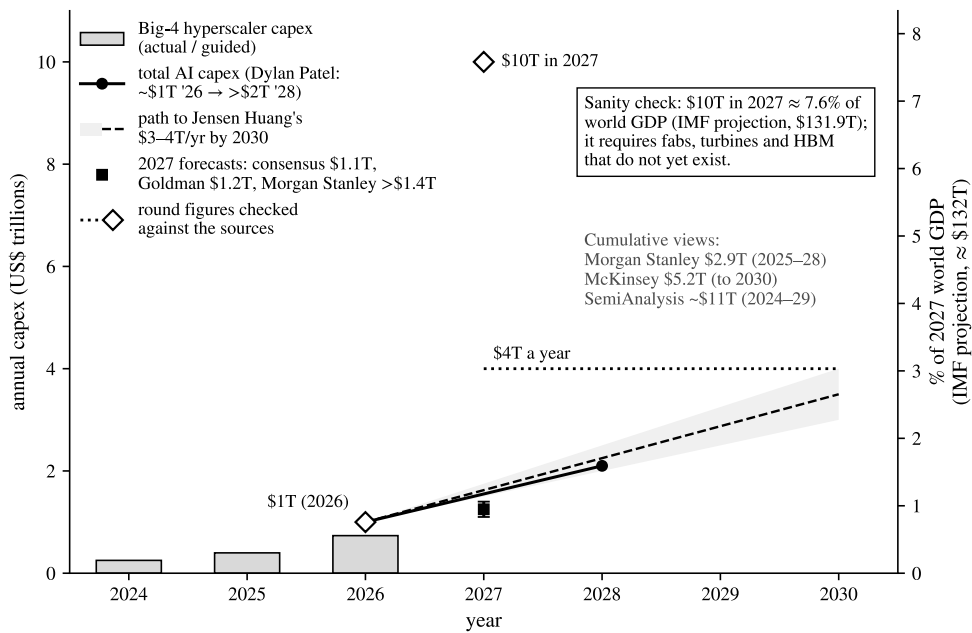


Figure 16.1. Annual AI capital spending, 2024–2030, in trillions of dollars and as a share of projected 2027 world output: big-four capex, with \$720–750B guided for 2026; Patel's whole-chain total; the 2027 forecasts; the path to NVIDIA's \$3–4T a year by 2030; and three often-quoted round figures (\$1T in 2026, \$4T a year, \$10T in 2027) drawn as hollow diamonds and tested against the sourced paths. The \$1T sits on the sourced line and \$4T a year only at the top of the 2030 band, while \$10T in 2027 stands seven to nine times above that year's forecasts, at about 7.6% of world output, and is about the whole build-out of 2024–29.

16:5 Capex is the gigawatt path in another unit. At \$50–60B of all-in capex per gigawatt (\$10.2), \$1T a year buys 17–20 GW, the order of the 30 GW the world adds in 2026; \$4T buys 67–80 GW, close to the 70 GW expected for 2028; and \$10T buys 170–200 GW, three to four times the 50 GW expected for 2027 (Patel). Two hundred gigawatts a year is about all that lithography could serve by 2030 if every extreme-ultraviolet (EUV) tool made AI chips: a gigawatt of Rubin-class chips needs about 3.5 of ASML's tools, so about \$1.2B of tools holds up \$50B of capex, and ASML ships about 70 a year, with about 700 installed by 2030 (Patel). The round figures make a coherent late-decade path and an impossible 2027 one, because the machines that make the machines do not yet exist.

16:6 What binds is what the money buys, because each item has a build time. GE Vernova has 116 GW of gas turbines under contract (53 GW firm, 63 GW of slot reservations), with agreements "into '31", against output of 20 GW a year rising to 30 GW by 2030 (GE Vernova). With Siemens Energy's 69 GW of backlog and 26 GW of reservations, the two hold about 122 GW firm and 211 GW with reservations against 30–40 GW a year of combined output: three to four years of output firm, five to seven with reservations, which carry down payments but are not firm orders. Turbine and step-up-transformer lead times have reached three to four years, against about 18 months historically (SemiAnalysis). PJM's 2027/28 capacity auction cleared at its \$333.44/MW-day cap and 6,623 MW short of its reliability requirement, citing "an unprecedented surge in data center load" (PJM), and new nuclear arrives over 2027–2035, with restarts such as Crane's in 2027 coming first (World Nuclear News).

16:7 In the master equation (0.2), capex is the physical form of $\dot{I}$, the rate at which intelligence supply is built, and so the compute term of the research loop, whose growth raises the odds of a fast takeoff in Experiment 11.1

(§B.7). The turbines, wafers and memory are the homeostat of Proposition 11.2, which caps the loop's long-run speed at the build-out's: the value chain is where the positive loop of research meets the negative loops of matter.

16.2 The stack, from electrons up

16:8 Every layer of the stack is necessary, since a gigawatt without transformers or high-bandwidth memory (HBM) serves nothing, and the layers earn very different returns. At the bottom, GE Vernova holds a \$176.3B backlog "on track to reach \$200 billion in 2027" ([GE Vernova](#)), and Caterpillar a record \$72.1B (+92%), with power-generation dealer sales up 72% on data-center demand ([Caterpillar](#)). Rehlko, the former Kohler Energy, reports about 1.7 GW of hyperscale backup-power awards in 60 days ([Rehlko](#)) and is doubling its Changzhou capacity to 11 GW a year ([Rehlko](#)), but Platinum Equity has owned a majority of it since 2024, so this rent is closed to public investors ([Rehlko](#)).

