

The Factories: What It Costs to Build Intelligence

Shay O'Kelly · August 2026

If the last essay was right that intelligence has become a manufactured good, then the obvious next question is what the factories cost, and the answer is that we are living through the largest infrastructure buildout in the history of capitalism. The four big hyperscalers alone, Amazon, Google, Meta, and Microsoft, have guided to roughly \$725 billion of capital expenditure for 2026, up 77% from about \$410 billion the year before¹. Add OpenAI's \$500 billion Stargate program, the sovereign projects, and the neocloud buildouts, and a trillion dollars a year is within sight. In the first quarter of 2026, AI data centers, hardware, and networking amounted to 1.4% of US GDP and drove roughly three quarters of the economy's growth². Relative to GDP this is already bigger than the telecom buildout at the peak of the dot-com era, and on the bullish forecasts it reaches 4% of GDP, which is railroad-boom territory³. That is the right comparison, and not as hyperbole. This is the canal-and-railroad phase of an industrial revolution. Enormous fixed investment, deployed ahead of the revenue, carrying the whole economy while it happens.

But the railroad analogy breaks in one place, and the break is where all the interesting questions live. The money is buying two completely different kinds of assets that get lumped into one capex line. The first is the durable layer: land, buildings, cooling, substations, transmission, and power generation, assets that will run for thirty to fifty years no matter whose chips sit inside. The second is the consumable layer, the chips themselves, which are roughly half the spend and carry accounting lives of three to six years. A railroad was bought once and used for a century. GPUs have to be repurchased forever. Which means intelligence does not have a construction cost, it has an ongoing production cost, like electricity. The capex line is really an opex line in disguise, and that is exactly what you would expect once intelligence becomes a manufactured good rather than a built one.

The consumable layer is where the bears attack, and the attack deserves to be taken seriously. Michael Burry and others have accused the hyperscalers of overstating earnings by stretching GPU depreciation schedules⁴, and the accounting point is real. Small changes in assumed useful life shift billions of dollars of reported profit without changing a dollar of cash. If a chip is economically dead in three years but depreciated over six, the industry's earnings are inflated and the buildout is more fragile than it looks.

Here is why I think the bears have the depreciation story backwards. Look at what old chips actually earn. The H100 launched in 2022 and rented for as much as \$8 an hour at the peak of the 2023 shortage. By late 2025 the price had collapsed to around \$1.70, which is the moment the depreciation bears point to. But then something happened that their model says should not happen, the price went back up, rising almost 40% to about \$2.35 by March 2026, because on-demand H100 capacity is effectively sold out across the market and renters who locked in capacity refuse to release it⁵. A four-year-old chip, two full generations behind the frontier, is getting more expensive to rent. The reason is that training and inference are different businesses. Training the next frontier model demands the newest silicon, but serving models

to users does not, and modern mixture-of-experts architectures only activate a fraction of their parameters per query, which makes older chips perfectly viable inference machines. So the installed base does not retire when a new generation ships. It migrates from training to inference and stays full. And the economics of that inference work are absurd by any labor standard. When a machine does fifty dollars an hour of work for two dollars, someone will always rent it.

There is also a supply-side reason obsolescence runs slower than the accounting debate assumes. Every leading-edge chip on earth passes through the EUV lithography machines of a single company, ASML, which will ship about 65 of them in 2026, up from 44 last year, and perhaps 85 next year⁶. Not thousands. Tens. The entire industry's ability to replace old chips with new ones is rationed by a production line measured in dozens of machines a year, which is why Blackwell lead times stretched into mid-2026 even as Nvidia ramped as fast as it could. Depreciation is not really a function of a chip's age. It is a function of replacement supply, and when replacement is rationed, old capital stays economically alive far longer than the spreadsheet says. In a supply-constrained world, the choice is not between old chips and new chips. It is between old chips and no chips.

The other thing changing is who writes the checks. The first phase of the buildout was funded from hyperscaler operating cash flow, arguably the safest financing in corporate history. That era is ending. Meta's Hyperion campus in Louisiana was financed through a \$30 billion special purpose vehicle, with \$27 billion of A+ rated debt anchored by PIMCO and BlackRock, the largest private credit deal ever done⁷. Stargate runs on stacked SPV debt from Blue Owl, JPMorgan, and others. Estimates of off-balance-sheet AI debt now run north of a trillion dollars, and senators on the Banking Committee have started writing letters about it⁸. I would resist the urge to treat this as a scandal, because it is how every great buildout in history was financed. The railroads ran on bonds too. But it does move risk from equity holders who can absorb losses to credit markets that have historically been bad at underwriting technology obsolescence. That leaves long-dated paper against assets whose useful life is the single most contested number in the market, and the tension is real. The old-chip economics above suggest the collateral lives longer than the bears claim. It does not suggest the debt is riskless.

The sharpest version of the financing critique is circularity. Nvidia invests in OpenAI and CoreWeave, which use the money to buy Nvidia chips, and Nvidia books the revenue. AMD handed OpenAI warrants on 160 million shares priced at a penny each, vesting as OpenAI buys AMD GPUs. By this summer, Nvidia was reportedly discussing guaranteeing up to \$250 billion of OpenAI's data center lease payments and financing another \$350 billion of chip purchases⁹. Bears call this vendor financing, and they are right. That is exactly what it is. It is the same structure that let Lucent and Nortel manufacture a demand boom in 1999, and the concern has escalated from an analyst complaint to an official one. The Bank for International Settlements now lists circular AI financing among the biggest risks to global financial stability¹⁰.