16:9 The electrical and cooling gear above the generators is less stiff. Vertiv's backlog was \$15.0B at the end of 2025 (+109%), about a year of sales ([Vertiv](#)), with a 22.6% adjusted operating margin ([Vertiv](#)); Eaton's data-center orders rose about 85% ([Eaton](#)), while its widely quoted "307 GW" of US data-center backlog is a project pipeline ([Eaton's call](#)). Backlog over annual output is a usable proxy for the inverse of supply elasticity, about a year for switchgear and cooling and five or more for gas turbines. What is slow to build sets each layer's elasticity, which is how inertia enters the price (Proposition 15.1).

16:10 Lithography, foundry and packaging come next. TSMC raised its 2026 capex to \$60–64B from \$52–56B, guides revenue growth "slightly above 40%" and reported a 67.7% gross margin, and its chief executive said in July that "our packaging capacity is so tight that now it limits my customers' growth" ([TSMC](#)). Analysts expect its CoWoS packaging capacity to grow from 70–80 thousand wafers a month at the end of 2025 to 120–140 thousand a year later and 190–200 thousand by the end of 2027 ([TrendForce](#); [Mizuho](#)).

16:11 Memory is the hottest layer. In the second quarter of 2026, Counterpoint put HBM revenue at 50% for SK hynix, 33% for Samsung and 18% for Micron, against 64%, 15% and 21% a year earlier ([Yonhap](#)). SK hynix earned ₩60.5T of operating profit on ₩79.3T of revenue, a 76% margin ([SK hynix](#)); Samsung earned ₩89.5T, about \$61.5B, nearly all of it from chips ([Samsung](#)); Micron's fiscal-2026 revenue was \$133.2B against \$37.4B a year earlier, and it has contracted "the vast majority" of its 2027 HBM supply "with significant price increases" ([Micron](#)). SanDisk is a NAND company that won through flash, with fiscal-2026 revenue of \$20.25B (+175%) and an 84.6% fourth-quarter gross margin ([SanDisk](#)); with SK hynix it is standardizing High Bandwidth Flash, stacked NAND that complements HBM, with samples due in 2027 ([SanDisk](#)).

16:12 Memory is hot for a physical reason. Over two decades, peak server FLOPS grew about 3.0× every two years and DRAM bandwidth 1.6× ([Gholami et al.](#)), and each decode step streams the active weights and every agent's cache from memory, so speed and agents per gigawatt are bought in bytes per second ((5.1); §10.3). Memory is also the most cyclical layer of semiconductors (§16.4).

16:13 Accelerators earn the margin of a layer that is itself short of supply. NVIDIA's quarter to 26 July brought \$96.2B of revenue (+106%) at a 75.0% gross margin ([NVIDIA](#)), and it expects "extreme pricing conditions in memory" to pull that margin to 71–72% ([NVIDIA's call](#)): part of its rent flows upstream to the memory makers.

16:14 Below the accelerators the economics invert. The big four's trailing free cash flow fell 24% in a year, and Alphabet's was $-\$5.9\text{B}$ in the second quarter (Platformonomics), its first negative quarter since it listed in 2004 (futurex). S&P cut Oracle to BBB– in July (UBS); Oracle's free cash flow was $-\$5.4\text{B}$ in its August quarter, and CoreWeave lost $\$626\text{M}$ on $\$2.6\text{B}$ of revenue in the second quarter (cdelta), with NVIDIA obligated to buy its unsold capacity through 13 April 2032 (CoreWeave). The capital-heavy middle owns no binding constraint and rents everyone else's with borrowed money. Above it sit the frontier labs ($\$16.6$); below the physical stack, the hands that wire it ($\$15.4$).

16:15

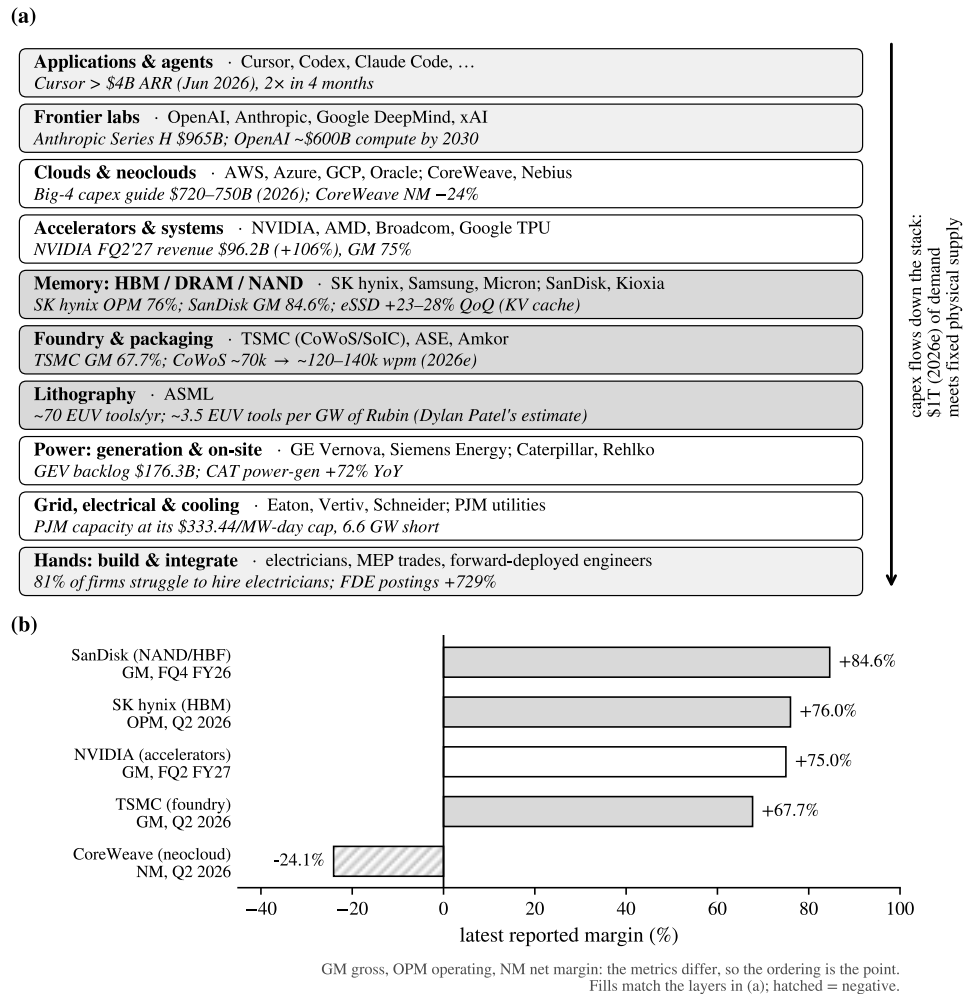


Figure 16.2. (a) The ten layers of the AI stack as of October 2026, from applications and agents down to the hands that build and integrate them, with their firms and a latest reported number for each, such as the big four's $\$720\text{--}750\text{B}$ capex guide for 2026; (b) the latest reported margins of five firms, from SanDisk's gross margin to CoreWeave's net loss. The margins differ in kind, so the ordering is the point: the physical layers that cannot expand quickly earn the most, and the capital-heavy cloud in the middle loses money.

16.3 Where the rent goes

16:16 Cash pools in the layers that cannot expand quickly. Goldratt's theory of constraints starts there, with a system's throughput set by its tightest constraint (Goldratt), and the price system spreads a scarcity without anyone

ordering it (§4.7). In a chain of complements the layer that binds charges for scarcity and the others cannot: Ricardo's rent, which can be made precise.

16:17

Definition 16.1 (A fixed-proportions chain). One unit of deployed AI service, a powered, cooled and networked gigawatt-year, needs n_j units of each layer j . Layer j has nameplate capacity K_j growing at rate g_j , unit cost c_j , price p_j and short-run supply elasticity $\varepsilon_j > 0$: by overtime, expediting and yield pushes it supplies $n_j Y = K_j (p_j/c_j)^{\varepsilon_j}$ for $p_j \geq c_j$. Final output Y grows at rate g_Y . The layer's markup is $m_j = p_j/c_j - 1$, its cost share $s_j = p_j n_j / \sum_k p_k n_k$, and its **scarcity index**

$$\mathcal{S}_j = \frac{g_Y - g_j}{\varepsilon_j}.$$

16:18 Hold the capacities fixed over a horizon and the chain becomes a linear program, with output $Y^* = \min_j K_j/n_j$. A layer is a **binding constraint** when its capacity sets that output, $n_j Y^* = K_j$, and its shadow price p_j^{sh} is what one more unit of it would add to the surplus. With the final good priced at p_Y , duality gives

$$p_j^{\text{sh}} > 0 \Rightarrow n_j Y^* = K_j, \quad \sum_j p_j^{\text{sh}} K_j = \left(p_Y - \sum_j c_j n_j \right) Y^* \quad (16.1)$$

16:19 The first statement is complementary slackness: only a binding constraint earns rent, and inputs in excess supply earn their cost and nothing more. The second is strong duality: the rents add up to the whole surplus. An input can bind over a horizon only if it cannot be built within it, so as the horizon lengthens the rent passes to the slowest-built input. In the toy chain of C:30, cognition, the scarcest of five inputs, takes the entire surplus of 15; made a hundred times cheaper and abundant, it lets output rise until memory binds, and the surplus, more than tripled to 47.8, goes entirely to the owner of the memory. Cheaper cognition enlarges the surplus and passes it to whoever owns the next binding constraint, which is the rule of §2.8, that once generation is free the check becomes scarce, stated for a whole economy.

16:20

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

Figure 16.3 (interactive). Each bar is the output one input allows; the shortest sets output, and only its shadow price is above zero. Make cognition cheaper and more plentiful, and watch the binding constraint and the whole rent jump to memory.

16:21

Proposition 16.1 (Bottleneck rents). With partly elastic supply the rents are graded. In the chain of Definition 16.1: (i) layer j 's markup grows at its scarcity index, $\frac{d}{dt} \ln(1 + m_j) = \mathcal{S}_j$; (ii) after a demand shock at fixed capacities, layer j captures the share

$$\omega_j = \frac{s_j/\varepsilon_j}{\sum_k s_k/\varepsilon_k}$$

of the incremental producer surplus, and what an inelastic layer collects this way is its **bottleneck rent**; (iii) as $\varepsilon_j \rightarrow 0$ for one layer, $\omega_j \rightarrow 1$, so a hard constraint takes the whole increment while the other layers price near cost, which recovers (16.1). Ranking rule: order the layers by $s_j \mathcal{S}_j$, and the top layer captures the marginal dollar of capex. Importance is not the criterion; elasticity is.