It matters, though, to separate two things the bears lump together. An equity stake in a customer is self-limiting. If Nvidia puts two billion dollars into CoreWeave and CoreWeave fails, Nvidia loses two billion dollars, takes a write-down, and moves on. A guarantee is different in kind, not in size. If Nvidia guarantees

\$250 billion of OpenAI's lease payments and OpenAI's revenue stalls, those obligations do not disappear with the customer. They land on Nvidia. The lenders who financed those data centers get paid by Nvidia or they do not get paid at all, and their debt is not held by venture capitalists who priced in failure. It is rated A+ and sits with insurers, pension funds, and credit funds that bought it as a safe investment. Equity losses stop with the holder. Credit losses cascade, because the holders are themselves leveraged, and a failure would not just slow the boom, it would ripple through the bond market from inside the most important company in the world. So my rule for reading these deals is simple. The investments are noise, the guarantees are the risk. Watch the guarantees.

But here is what the circularity debate keeps missing, and where I part ways with the bears entirely. Vendor financing is only fatal when the end customer never shows up. That is what actually killed the telecom vendors. They financed carriers building networks for traffic that did not exist, so when the lending stopped there was no revenue under any layer of the stack. Which means the whole question collapses to one measurable thing. Is there real, paying, arms-length demand at the end of the loop? That is no longer a matter of opinion. Google processed 3.2 quadrillion tokens in May, seven times more than a year earlier and more than three hundred times the level of two years ago¹¹. Microsoft was running over 100 trillion tokens a quarter, five times its prior year, as far back as the spring of 2025¹². Token consumption is already running at more than double what the equipment makers were forecasting for 2028. On revenue, Anthropic went from one billion to forty-seven billion dollars of annualized revenue in less than a year and a half, the fastest revenue scaling in corporate history, with OpenAI around twenty-five billion¹³, and that money comes overwhelmingly from businesses paying market prices, not from Nvidia's balance sheet. Nobody is subsidizing the two dollar an hour H100 renter, and the market for those four-year-old chips is sold out anyway. The sold-out market for old chips is the fifty-for-two trade already happening at scale. So my bull case is simple. The loop is real, and it is being lapped by demand. Compute consumption is compounding at several hundred percent a year while the physical capacity to serve it grows at fifty. The vendor financing exists not because demand is missing, but because demand is arriving faster than any balance sheet can build for it. The 1999 comparison fails on the only fact that matters. This time the end customer showed up. What we have is a shortage wearing a bubble costume.

Which leaves the ceiling question. How big can the demand actually get? The bear math stacks a trillion of annual capex against AI revenues measured in the low hundreds of billions and calls it a bubble. I think the reference class is wrong, the same way it was wrong in the last essay. This infrastructure is not underwritten against software budgets. It is underwritten against the global wage bill for cognitive work, which runs to tens of trillions of dollars a year. Do the division. A trillion dollars of annual infrastructure spend needs to capture only low single digits of that wage bill to clear its cost of capital. Priced against SaaS, the buildout looks indefensible. Priced against labor, it looks conservative.

So the checks are being written, the debt is being raised, and the spending already shows up in national output as the largest single driver of American growth. And yet consensus forecasts for the decade ahead barely move. The same two percent growth, the same productivity trend, as if a railroad-scale buildout were a rounding error. Someone is badly wrong, and the next essay is about why I think it is the forecasters, not the builders.

SOURCES

1. Hyperscaler 2026 capex guidance (~\$725B combined, up 77% from ~\$410B). Company guidance compiled by ValueAdd VC, CreditSights, and Yahoo Finance.
2. AI capex at 1.4% of US GDP and ~75% of Q1 2026 growth. Via Paul Kedrosky, "Honey, AI Capex is Eating the Economy," Epoch AI, and TECHi.
3. Railroad and telecom era comparisons as a share of GDP. Via Paul Kedrosky and Epoch AI.
4. GPU depreciation debate and stretched useful lives. Deep Quarry via National Law Review, and Michael Burry public commentary.
5. H100 rental price history (\$8 peak 2023, ~\$1.70 October 2025, ~\$2.35 March 2026, sold-out on-demand capacity). Via GPUSmith and IntuitionLabs.
6. ASML EUV shipments (44 in 2025, ~65 in 2026, 80-85 guided 2027). ASML guidance via TechPowerUp and CryptoBriefing.
7. Meta Hyperion \$30B SPV with \$27B A+ debt (PIMCO, BlackRock). Via Global Datacenter Hub and Capacity.
8. Off-balance-sheet AI debt above \$1T, and the US Senate Banking Committee letter (January 2026). Via Techerati and the US Senate Banking Committee.
9. Nvidia-OpenAI discussions (\$250B lease guarantees, \$350B chip financing). Wall Street Journal and Bloomberg reporting, July 2026. Nvidia \$2B CoreWeave investment plus \$6.3B capacity agreement, and AMD-OpenAI 160M share warrants. Via Yahoo Finance and EBC Financial Group.
10. BIS 2026 Annual Report naming circular AI financing a top financial stability risk. Bank for International Settlements.
11. Google 3.2 quadrillion tokens per month (7x YoY). Google I/O 2026.
12. Microsoft 100T+ tokens per quarter (5x YoY). Microsoft FY2025 Q3 earnings call, April 30, 2025.
13. Anthropic ~\$47B and OpenAI ~\$25B annualized revenue. Via Trending Topics, ValueAdd VC, and Epoch AI.

No Offer or Solicitation. The only purpose of this site is to provide general background information on RAXIA. This is not an offer to sell or the solicitation of any offer to buy any securities. Any such offer would be made only to qualified investors after distribution and return of the Private Placement Memorandum and governing agreement for the applicable Fund (the "Offering Materials"). Nothing on this site is investment advice or a recommendation of any security. The author may hold positions in companies discussed. Past performance is not indicative of future results.