▼ *Proof*

Inverting supply gives $p_j = c_j (n_j Y / K_j)^{1/\varepsilon_j}$, whose log-derivative in time is $(g_Y - g_j)/\varepsilon_j$, which is (i) because $1 + m_j = p_j/c_j$. Producer surplus, the area above the supply curve, is $\text{PS}_j = (p_j n_j Y - c_j K_j)/(1 + \varepsilon_j)$ (PS_j local); at fixed capacities $d \ln p_j = d \ln Y / \varepsilon_j$, so $d \text{PS}_j = p_j n_j Y d \ln Y / \varepsilon_j$, and normalizing by $\sum_k d \text{PS}_k$ gives (ii). The j -th term dominates the sum as $\varepsilon_j \rightarrow 0$, which gives (iii). ■

16:22 The 2026 record fits the ranking rule. The stiffest layers by the proxies of §16.2 (turbines booked five or more years ahead, packaging that limits customers' growth, HBM contracted a year ahead) show the widest margins or longest order books, and the most elastic large layer, rentable cloud capacity, the weakest cash flows (Figure 16.2). The rents also have an address: the stiffest layers sit in Korea, Taiwan, the Netherlands and Japan, so the asymmetry points abroad at least as often as to the United States.

16:23 The same algebra predicts a new layer. Once generation is cheap, the scarce input in many workflows is checking: GDPval's deliverables still need about 109 minutes of expert review each ((2.3)), and Catalini, Hui and Wu argue that "the binding constraint on growth is no longer intelligence. It is human verification bandwidth" ([arXiv](#)). Verification capacity, human and machine, then behaves like a layer of the stack with its own rent and its own cycle: the verifier cap of Proposition 9.2 at the scale of the economy.

16:24

Forecast 16.1 (Verification becomes a layer). By the end of 2028, verification capacity (expert review, evaluators, test infrastructure, audit) becomes a priced, low-elasticity layer of the AI stack, and its suppliers report rising margins or order books.

Horizon 2029-03-31

Probability 50%

Check Read the annual results for 2026 and 2028 of listed suppliers of AI evaluation, expert review and data annotation, such as Innodata and Appen, published by the horizon, and the reported figures of private ones, such as Scale AI, Surge AI and Mercor. The forecast holds if at least two report higher gross margins or larger contracted backlogs for 2028 than for 2026; it fails otherwise.

16.4 Rents are self-liquidating

16:25 A markup is also a signal to build, and every layer's capacity answers it after its own build time.

16:26

Proposition 16.2 (Rents are self-liquidating). In the chain of Definition 16.1, let capacity growth answer the markup after a build time Δ_j ,

$$g_j(t) = \bar{g}_j + \nu_{\text{inv},j} m_j(t - \Delta_j),$$

with $\bar{g}_j$ the growth at zero markup and $\nu_{\text{inv},j}$ the strength of the investment response. Near the resting markup $\bar{m}_j$, where $\bar{g}_j + \nu_{\text{inv},j} \bar{m}_j = g_Y$, the deviation $\epsilon(t) = m_j(t) - \bar{m}_j$ obeys the delayed brake of (4.3) with $g = 0$, lag Δ_j and strength

$$\nu_{\text{brk}} = \frac{\nu_{\text{inv},j} (1 + \bar{m}_j)}{\epsilon_j}.$$

By Proposition 4.3 (iv), the markup relaxes smoothly when $\nu_{\text{brk}} \Delta_j \leq 1/e$, overshoots in damped cycles when $1/e < \nu_{\text{brk}} \Delta_j < \pi/2$, and cycles with growing amplitude beyond $\pi/2$ ($\bar{g}_j$ and $\bar{m}_j$ local). A low elasticity raises the strength and a long build time raises the lag, so the stiffest layers earn the largest rents and cycle hardest.

▼ Proof

Put the investment rule into (i) of Proposition 16.1: $\frac{d}{dt} \ln(1 + m_j) = (g_Y - \bar{g}_j - \nu_{\text{inv},j} m_j(t - \Delta_j)) / \epsilon_j$. The right side vanishes at the resting markup; writing $m_j = \bar{m}_j + \epsilon$ and keeping first-order terms gives $\dot{\epsilon}(t) = -\nu_{\text{brk}} \epsilon(t - \Delta_j)$. This is Ezekiel's cobweb theorem (1938) in continuous time. ■

16:27 The proposition is the bear case written in the bull case's variables: the stiffness that creates a rent also makes the layer that earns it the most prone to boom and bust. Memory's history is this proposition: after each earlier peak Micron's gross margin gave up more than 30 points within about a year, and SK hynix's operating profit fell 87% from 2018 to 2019 ([StockTitan](#)). Goldratt's fifth focusing step says it in management language: once a constraint is broken, "go back to step one, but do not allow inertia to cause a system constraint."

16:28 The responses are already moving. Samsung's share of HBM revenue rose from 15% to 33% in a year, and SemiAnalysis models industry DRAM wafer additions rising from about 190,000 a month in 2026 to 395,000 in 2028, with China's CXMT approaching 17% of DRAM supply ([StockTitan](#)); TSMC's capex and GE Vernova's turbine output are rising too (16:10; 16:6). The forward question is whose capacity response arrives first, and how much of it at once. A thesis that buys the rent without pricing the response has bought half an equation.

16:29

Forecast 16.2 (Memory turns first). By the end of 2028, memory is the first layer of the stack whose markup cycles down as new capacity arrives: DRAM contract prices fall for two consecutive quarters before any of the big four cuts its capex guidance.

Horizon 2028-12-31

Probability 50%

Check Read TrendForce's quarterly DRAM contract prices and the capex guidance of Alphabet, Amazon, Meta and Microsoft through 2028. The forecast holds if contract prices fall for two consecutive quarters before any of the four lowers its guidance for 2027 or 2028; it fails if a guidance cut comes first, or if contract prices have not fallen for two consecutive quarters by the horizon.

16.5 Who captures a closed gap

16:30 Seen from the stack, the half cycle of closing (0:13) is a cycle of markups and capacity: a markup summons capacity after its build time, the new capacity erodes the markup, and the rent re-forms at whichever layer binds next (Proposition 16.2; (16.1)). Wiener drew the dynamic conclusion that free competition has no homeostasis (4:48).

16:31

Proposition 16.3 (Who captures a closed gap). A **closer** is anyone who pays to turn a gap into realized value: a trader, a firm, a swarm, a model. Let closers pay $c_{\text{close},k}$ per unit of effort aimed at gap k , and let the return per unit of effort fall as closers enter, because each closure leaves less gap. Free entry stops where that return equals $c_{\text{close},k}$, at a residual gap $G_k^* > 0$ whenever $c_{\text{close},k} > 0$, which Grossman and Stiglitz called "an equilibrium degree of disequilibrium". If the closed gap's surplus is produced through a chain like Definition 16.1, then with free entry: (i) the closers' share of the surplus tends to zero; (ii) the rest divides between buyers, through lower prices, and the owners of the chain's inputs, in proportion to s_j/ϵ_j ; (iii) in a task i that compute can reproduce, the wage is capped by the compute cost of reproducing it,

$$w_i \leq C_i p_C,$$

with C_i the compute per unit of task i and p_C its price. Labor earns a rent only where it is a low-elasticity complement.

▼ Proof

(i) Entry continues while the return per unit of effort exceeds $c_{\text{close},k}$, and competition drives that cost toward the price of the intelligence closers buy; their margin is the temporary extra surplus value of §3.7. (ii) is Proposition 16.1 applied to the closers' inputs. (iii) is arbitrage: if $w_i > C_i p_C$, a firm replaces the work with compute. Restrepo finds the same bound with AGI, where "wages converge to the opportunity cost of computational resources required to reproduce human work" (NBER w34423). That is Marx's accounting in reverse: in *Capital* the reproduction cost of labor-power sets the wage, and here the machine's cost of reproducing the work caps it. ■

16:32 AI lowers $c_{\text{close},k}$ for every gap at once, so it shrinks every residual G_k^* , which is the precise sense in which it is a general-purpose closer, and it makes closings worth doing that were not before: the regeneration term of

(0.2). In the synthetic market of Experiment 3.1, adding three cheaper model tiers cuts inference spend by 89% while surplus rises 19%, so in the model cheaper intelligence moves money from model providers to the firms that deploy them, and competition should pass it on to buyers.

16:33 The cap in (iii) can be priced. On the measured counts an always-on frontier agent carries about \$4,000–9,000 a year of rent, or about \$0.5–1 an hour (10:6). Dworkesh Patel's picture of a gigawatt that sustains about a million white-collar workers puts their wages near \$100B a year, against a base rent of \$10–15B for the gigawatt (10:25). Whatever the exchange rate between agents and workers, that distance is the surplus of (3.1) at the scale of an economy, and the proposition says where it lands: with buyers and the owners of the stiffest inputs, with a lab only while it stays ahead, and with a worker only where she is the binding constraint.

16:34 Of the three destinations of an agent's surplus (3:62), the duality of (16.1) makes the last two, the owners of slow complements and the owners of verification, the same kind of place, since both are binding constraints. Labor, so far, is not among them (§17.1).

16.6 Valuations and base rates

16:35 Valuations test the same algebra at the top of the stack. Anthropic's last priced round was its Series H at \$965B post-money on 28 May 2026, when its run-rate revenue had "crossed \$47 billion" ([Anthropic](#)), up from about \$9B at the end of 2025; by late July it had passed \$65B ([Reuters](#)). Two trillion dollars is IPO talk: Reuters, which saw the draft prospectus, wrote that the listing "could value it at more than \$2 trillion" ([Reuters](#)), and prospective investors talk of \$1.8–2T ([Mint](#)). The prospectus shows 2025 revenue of about \$4.6B, a net loss of about \$42B that includes a \$34B non-cash charge, and about \$518B of compute commitments ([Reuters](#)), some 80% of them non-cancelable ([PitchBook](#)).

16:36 At \$2T Anthropic would be priced at about 31 times its July run-rate and 435 times its 2025 revenue, against about 20 times for OpenAI's reported \$1.4T ask on a run-rate near \$70B ([TechCrunch](#); [Tech Funding News](#)). PitchBook judged that the leak "supports a valuation well above \$1 trillion and falls short of justifying \$2 trillion" ([PitchBook](#)). A trailing multiple goes stale within a quarter for a firm whose run-rate rose about sevenfold in seven months, so the useful question is what must be true in five years for \$2T to be fair.

▼ What \$2T must become

A buyer who pays val_0 breaks even if, after T years, the firm earns a net margin s_{net} on revenue rev_T and trades at a price-to-earnings multiple pe , discounted at g_{disc} a year, so the required revenue and growth from today's run-rate rev_0 are

$$\text{rev}_T^* = \frac{\text{val}_0 (1 + g_{\text{disc}})^T}{\text{pe} \cdot s_{\text{net}}}, \quad g_{\text{req}} = \left(\frac{\text{rev}_T^*}{\text{rev}_0} \right)^{1/T} - 1$$

(all symbols local). With \$2T today, five years and \$65B of run-rate, a bull setting (multiple 30, margin 30%, discount 10%) needs \$358B of revenue in 2031, 41% growth a year; a base setting (25, 25%, 12%) \$564B, 54% a year; a bear setting (15, 15%, 15%) \$1.79T, 94% a year. At the \$30–50B of revenue per gigawatt-year that [Hock Tan](#) of Broadcom and [Patel](#) cite, the base case needs 11–19 GW of paid traffic in 2031, within reach of Anthropic's fleet, which grows from under 2 GW at the start of 2026 to above 5 GW at its end and about 10 GW by the end of 2027 (10:2). The 2030 build-out needs about \$2T a year of AI revenue ([Bain](#)), so the base case

asks one lab for more than a quarter of it. If in five years the firm is either a \$10T franchise or a \$0.3T commodity supplier, \$2T is fair at 12% a year when the franchise has probability $(2 \times 1.12^5 - 0.3)/(10 - 0.3) \approx 0.33$, or about 0.57 for a \$6T franchise.

^{16:37} In the terms of (16.1), frontier capability binds on tasks the cheapest sufficient model cannot do, and the treadmill of §3.8 moves that line every quarter. A lab keeps its rent only by staying ahead, which means buying the capex whose rents go to memory and power. The bull case is that research capability compounds through §11.1, which would make it the one input whose owner keeps it binding (Proposition 12.1). The bear case adds three pressures to the treadmill: open-weight models already cost far less than comparable closed ones (3:61); \$518B of commitments is a fixed cost while the price of a fixed capability keeps falling (3:48); and the labs may slow themselves, as on 14 September 2026, when AI stocks fell worldwide after Dario Amodei wrote "We must slow the pace at which we improve the capabilities of AI models" and Sam Altman agreed ([CNBC](#)). Calling \$2T cheap is a bet on the dates of §11.8.

^{16:38} Forecasts that some listed AI suppliers will rise 1,000% start from base rates. [Bessembinder's](#) study of 29,754 US common stocks from 1926 to 2025 finds a median lifetime buy-and-hold return of -6.9%; only 41.2% of the stocks beat one-month Treasury bills, and 1,082 firms, 3.72% of the total, account for all \$91T of net wealth creation. The best annualized return among stocks with twenty or more years of data was NVIDIA's 37.0% a year since its 1999 listing, and "identifying such stocks in advance is a formidable challenge, to say the least."

^{16:39} A rise of 1,000% is an elevenfold multiple; a thousandfold rise is ninety times larger. At NVIDIA's century-best 37% a year, eleven times takes $\ln 11 / \ln 1.37 \approx 7.6$ years and a thousand times $\ln 1000 / \ln 1.37 \approx 21.9$, about as long as NVIDIA took, and a thousandfold rise would make a \$1T company worth several times world output. Elevenfold is plausible for small companies and has already been exceeded: SanDisk is up about 3,558%, roughly 37-fold, since its spin-off began trading in February 2025 ([24/7 Wall St](#)). By Proposition 16.1, a candidate needs a small base in a layer whose cost share is rising and whose scarcity index stays positive for years, competitors slow to build, a balance sheet that survives Proposition 16.2, and a price that does not already discount all of it. The last is the hard one: once an asymmetry is visible on an earnings call its price usually encodes it, which is Proposition 16.3 applied to investors. Nothing here recommends a security.

16.7 Checkpoints of the economy

^{16:40} A capex thesis needs a verifier: observable checkpoints, cheap to check and expensive to fake, that confirm or refute it before its conclusion arrives (Definition 1.1). If cognition's falling price shows up instead as falling prices for memory, power and skilled labor, this chapter's account of where the rent goes is wrong (§15.6).

^{16:41} Five readings are the leading edges of Proposition 16.2. Hyperscaler capex exceeds operating cash flow ([Goldman Sachs](#)); the hyperscalers' net investment-grade issuance reached about \$274B in 2026 to date against \$130B in 2025, ten-year spreads widened from 40–75 bp toward 90 bp, and UBS counts about \$3.5T of obligations once leases and commitments join \$0.8T of reported debt ([UBS](#)). Free cash flow has turned negative at Alphabet and Oracle (16:14), and loans for an Oracle-leased campus traded at 89–91¢ on the dollar ([Disruption Banking](#)). Morgan Stanley counted "\$156 billion of projects ... canceled or delayed in 2025" and "almost that same exact number" in the first quarter of 2026 ([Morgan Stanley](#)). Goldman's break-even of about \$300B a year of AI revenue compares with cloud revenue about \$70B above its pre-AI trend ([Startup Fortune](#)). And part of the

financing is circular: NVIDIA's guarantees on OpenAI's Ohio leases are capped at \$105B ([NVIDIA](#)), and its CFO expects demand it supports with its balance sheet to be "roughly a quarter of our business next year" ([NVIDIA's call](#)). The [BIS](#) warns that disappointment "could trigger a sudden pullback in financing and turn the capex boom into a protracted investment bust."

16:42

Checkpoint	The binding constraint holds if	It moves if
Big-four capex guidance	2027 guides reach \$1.1–1.4T	guides are cut, or leases are reclassified to hide cuts
DRAM and HBM contract prices	contracts extend into 2028 at rising prices	prices fall two quarters running
Turbine and transformer lead times	reservations turn into firm orders	reservations lapse; lead times fall below three years
CoWoS capacity	packaging still limits accelerator shipments	capacity outruns orders
PJM capacity price	auctions clear at or near the cap	auctions clear well below it
Hyperscaler credit	spreads hold near 90 bp	spreads widen; project loans fall further below par
AI revenue against break-even	lab run-rates keep compounding	run-rate growth stalls for two quarters
Price of a fixed capability	it keeps falling while input prices rise	it falls together with input prices

16:43 The chain also predicts the order in which the binding constraint moves, since the scarcity index falls first where capacity answers fastest: power equipment and packaging bind in 2026–27, lithography and memory in 2028–29, where Patel places the ultimate limit, and then the cost of capital, as the debt-funded \$5T of SemiAnalysis's \$11T comes up for refinancing ([Patel](#)). The richest layer of one phase need not be the richest of the next. I put the whole sequence at one in five, since every step must arrive in order.

16:44

Forecast 16.3 (The constraint migrates). By the end of 2029, the binding constraint of the AI stack has moved from power equipment and packaging (2026–27) to lithography and memory (2028–29), and then to the cost of capital, with rents migrating in the same order.

Horizon 2029-12-31

Probability 20%

Check The forecast holds only if the steps arrive in order: at the end of 2027, turbine and transformer lead times are still above three years or packaging still limits accelerator shipments; by the end of 2028, those lead times are below three years while DRAM or HBM contract prices are still rising or EUV tool output is the reported limit on accelerator supply; and during 2029, hyperscaler ten-year credit spreads widen while their issuance keeps growing. It fails otherwise.

References

1. Platformonomics. <https://platformonomics.com/2026/07/follow-the-capex-q2-2026-scoreboard/>
2. TMT Finance. <https://www.tmtfinance.com/intel/2026-hyperscaler-capex-tops-us700bn-analysis>
3. lowdown. <https://lowdown.today/t/data-centres/13/the-big-fours-combined-capex-holds-near-732bn/>

4. Morgan Stanley. <https://www.morganstanley.com/insights/podcasts/thoughts-on-the-market/ai-data-centers-political-pushback-capital-spending-outlook-ariana-salvatore>
5. Dwarkesh Patel's podcast. <https://www.dwarkesh.com/p/dylan-patel-3>
6. Dell'Oro. <https://www.delloro.com/news/ai-boom-drives-data-center-capex-to-1-7-trillion-by-2030/>
7. Epoch. <https://epoch.ai/gradient-updates/frontier-labs-dont-use-most-ai-compute>
8. note. <https://doc.mbalib.com/view/102f41fce8562ccf7086e0124c5b5cf6.html>
9. Goldman Sachs. <https://themalaysianreserve.com/2026/09/26/goldman-sees-hyperscaler-ai-capex-rising-50-to-us1-2t/>
10. Morgan Stanley. <https://www.listennotes.com/podcasts/thoughts-on-the-can-the-ai-spending-boom-pay-9FwTmYCIEd1/>
11. Yahoo Finance. <https://finance.yahoo.com/markets/article/nvidia-ceo-jensen-huang-just-doubled-down-on-his-big-2030-prediction-115945089.html>
12. Dell'Oro. <https://www.delloro.com/news/ai-buildout-maintains-momentum-as-data-center-capex-surpasses-3-trillion-by-2030/>
13. McKinsey's. <https://web.archive.org/web/20260518031027/https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers>
14. IMF projects. <https://www.imf.org/-/media/files/publications/weo/2026/april/english/tablea.pdf>
15. March. <https://www.dwarkesh.com/p/dylan-patel>
16. GE Vernova. https://www.gevernova.com/sites/default/files/gev_webcast_pressrelease_07222026.pdf
17. Siemens Energy. <https://www.siemens-energy.com/global/en/home/press-releases/earnings-release-q3-fy-2026.html>
18. SemiAnalysis. <https://newsletter.semianalysis.com/p/us-grid-constraints-towards-40gw>
19. PJM. <https://www.pjm.com/-/media/DotCom/about-pjm/newsroom/2025-releases/20251217-pjm-auction-procures-134479-mw-of-generation-resources.pdf>
20. World Nuclear News. <https://www.world-nuclear-news.org/articles/nrc-completes-environmental-review-of-crane-restart>
21. Caterpillar. <https://www.sec.gov/Archives/edgar/data/18230/000001823026000040/ex992toformcat2q2026retail.htm>
22. Rehlko. <https://www.rehlko.com/newsroom-backup-power-orders-secured-for-data-center-growth>
23. Rehlko. <https://www.prnewswire.com/news-releases/rehlko-doubles-annual-backup-power-capacity-at-changzhou-manufacturing-facility-302883757.html>
24. Rehlko. <https://www.rehlko.com/kohler-energy-is-now-rehlko>
25. Vertiv. <https://investors.vertiv.com/news/news-details/2026/Vertiv-Reports-Strong-Fourth-Quarter-with-Organic-Orders-Growth-of-252-and-Diluted-EPS-Growth-of-200-Adjusted-Diluted-EPS-37/>
26. Vertiv. <https://www.sec.gov/Archives/edgar/data/1674101/000162828026050323/q22026exhibit991vrt07292026.htm>
27. Eaton. <https://www.eaton.com/content/dam/eaton/company/investor-relations/quarterly-earnings/filings/2026/q2/q2-2026-analyst-presentation.pdf>
28. Eaton's call. <https://stockanalysis.com/stocks/etn/transcripts/660780-q2-2026/>
29. TSMC. https://investor.tsmc.com/english/encrypt/files/encrypt_file/reports/2026-07/547d1696765e05ce3adb81c108ce1c8c1682b80c/TSMC%202Q26%20Transcript.pdf
30. TrendForce. <https://www.trendforce.com/news/2026/06/15/news-tsmc-cowos-supply-demand-gap-reportedly-seen-narrowing-from-20-to-10-by-end-2026-as-capacity-expands/>
31. Mizuho. <https://www.investing.com/news/stock-market-news/mizuho-lifts-tsmc-cowos-capacity-forecasts-as-server-cpu-demand-surges-4769157>
32. Yonhap. <https://en.yna.co.kr/view/AEN20260903010700320>
33. SK hynix. <https://news.skhynix.com/en/q2-2026-business-results/>
34. Samsung. <https://news.samsung.com/global/samsung-electronics-announces-second-quarter-2026-results>
35. Micron. <https://investors.micron.com/news/press-release/2026/Micron-Technology-Inc--Reports-Record-Fiscal-Fourth-Quarter-and-Full-Year-2026-Results/default.aspx>
36. SanDisk. <https://investor.sandisk.com/news-releases/news-release-details/sandisk-reports-fiscal-fourth-quarter-2026-financial-results>
37. SanDisk. <https://www.sandisk.com/company/newsroom/press-releases/2026/2026-08-03-Sandisk-and-sk-hynix-advance-global-standardization-of-hbf>

38. Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, Kurt Keutzer (2024). AI and Memory Wall. *arXiv*. <https://arxiv.org/abs/2403.14123>
39. NVIDIA. <https://investor.nvidia.com/news/press-release-details/2026/NVIDIA-Announces-Financial-Results-for-Second-Quarter-Fiscal-2027/default.aspx>
40. NVIDIA's call. https://s201.q4cdn.com/141608511/files/content_files/TRANSCRIPT_NVIDIA-Corp-NVDA-US-Q2-2027-Earnings-Call-26-August-2026-5_00-PM-ET.pdf
41. futurex. <https://futurex.capital/en/ai-lab/reports/big-tech-ai-capex-2026q2>
42. UBS. <https://www.ubs.com/cz/en/assetmanagement/insights/investment-outlook/the-red-thread/trt-end-year-2026/articles/webs-of-ai-debt.html>
43. cdelta. <https://www.cdelta.ch/publications/the-circular-machine/>
44. CoreWeave. <https://www.sec.gov/Archives/edgar/data/1769628/000176962825000047/crwv-20250909.htm>
45. Goldratt. <https://northriverpress.com/wp-content/uploads/2018/01/Free-download-5FS.pdf>
46. Christian Catalini, Xiang Hui, Jane Wu (2026). Some Simple Economics of AGI. *arXiv*. <https://arxiv.org/abs/2602.20946>
47. ideas.repec.org, 1938. <https://ideas.repec.org/a/oup/qjecon/v52y1938i2p255-280.html>
48. StockTitan. <https://www.stocktitan.net/articles/micron-q4-fy2026-earnings-memory-cycle>
49. Grossman and Stiglitz. <https://www.aeaweb.org/aer/top20/70.3.393-408.pdf>
50. Pascual Restrepo (2025). We Won't be Missed: Work and Growth in the AGI World. *National Bureau of Economic Research*. <https://www.nber.org/papers/w34423>
51. Anthropic. <https://www.anthropic.com/news/series-h>
52. Reuters. <https://www.reuters.com/technology/anthropic-revenue-run-rate-tops-65-billion-source-says-2026-08-17/>
53. Reuters. <https://live.euronext.com/en/financial-news/anthropics-ipo-prospectus-shows-sweeping-ai-vision-surging-costs>
54. Mint. <https://www.livemint.com/companies/news/anthropic-ipo-may-come-in-november-why-its-2-trillion-valuation-is-raising-eyebrows-11790902031997.html>
55. Reuters. <https://finance.yahoo.com/technology/ai/articles/exclusive-anthropics-ipo-prospectus-shows-231722972.html>
56. PitchBook. <https://pitchbook.com/news/reports/q3-2026-anthropic-ipo-impact>
57. TechCrunch. <https://techcrunch.com/2026/09/29/openai-reportedly-in-talks-to-raise-30b-round-at-1-4t-valuation/>
58. Tech Funding News. <https://techfundingnews.com/openai-eyes-30b-at-1-4t-valuation-as-bridge-financing-while-ipo-waits-report/>
59. Hock Tan. <https://stockanalysis.com/stocks/avgo/transcripts/685349-q3-2026/>
60. Bain. <https://www.prnewswire.com/news-releases/2-trillion-in-new-revenue-needed-to-fund-ais-scaling-trend---bain--companys-6th-annual-global-technology-report-302563362.html>
61. CNBC. <https://www.cnbc.com/2026/09/14/ai-stocks-slowdown-amodei-altman.html>
62. Bessembinder's. <https://www.fundresearch.de/fundresearch-wAssets/docs/100-Jahre-Studie-Aktien-ssm-6438198.pdf>
63. 24/7 Wall St. <https://247wallst.com/investing/2026/10/02/10000-put-into-septembers-best-performing-large-cap-is-worth-a-lot-more-now-is-there-room-left-in-october/>
64. Disruption Banking. <https://www.disruptionbanking.com/2026/09/24/could-ai-data-centre-financing-become-a-systemic-risk/>
65. Startup Fortune. <https://startupfortune.com/goldman-sachs-says-hyperscalers-will-spend-12-trillion-on-ai-in-2027/>
66. NVIDIA. <https://www.sec.gov/Archives/edgar/data/1045810/000104581026000069/nvda-20260817.htm>
67. BIS. <https://www.bis.org/publications/aer-2026/progress-peril>